UNIVERSIDADE FEDERAL DE MINAS GERAIS

INSTITUTO DE CIÊNCIAS BIOLÓGICAS

PROGRAMA DE DOUTORADO EM BIOINFORMÁTICA

TESE

A GENÔMICA COMO FERRAMENTA PARA SELEÇÃO DE ALVOS CONTRA A LINFADENITE CASEOSA

ORIENTADO: **Anderson Rodrigues dos Santos**

ORIENTADOR: **Prof. Dr. Vasco Azevedo**

CO-ORIENTADOR: **Prof. Dr. Artur Silva**

BELO HORIZONTE - MG

Abril de 2012

A GENÔMICA COMO FERRAMENTA PARA SELEÇÃO DE ALVOS CONTRA A LINFADENITE CASEOSA

Manuscrito apresentado como requisito parcial para obtenção do grau de Doutor pelo programa de Pós-Graduação em Bioinformática, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais.

ORIENTADO: **Anderson Rodrigues dos Santos**

ORIENTADOR: **Prof. Dr. Vasco Azevedo**

CO-ORIENTADOR: **Prof. Dr. Artur Silva**

BELO HORIZONTE - MG

Abril de 2012

Dedicatória

Dedico este manuscrito a todos que:

acreditaram na nossa capacidade de fazê-lo;

contribuíram para a viabilização de todas as etapas dessa longa jornada;

e, por último e igualmente importante, todos que

não criaram dificuldades para a sua realização.

"Quando avançamos nas fronteiras do conhecimento
nos conscientizamos que pouco sabemos"

AGRADECIMENTOS

Aos meus orientadores, pela oportunidade única de crescimento;

Às agências de fomento FAPEMIG e CNPQ;

Aos colegas do LGCM que gastaria páginas citando;

Aos familiares e amigos.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AA	Aminoácidos
BAC	*Bacterial Artificial Chromosome*
CDS	*Coding Sequence*
DNA	Acido Desoxirribonucleico
EMBL	*European Molecular Biology Laboratory*
GO	*Gene Ontology*
HLA	*Human Leukocyte Antigen*
HMM	*Hidden Markov Model*
LC	Linfadenite Caseosa
LGCM	Laboratório de Genética Celular e Molecular
MED	*Mature Epitope Density* ou Densidade de Epitopos Maduros = [(50-Média(Afinidade de ligação forte ao MHC))*Quantidade de epitopos preditos/Extensão de AA's – 9 + 1]
MHC	*Major Histocompatibility Complex*
NCBI	*National Center for Biotechnology Information*
NR	*Non Redundant*
ORF	*Open Reading Frame*
PERL	Linguagem de programação dinâmica de alto nível e interpretada
PSE	*Potencialy Surface Exposed* ou Potencialmente Exposta na Superfície
RGMG	Rede Genômica de Minas Gerais
RPGP	Rede Paraénse de Genômica e Proteômica
SQL	*Structured Query Language*
VR	Vacinologia Reversa

RESUMO

Foram depositadas no sítio do NCBI os genomas das linhagens 1002, C231, I19, PAT10 e FRC41 da *Corynebacterium pseudotuberculosis,* entre os anos de 2009 e 2011. Utilizou-se a pangenômica aliada à predição de proteínas exportadas, primeira etapa da Vacinologia Reversa (VR), sobre esses cinco genomas de *C. pseudotuberculosis*, uma metodologia que ao buscar alvos vacinais parte de um conjunto genomas completos, de predições *in silico*, ao invés de antígenos isolados por abordagens tradicionais. Dentre estas linhagens, adotou-se como modelos as linhagens 1002 e C231 que infectam caprinos e ovinos, respectivamente. Apesar da *C. pseudotuberculosis* causar outras doenças em animais como, por exemplo, equinos e bovinos, a escolha das linhagens 1002 e C231 como modelo baseou-se na importância da enfermidade linfadenite caseosa (LC). A LC é uma doença contagiosa crônica associada com perdas econômicas consideráveis e distribuída mundialmente. A inexistência de vacinas eficientes e métodos de diagnóstico contra a LC impulsionaram a pesquisa sobre os mecanismos moleculares da patogênese dessa bactéria. Para realizar a predição *in silico* de proteínas exportadas nas cinco linhagens de *C. pseudotuberculosis*, foi utilizado um *pipeline* com programas de predição de motivos proteicos de exportação combinados, no qual o resultado de processamento de um programa serve como parâmetro para iniciar o processamento de outros programas. De um total aproximado de 17 mil proteínas, nos cinco genomas, essa análise resultou em 750 proteínas preditas como secretadas. Dentre estas, foram preditas 139 e 149 proteínas secretadas nos genomas de *C. pseudotuberculosis* linhagens 1002 e C231, respectivamente. Dentre essas predições, foram confirmadas 87 e 77 proteínas como sendo secretadas, por meio de trabalhos de proteômica, nas linhagens 1002 e C231, respectivamente. Dentre essas proteínas, 55 são comuns a ambas as linhagens. Ao considerarmos a grande similaridade entre os genomas de *C. pseudotuberculosis*, supõem-se que os resultados experimentais das linhagens modelos são válidos para as demais linhagens da espécie. Análises *in silico* sobre o potencial imunogênico do proteoma predito, juntamente com o proteoma comprovado experimentalmente, evidenciaram quinze proteínas detentoras de uma concentração maior de epitopos preditos por extensão da porção madura dessas proteínas. Dentre essas quinze proteínas, quatro se destacaram em termos de seus potenciais imunogênicos analisados *in silico* e estão em fase de testes experimentais na busca de vacinas, drogas e diagnósticos contra LC.

ABSTRACT

Between the years of 2009 and 2011, the genomes of strains 1002, C231, I19, PAT10 and FRC41 of *Corynebacterium pseudotuberculosis* were deposited at the site of NCBI. We used pangenomics allied to the prediction of exported proteins on these five genomes as the first step of Reverse Vaccinology (RV), a methodology that aims at finding vaccines starting from *in silico* predictions of full genomes rather than isolated antigens by traditional approaches. Among all five strains, strains 1002 and C231, that infect goats and sheep respectively, were adopted as models. Although *C. pseudotuberculosis* also causes other diseases in animals, such as horses and cattle, the selection of strains 1002 and C231 as models was based on the importance of the disease caseous lymphadenitis (CLA) and also due to the high similarity degree between these genomes. CLA is a worldwide distributed chronic infectious disease associated with considerable economic losses. The absence of effective vaccines and diagnostics against CLA boosted research on the molecular mechanisms of pathogenesis of this bacterium. To predicted *in silico* exported proteins into five *C. pseudotuberculosis* strains, a combination of exportation motifs predicting programs were arranged in a pipeline, a schema were a program's processing result serves as initial parameter to initiate the processing of the other programs. We obtained 750 proteins predicted as secreted, out of approximately 17.000 proteins of the five genomes. Among these results, 139 and 149 were predicted as secreted proteins in strains 1002 and C231 respectively. Within these predictions, there were 87 and 77 proteins confirmed as secreted by proteomics studies in the strains 1002 and C231 respectively. Within these proteins, 55 are common to both model strains. The results obtained for the model strains were considered valid also for the other strains of this species. *in silico* analysis of the predicted immunogenic potential of the exoproteome, allied with the experimentally proven exoproteome demonstrated fifteen proteins owning a differential predicted epitope's concentration per protein's mature portion. Among these fifteen proteins, four stood out in terms of their *in silico* immunogenic potential and are currently under experimental tests in the search for vaccines, drugs and diagnosis against CLA.

Apresentação da Tese

Colaboradores

Este trabalho foi realizado no Laboratório de Genética Celular e Molecular (LGCM) do instituto de Ciências Biológicas (ICB) da Universidade Federal de Minas Gerais (UFMG). Contou com a parceria do Laboratório de Polimorfismos de DNA (LPDNA) da Universidade Federal do Pará. Os pesquisadores envolvidos foram:

- Prof. Dr. Artur Silva, pesquisador e professor do Laboratório de Polimorfismos de DNA do Instituto de Ciências Biológicas da UFPA e coordenador da Rede Paraense de Genômica e Proteômica (RGPG), Belém, Pará;

- Prof. Dr. Roberto Meyer, chefe do Laboratório de Imunologia e Biologia Molecular do Instituto de Ciências da Saúde da UFBA, Salvador, Bahia;

- Profa. Dra. Débora de Oliveira Lopes, pesquisadora e professora do Laboratório de Genética Molecular da Universidade Federal de São João del-Rei, Divinópolis, Minas Gerais;

- Dr. Robert Moore, pesquisador do CSRIO *Livestock Industries*, Austrália.

- Prof. Dr. Andreas Tauch, pesquisador do Centro de Biotecnologia (CeBiTec) da Universidade de Bielefeld, Bielefeld, Alemanha;

- Dr. Jan Baumbach, pesquisador do Instituto de Informática Max-Planck, Saarbrücken, Alemanha;

- Dr. Debmalya Barh, pesquisador do Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Índia;

- Prof. Dr. Nahum Shpigel, pesquisador e professor da *Hebrew University of Jerusalem, Koret School of Veterinary Medicine*, Jerusalém, Israel.

Esse trabalho contou com o apoio financeiro das instituições: Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e Fundação de Amparo à Pesquisa do Estado do Pará (FAPESPA).

Delineamento

Com o advento de novas tecnologias de sequenciamento de genomas; transcriptomas e proteomas, surgiram novas disciplinas como a Patogenômica. A Patogenômica utiliza dados genômicos coletados através de tecnologias de *Next Generation Sequence* para estudar a diversidade microbiana e tentar entender as interações entre o agente infeccioso e o hospedeiro, bem como o desenvolvimento da doença. O nosso grupo de pesquisa se especializou na montagem de genomas, anotação funcional gênica, busca de candidatos a alvos para drogas e vacinas, bem como a busca por genes importantes para a sobrevivência de bactérias patogênicas. Esse conhecimento adquirido foi convertido em 16 artigos científicos, dois capítulos de livro, dois trabalhos de vulgarização e depósito em bancos de dados públicos de 15 genomas da espécie *Corynebacterium pseudotuberculosis*. Estas publicações foram divididas em três grupos compreendendo trabalhos sobre Genômica Estrutural, Genômica Funcional e Aprendizado de Máquina. A última etapa da cadeia de publicações em Genômica foi a busca por alvos vacinais para *C. pseudotuberculosis*. O presente manuscrito de tese encontra-se subdivido em seções:

1. A primeira seção é uma introdução geral, estando dividida em quatro subseções. Na primeira subseção é descrita a bactéria *C. pseudotuberculosis,* agente etiológico da doença linfadenite caseosa (LC), principal problema abordado nessa tese; na segunda divisão são abordadas: a história da arte da vacinação, os tipos de vacina, os esforços realizados na tentativa de criar uma solução vacinal contra a LC por meio de diferentes estratégias vacinais e como fazer uso de genomas para a predição de alvos vacinais por meio da técnica de Vacinologia Reversa (VR); na terceira divisão os conceitos, cenário atual e perspectivas da pangenômica aliada à VR são revistos; por fim, na quarta divisão, são abordados conceitos de imunoinformática, uma técnica que geralmente é adotada em conjunto com a predição de alvos exportados.

2. Na segunda seção são apresentadas as justificativas que levaram à realização deste trabalho de tese e os objetivos a serem alcançados.

3. A terceira seção é composta por resultados e a discussão dos mesmos, estando dividida em duas partes. A primeira parte se refere ao uso da **Genômica Estrutural**. Nessa parte utilizou-se sequenciadores de próxima geração e análises de plasticidade genômica para montar diversos genomas de *C. pseudotuberculosis*, bem como anotar por completo os genomas montados,

tópico no qual nosso grupo de pesquisa não se atém apenas a inferir um produto gênico para ORF's, mas predizer todas as estruturas possíveis de um genoma como, por exemplo, RNA's, operons, motivos de exportação, domínios proteicos, local subcelular de proteínas e capacidade de peptídeos de proteínas induzirem respostas imunes em hospedeiros. A segunda parte dos resultados se refere ao uso da **Genômica Funcional**, na qual em posse de genomas completamente montados e anotados é possível analisar o genótipo buscando o entendimento do fenótipo de diferentes linhagens de *C. pseudotuberculosis*. Nesse contexto sugeriu-se a pangenômica de proteínas exportadas sobre os genomas de *C. pseudotuberculosis*, estratégia de busca por alvos vacinais. Em cada uma das subseções dessas estratégias estarão anexadas as respectivas publicações de nosso grupo de pesquisa. Também explorou-se uma técnica de Aprendizado de Máquina não-supervisionado para o tratamento de sequências biológicas que devem ser agrupadas. Essa estratégia explora técnicas de Inteligência Artificial para mineração de dados biológicos, permitindo-nos extrair informações não evidentes à primeira vista em um genoma.

4. Na quarta seção estão contidas as conclusões e perspectivas desse trabalho de tese;

5. A quinta seção é composta das referências bibliográficas;

6. Na sexta seção estão os anexos, nela encontramos as metodologias adicionais às utilizadas nos artigos científicos. Também estão alguns resultados que não entraram no escopo principal da tese, mas foram utilizados ou gerados nos estágios intermediários da terceira seção. Dentre esses subprodutos estão técnicas computacionais para identificar diferenças entre genomas, um programa para criação de mapas de BAC's e uma análise da utilização de códons de espécies do gênero *Corynebacterium*. Por último, é apresentado o *Curriculum Vitae* do doutorando.

1. Introdução

1.1 O agente etiológico e a doença

1.1.1 Aspectos microbiológicos

Corynebacterium pseudotuberculosis é um patógeno intracelular facultativo pertencente ao grupo de actinobactérias conhecido como CMNR (*Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus*). Esse grupo possui características em comum como: (i) parede celular composta principalmente de peptideoglicanos, arabinogalactano e ácidos micólicos, e (ii) o alto conteúdo G+C do genoma (47-74%). A esse grupo também pertencem a *Corynebacterium diphtheriae* e *Mycobacterium tuberculosis* que são importantes patógenos que acometem humanos. Vários genomas do grupo CMNR têm sido depositados no NCBI indicando a crescente importância médica, veterinária e biotecnologia desses organismos (Dorella e cols., 2006).

A *C. pseudotuberculosis* é uma bactéria pleomórfica, detentora de fímbrias, imóvel e anaeróbica facultativa. Exibe formas que variam desde cocóides a bastões filamentosos, não possui cápsula e não esporula. Suas dimensões variam de 0.5 a 0.6 micrômetros por 1.0 a 3.0 micrômetros (Dorella e cols., 2006). A Figura 1 mostra uma foto de microscopia eletrônica de transmissão dessa bactéria.

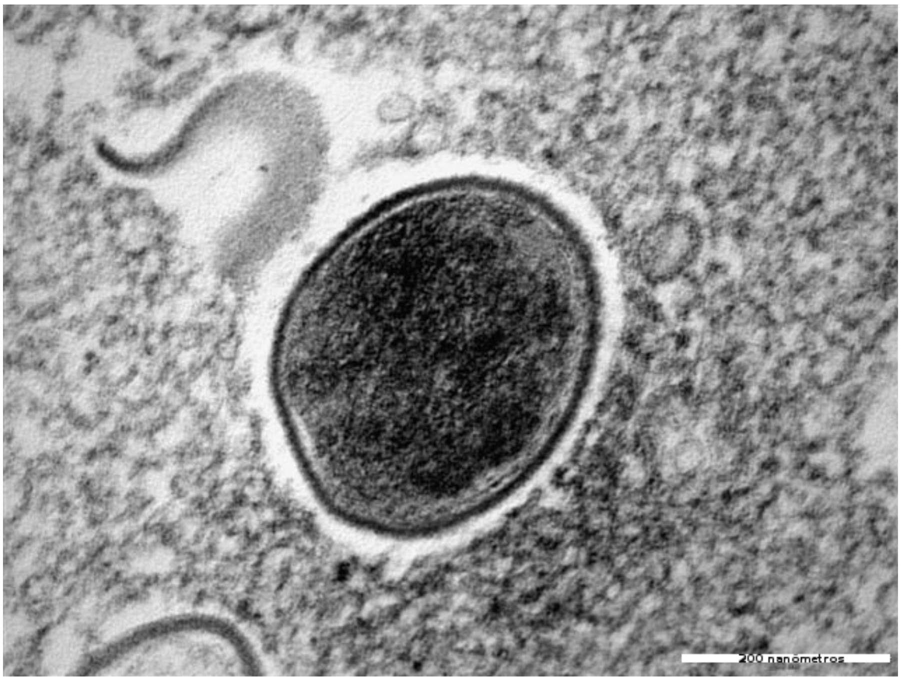


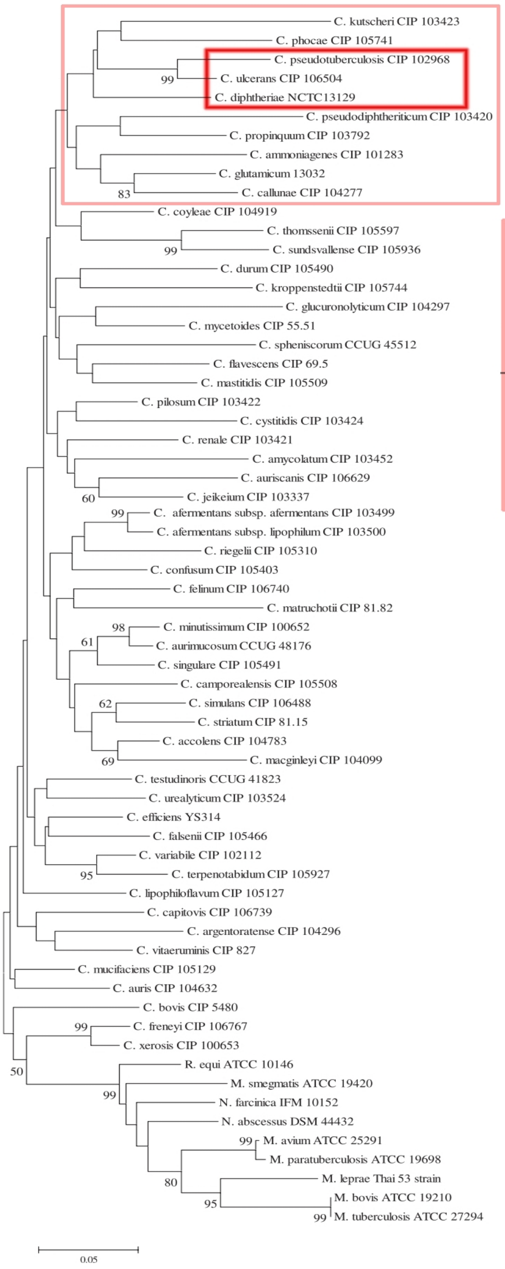
Figura 1: Microscopia eletrônica de transmissão da *C. pseudotuberculosis,* obtida pelo microscópio Zeiss-EM10 do CEMEL-UFMG. A barra de escala, no canto inferior direito, representa 200 nanômetros.

O peptideoglicano da parede celular apresenta, além de outros constituintes, o ácido meso-diaminopimélico relacionado à resistência da parede contra a maioria das peptidases, mas arabinose e galactose são os principais açúcares da parede (Dorella e cols., 2006a). As reações bioquímicas de isolados de *C. pseudotuberculosis* variam, principalmente em sua habilidade de fermentação. Todas as linhagens produzem ácido, mas não gás, a partir de fontes de carbono incluindo glicose, frutose, maltose, manose, e sacarose (Holt e cols., 1994).

1.1.2 Taxonomia

A classificação taxonômica de *C. pseudotuberculosis* foi baseada primeiramente pelas características bioquímicas e morfológicas (Muckle e Gyles, 1982). Quanto à capacidade de redução de nitrato a nitrito, dois biovares foram identificados: tipo I, isolada, preferencialmente, de ovino ou caprino, nitrato negativo e tipo II, equino e bovino, nitrato positivo. Porém, outros autores têm relatado que ambos os tipos podem ser isolados de bovinos (Barakat e cols., 1984). Ainda não está claro se a disseminação e doença visceral estão relacionadas à diferença na linhagem, aos fatores do hospedeiro, ou ambos (Shpigel e cols., 1993). Para diferenciação entre gêneros de *Corynebacterium* os dois biotipos, tipo I e tipo II, a técnica de análise de restrição de rDNA amplificado (ARDRA), que estuda o gene da subunidade 16S do rRNA bacteriano foi utilizada e mostrou-se eficiente (Vaneechoutte e cols., 1995). De acordo com estes estudos, *C. pseudotuberculosis* está filogeneticamente mais próximo de *C. ulcerans* do que de *C. diphtheriae*. A proximidade destes organismos também foi sugerida pelo fato de serem os únicos do gênero *Corynebacterium* a produzirem fosfolipase D (Buck e cols., 1985).

Análises de sequenciamento do gene da subunidade β da RNA polimerase (*rpoB*) mostraram-se mais acuradas para a identificação das espécies de corinebactérias do que análises baseadas no rDNA 16S (Khamis e cols., 2004). Esses autores demonstraram, através da árvore filogenética construída baseada na sequência de gene *rpoB* (Figura 2), a clara relação entre *C. pseudotuberculosis* e *C. ulcerans* mostrando que o gene *rpoB* é uma poderosa ferramenta de identificação. Porém, muitos autores propõem que as análises de subunidade β da RNA polimerase seja utilizada de forma complementar as análises do gene rRNA 16S para estudos filogenéticos das espécies de *Corynebacterium* e *Mycobacterium* (Khamis e cols., 2004).



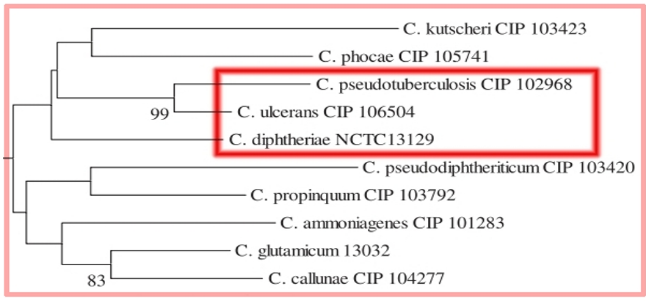


Figura 2: Dendrograma adaptado de Dorella e cols. 2006 com espécies do grupo CMNR.

Representa relações filogenéticas de espécies dos gêneros *Corynebacterium*, *Mycobacterium*, *Nocardia* e *Rhodococcus* obtidos pelo método de *neighbor-joining*. A árvore foi derivada de alinhamentos de sequências do gene *rpoB*. A confiabilidade de cada ramo, foi determinada a partir de 1000 amostras (*bootstrap*), e está indicada em porcentagem em cada nó da árvore.

1.1.3 Agente etiológico de diferentes doenças

C. pseudotuberculosis é um patógeno importante na pecuária mundial. É o agente etiológico de doenças como: a linfadenite caseosa, linfangite ulcerativa, e acne contagiosa ou dermatite ulcerativa que acomete caprinos/ovinos, equinos e bovinos, respectivamente. Essas são doenças com maior impacto econômico por afetarem rebanhos ou animais de considerável valor econômico (Brown e Olander, 1987). Além dessas doenças, também é o agente etiológico da linfangite piogranulomatosa, dermatite ulcerativa granulomatosa, pneumonia, mastite superficial e visceral (Yeruham e cols., 1997; Yeruham e cols., 2003). *C. pseudotuberculosis* também foi isolada de espécies como camelos, cervos, alpacas, rinocerontes, porcos, roedores, macacos, búfalos e lhamas. Em humanos, foram relatados aproximadamente 25 casos de contaminação por esse microrganismo: Peel e cols. (1997) revisaram 22 relatos nos quais observaram a presença de linfadenite e abscessos; Mills e cols. (1997) descreveram uma linfadenite granulomatosa supurativa em um garoto devido ao contato com animais contaminados; um paciente de 63 anos de idade foi acometido por uma infecção no olho devido a um implante ocular (Liu e cols., 2005); uma criança de 12 anos da França apresentou linfadenite necrosante na virilha causada por *C. pseudotuberculosis*, que foi tratada de maneira errônea pela falta de conhecimento do microrganismo e da patologia (Join-Lambert e cols., 2006; Trost e cols., 2010).

1.1.4 Transmissão

A transmissão da *C. pseudotuberculosis* ocorre principalmente pela ingestão de água e alimentos contaminados bem como através de ferimentos superficiais na pele, os quais podem ser causados tanto por procedimentos de manejo como tosquia, castração, tratamento do cordão umbilical e agulhas contaminadas quanto por fatores naturais como arbustos pontiagudos (Alves & Pinheiro, 1997).

A contaminação ambiental é considerada um fator de grande importância na disseminação da enfermidade já que *C. pseudotuberculosis* é capaz de sobreviver no ambiente por longos períodos. Sob baixas temperaturas e condições de umidade, o tempo de sobrevivência pode ser prolongado. O microrganismo pode sobreviver em frestas de piso à temperatura ambiente por até 10 dias e cerca de 8 meses ou mais de um ano em fômites, principalmente em baixas temperaturas ambiente (Collett e cols., 1994).

1.1.5 Determinantes moleculares de virulência e patogenicidade

1.1.5.1 Fosfolipase D

Existem poucos determinantes moleculares bem caracterizados para patogenias causadas pela *C. pseudotuberculosis*: a fosfolipase D (PLD), uma fosfolipase secretada com ação de esfingomielinase, o operon f*agABC* e o gene *fagD*, constituídos de componentes do tipo ABC-transportador relacionados à captação de ferro (Billington e cols., 2002); e lipídios da superfície celular bacteriana (Hard, 1975). A PLD tem sido considerada o principal fator de virulência para a *C. pseudotuberculosis* (Hodgson e cols., 1992; Lipsky e cols., 1982; Carne e cols., 1956).

C. pseudotuberculosis e *C. ulcerans* são as únicas espécies do gênero *Corynebacterium* que produzem a exotoxina fosfolipase D (Barksdale e cols., 1981). Essa exotoxina permite a persistência e a disseminação do organismo dentro do sistema linfático do hospedeiro (McNamara e cols., 1995). Ela é um fator de permeabilidade, que promove a hidrólise de ligações éster na esfingomielina nas membranas de células de mamíferos, podendo contribuir para a disseminação da bactéria a partir do local inicial da infecção para sítios secundários dentro do hospedeiro (Carne e Onon, 1978; Coyle e Lipsky, 1990; McNamara e cols., 1995). Além disso, provoca lesões dermonecróticas, e em doses mais elevadas, é letal para um número de diferentes espécies de animais domésticos e de laboratório (Egen e cols., 1989; Songer, 1997). Foram observados danos e destruição dos macrófagos de caprinos durante a infecção por *C. pseudotuberculosis*. Este efeito letal é devido à ação da fosfolipase D (Tashjian e Campbell, 1983).

A fosfolipase D é uma enzima encontrada em bactérias, fungos, plantas e animais (Pépin e cols., 1993). A esfingomielina hidrolisada por essa enzima pertence a uma classe de lipídeos cuja função parece estar ligada à transmissão de sinais e reconhecimento celular através da membrana de células. Sabe-se que desordens no metabolismo dessa classe de lipídeos podem causar impactos no tecido nervoso. A análise da sequência do gene *pld* de *C. pseudotuberculosis* revela que este codifica uma proteína de 31.4 kDa, compartilhando 80% e 64% de identidade, respectivamente, com a proteína de *C. ulcerans* e *Arcanobacterium haemolyticum* (Hodgson e cols., 1990; McNamara e cols., 1995). A bactéria *A. haemolyticum* chegou a ser confundida com pertencente ao gênero *Corynebacterium* e causa faringite aguda em humanos (Linder, 1997).

Várias das atividades biológicas da PLD de *C. pseudotuberculosis*, bem como a sua estrutura molecular, também foram encontradas em esfingomielinases no veneno da aranha

Loxosceles intermedia (Bernheimer e cols., 1985; Coyle e Lipsky, 1990; Tambourgi e cols., 2002) e o uso de antitoxina tem reduzido a disseminação de *C. pseudotuberculosis* no hospedeiro (Williamson, 2001).

1.1.5.2 Operon *fagABC* e o gene *fagD*

Os genes denominados *fag A, B, C* e *D* de *C. pseudotuberculosis* estão localizados próximos ao gene *pld*, conforme exibido na Figura 3 do genoma da linhagem 1002 de *C. pseudotuberculosis*.

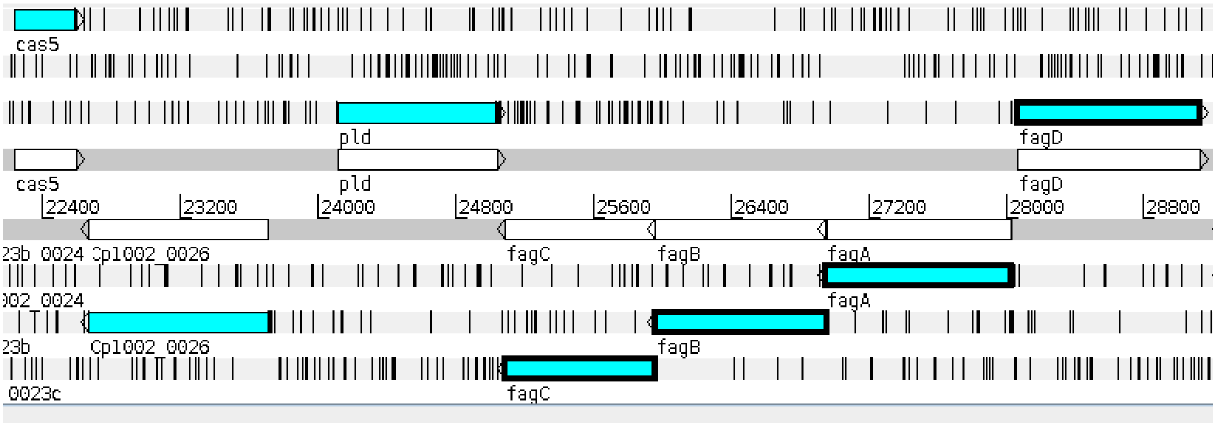


Figura 3: Disposição dos genes com prefixo *fag* no genoma completo da linhagem 1002.

Os genes do operon fagABC apresentam grande similaridade com proteínas de outras bactérias, como os transportadores ABC (Billington e cols. 2002). Os genes *fagA* e *fagB* se assemelham com permeases da membrana citoplasmática; *fagD* se assemelha a uma proteína sideróforа ligante ao ferro e *fagC* se assemelha à proteína citoplasmática ligante de ATP. Para investigar os efeitos da disponibilidade de ferro na regulação desses genes foi feita uma fusão com *lacZ* no início da primeira janela de leitura do operon. A atividade do operon *fagABC* foi baixa em um meio rico em ferro e três vezes maior quando em meio pobre em ferro, indicando que esse operon é induzido por condições que limitam a quantidade de ferro em um meio (Billington e cols. 2002). Foi demonstrado que os genes *fagB* e *fagC* afetam a patogenicidade de *C. pseudotuberculosis* e a introdução de um cassete de canamicina impediu a expressão do gene *fagB,* bem como reduzindo a expressão do gene *fagC.* A utilização de ferro não foi modificada na linhagem mutante, mas a virulência foi reduzida quando comparadas com linhagens selvagens (Billington e cols. 2002). Esse mesmo experimento não conseguiu reproduzir esses resultados nos genes *fagA* e *fagD*, permanecendo indeterminado a influência desses genes na virulência da *C. pseudotuberculosis*.

1.1.5.3 Lipídeos tóxicos da parede celular

Os lipídios de superfície de *C. pseudotuberculosis* têm sido descritos como fatores importantes que contribuem para a patogênese (Carne e cols., 1956; Hard, 1972; Hard, 1975). A toxicidade do material lipídico extraído foi demonstrada pela indução de necrose hemorrágica após a injeção intradérmica de cobaias. Em análise do fluído peritoneal de cobaias infectadas, foram encontrados macrófagos altamente suscetíveis à ação necrosante de lipídeos de superfície de *C. pseudotuberculosis* (Hard, 1975). No entanto, a infecção com *C. pseudotuberculosis* em cobaias progride até a morte, invariavelmente, enquanto os macrófagos de cobaias não são sensíveis à ação citotóxica dos lipídios bacterianos (Hard, 1975). Tashjian e Campbell (1983) observaram que *C. pseudotuberculosis* foi resistente a morte e digestão por macrófagos caprinos devido a seu revestimento lipídico.

Um estudo realizado em ratos com 25 linhagens isoladas de *C. pseudotuberculosis* propôs que existe uma relação direta entre a porcentagem de lipídios de superfície com a indução de abscessos crônicos (Muckle e Gyles, 1983).

1.1.6 Linfadenite caseosa

1.1.6.1 A doença

A linfadenite caseosa (LC) é uma doença infecto contagiosa crônica, que se caracteriza pela hipertrofia dos gânglios linfáticos localizados ao longo do corpo do animal, principalmente caprinos e ovinos. Esta doença é responsável por perdas econômicas significativas relacionadas à redução da produtividade, carne, leite, lã, pele, e eficiência reprodutiva dos animais infectados (Dorella e cols., 2006a e 2009).

A forma mais frequente da doença, a LC externa, é caracterizada pela formação de abscessos em nódulos linfáticos superficiais e em tecidos subcutâneos, assim como mostrado na Figura 4. Esses abscessos podem também se desenvolver em órgãos internos, tais como pulmões, rins, fígado e baço, caracterizando a LC visceral (Piontkowski & Shivvers, 1998). Em alguns casos, a infecção produz poucos sinais clínicos no animal, o que leva à impossibilidade de identificá-los, tornando difícil a obtenção de dados sobre a prevalência dessa doença (Arsenault e cols., 2003).

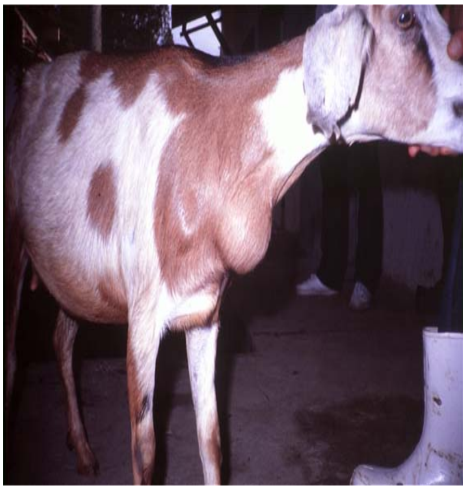


Figura 4: Caprino com quadro clinico de LC, apresentando abcesso na região ventral do pescoço.

1.1.6.2 Epidemiologia

A LC foi identificada em países detentores de grandes criações de ovinos e caprinos como Austrália, Argentina, Nova Zelândia, África do Sul e Estados Unidos e em países da Comunidade Européia (França, Itália, Grã-Bretanha, União Soviética), Chile, Uruguai, Canadá, e Sudão (Dorella e cols., 2009b).

No Brasil, a *C. pseudotuberculosis* foi isolada e caracterizada em diversos estados (Oliveira, 2007) que detêm 3% do rebanho mundial de caprinos e ovinos. (EMBRAPA Semi-Árido, 2008). Na região nordeste, a incidência de LC foi verificada ao examinarem 656 caprinos periodicamente, durante dois anos, e 41,6% dos animais apresentaram abscessos superficiais palpáveis. Os prejuízos provocados pela doença são grandes, uma vez que muitos dos pequenos criadores têm a caprinovinocultura como uma das suas principiais atividades econômicas. Os abscessos causam danos à pele do animal e levam à condenação da carne, principalmente quando o produto é destinado ao comércio exterior. Por isso, além do impacto econômico, a doença tem um grave impacto social (Oliveira, 2007).

Estudos epidemiológicos recentes mostram que existe alta incidência da doença também na região sudeste. A prevalência da LC entre ovelhas no estado de São Paulo chega a 71%, enquanto a prevalência em Minas Gerais, estado onde a ovinocultura tem apresentado grande crescimento, é de aproximadamente 75,8% entre ovinos (Guimarães e cols., 2011) e 78,9% entre caprinos (Seyffert e cols., 2010).

1.1.6.3 Diagnóstico

Apesar de sua importância, os diagnósticos, tratamentos ou vacinas que existem não são eficientes ou práticos no combate à LC. O diagnóstico atualmente utilizado é baseado em cultura bacteriológica do material purulento, recolhido de animais com abscessos externos, e posterior identificação bioquímica. Esse procedimento demanda tempo e custo sendo incapaz de identificar infecções subclínicas. Entretanto, Pacheco e cols. (2007) implantaram o diagnóstico por meio de uma técnica molecular denominada multiplex PCR utilizando três genes, *rpoB*, rRNA 16S e *pld,* obtendo resultados 4,35% melhores do que a cultura bacteriológica, o atual padrão ouro de diagnóstico, apesar de também não identificar infecções subclínicas.

1.1.6.4 Tratamento

O uso de antibióticos não é aconselhável visto que, além de bastante longo (dura de semanas a meses), não é totalmente eficaz, uma vez que os fármacos são incapazes de penetrar na cápsula dos abscessos. Inspeções periódicas do rebanho, isolamento dos animais doentes, tratamento e desinfecção de qualquer tipo de ferimento superficial, além da limpeza das instalações, são algumas das medidas profiláticas que podem ajudar a conter a doença (Williamson, 2001; Dorella e cols., 2006a; Dorella e cols., 2009).

1.2 Vacinas: História da Arte

1.2.1 As primeiras vacinas

Há mais de duzentos anos deu-se início à era das vacinas com a descoberta, por Edward Jenner (1749 – 1823), em 1796, do princípio da vacinação. Ele observou que existiam, nas tetas das vacas, lesões similares às causadas pela varíola em humanos. Jenner percebeu que as mulheres responsáveis pela ordenha das vacas, quando expostas ao vírus da varíola, desenvolviam uma versão mais branda da doença.

Para testar sua hipótese, Jenner recolheu o exsudato que saía de tais feridas e escarificou a pele e expôs um menino saudável ao material contaminado. O garoto teve febre e algumas lesões, mas se recuperou rapidamente. Semanas depois, o cientista expôs novamente o menino ao material da ferida de um paciente e desta vez o menino passou incólume à doença. Estava descoberta a primeira vacina de que se tem notícia (Morgan e Parker, 2007).

Cem anos depois, Louis Pasteur (1822-1895), entre pesquisas sobre fermentação alcoólica, anaerobiose e doenças infecciosas, descobriu que existiam formas menos e mais virulentas de micróbios e que as primeiras poderiam ser usadas como agentes imunizantes contra a infecção pelas últimas. Iniciou-se assim a utilização da vacinação como forma de prevenção de doenças. Entre os diversos campos da ciência em que atuava, coube a Pasteur a descoberta do agente transmissor da raiva bem como da vacina antirrábica. Esta e outras descobertas científicas levaram à fundação do renomado Instituto Pasteur (Arroio, 2006) e à época de ouro da vacinologia.

As descobertas de Jenner e Pasteur fizeram parte do início de uma era muito importante para a medicina: a era da vacinação. Até então, ao longo da história, grandes epidemias de doenças infectocontagiosas atingiram inúmeros países, cada uma com suas particularidades. Hoje em dia, doenças consideradas controladas, ou até mesmo extintas, estão reemergindo no cenário mundial. E no combate a essas doenças, a vacinação ainda é a melhor alternativa (Mahmouda e Levin, 2007).

1.2.2 Tipos de vacinas

Basicamente existem três tipos de vacinas, levando-se em consideração o modo como são formuladas.

1.2.2.1 Primeira geração

As vacinas de primeira geração consistem na utilização de patógenos vivos atenuados ou mortos, capazes de induzir uma resposta imune protetora contra estes mesmos agentes em sua forma selvagem. A vacina BCG (Bacilo de Calmette-Guérin) amplamente utilizada no combate à tuberculose faz parte da categoria de vacinas de primeira geração viva e atenuada (Silva e cols., 2004). Além dela, vacinas contra varíola, raiva, peste bubônica, difteria, coqueluche, febre amarela, poliomielite, sarampo e rubéola são vacinas de primeira geração. Embora este tipo de vacina seja predominante no mercado, atualmente sua eficácia e segurança é bastante contestada bem como a questão ética na utilização de organismos vivos, mesmo que atenuados, em seres humanos (Schatzmayr, 2003; Movahedi e Hampson, 2008). Essas questões de biossegurança constituem uma importante plataforma de trabalho para o desenvolvimento de vacinas mais seguras (Dougan, 1994; Foss e Murtaugh, 2000). Para contornar estes problemas, hoje em dia são empregadas técnicas moleculares que permitem torná-las seguras, diminuindo a virulência do patógeno e evitando que haja reversão. Entre essas técnicas, a remoção de múltiplos genes, relacionados à virulência, ou mesmo inserções de fragmentos que interrompam tais genes no patógeno, diminuem o risco de um patógeno reverter a sua forma virulenta (Schatzmayr, 2003).

1.2.2.2 Segunda geração

As vacinas de segunda geração surgiram na década de 80 e substituíram o microrganismo como um todo por subunidades proteicas do mesmo. Estas subunidades são antígenos purificados ou recombinantes, isolados de cultura do próprio patógeno, responsáveis por estimular mais fortemente o sistema imune do hospedeiro. No caso da tecnologia do DNA recombinante esse tipo de vacina se baseia no isolamento do gene, clonagem, expressão e capacidade da proteína recombinante de induzir uma resposta imune protetora no hospedeiro. As vantagens deste tipo de vacina são a de que uma única proteína pode ter vários epitopos imunodominantes capazes de induzir uma imunidade protetora sem riscos e adversidades de se administrar um patógeno vivo; vacinas de subunidades não se replicam no hospedeiro, portanto não há risco de patogenicidade, além da possibilidade de produção em larga escala (Movahedi e Hampson, 2008; Silva e cols., 2004). A vacina contra a hepatite B, que contém fragmentos antigênicos da superfície do vírus da hepatite, isolados do sangue de portadores de hepatite, é um exemplo bem sucedido de vacina de segunda geração (Briles e cols., 2003). No entanto, existem dificuldades na produção desse tipo de vacina como a purificação dos antígenos e a expressão de proteínas eucarióticas em procariotos. Além disso, a escolha de antígenos deve ser cuidadosa visto que é necessário que os epitopos eleitos induzam imunidade humoral e/ou celular no hospedeiro de acordo com o tipo de organismo a ser combatido (Movahedi e Hampson, 2008).

1.2.2.3 Terceira geração

No início da década de 1990 observou-se que o DNA plasmidial poderia ser transfectado em células animais *in vivo*. Em 1993, uma molécula de DNA *naked* codificando genes virais conferiu imunidade protetora e trouxe surpresa aos vacinologistas (Manickan e cols., 1997). A partir desta observação surgiu o questionamento: genes poderiam virar vacinas? Desde então surgiram as vacinas de terceira geração, compostas de genes ou fragmentos gênicos que codificam antígenos potencialmente imunogênicos carreados por DNA plasmidial visando induzir resposta imune protetora através da injeção direta do DNA no tecido animal (Donnelly, 2003). As vantagens oferecidas pelas vacinas de terceira geração são bastante atrativas: não apresentam risco de causar infecção, uma vez que os genes utilizados são específicos para uma dada proteína antigênica; pode-se empregar um mesmo plasmídeo carreando vários genes, o que é vantajoso quando o organismo é altamente variável como um vírus; a imunidade adquirida persiste por longo tempo devido à constante produção do antígeno dentro da célula hospedeira. Além de boa estabilidade em baixas ou altas temperaturas, o que facilita enormemente a estocagem e distribuição (Oliveira, 2004). E finalmente, as vacinas são de fácil preparação e baixo custo de produção, o que traz grandes esperanças no campo da vacinologia (Beláková e cols., 2007). Uma das principais desvantagens deste tipo de vacina é o fato de não apresentar resultados tão promissores no hospedeiro final quanto em modelos laboratoriais. Por exemplo, em estudos realizados com vacinas de DNA no combate à malária, os excelentes resultados obtidos em modelo murino não se reproduziram em humanos (Dunachie e Hill, 2003).

Os benefícios resultantes do desenvolvimento de vacinas gênicas são inúmeros, como mencionado anteriormente. E o impacto gerado sobre o controle das doenças infecciosas que podem ser prevenidas por imunização gênica será, provavelmente, uma das aquisições mais importantes advindas da utilização desta tecnologia.

A imunização baseada em DNA deu início a uma nova era no ramo da vacinologia, e com isso a busca pelo desenvolvimento de novas alternativas vacinais capazes de controlar doenças infecciosas de modo profilático.

1.2.3 Vacinas contra linfadenite caseosa

Diversas estratégias vacinais contra a LC foram experimentadas, havendo inclusive vacinas comerciais disponíveis, porém nenhuma alcançou uma vacina eficaz e de fácil administração.

Em 1971 foi isolada uma linhagem naturalmente atenuada de *C. pseudotuberculosis*, denominada linhagem 1002, permitindo à Empresa Baiana de Desenvolvimento Agrícola S.A (EBDA) criar uma vacina utilizada em caprinos. Em 1987, LeaMaster e cols. fizeram ensaios de vacinação em ovinos utilizando uma cultura inativada da bactéria, também conhecida como bacterina. Em 1989, Holdstad e cols. fizeram um ensaio de vacinação em um rebanho de 247 caprinos jovens utilizando como vacina uma solução contendo o organismo inteiro do agente etiológico misturado com o toxoide PLD filtrado. Brogden e cols. (1990) avaliaram concentrações de células inativadas com e sem a presença do dipeptídeo muramil (MDP) capaz de conferir relativa proteção em cobaias. Eggleton e cols. (1991) combinaram em uma vacina o toxoide da *C. pseudotuberculosis* com cinco antígenos clostridianos e mostrou que existe relação entre a quantidade de toxoide e a severidade dos efeitos da LC em ovinos. Nesse mesmo ano, Ellis e cols. ministraram anticorpos neutralizantes de exotoxinas obtidas de culturas da bactéria conseguindo diminuir *in vitro* o crescimento da bactéria. Hodgson e cols. (1992) deletaram o gene *pld* por meio de mutagênese sítio específica criando uma linhagem atenuada de *C. pseudotuberculosis* denominada *Toxminus* e caracterizando o gene *pld* como importante para o estabelecimento de uma infecção. Pépin e cols. (1993) inocularam uma linhagem de *C. pseudotuberculosis* sensível ao antibiótico estreptomicina na orelha de ovinos que após o desafio com uma linhagem resistente a esse mesmo antibiótico não desenvolveram lesões. Em 1994, Hodgson e cols. utilizaram a linhagem *Toxminus* de *C. pseudotuberculosis* previamente desenvolvida (Hodgson e cols., 1992) como vetor vacinal expressando uma forma inativa de PLD em ensaios de imunização oral em ovinos. Alcançou-se uma proteção de cerca de 50% dos animais, enquanto que os animais imunizados somente com *Toxminus* não alcançaram proteção. Pogson e cols (1996) deletaram o gene *recA* responsável por recombinação gênica em *C. pseudotuberculosis* com o intuito de utilizar a bactéria como um vetor capaz de carrear antígenos para o interior de células de cobaias. O experimento foi bem sucedido por aumentar a expressão gênica dos antígenos e por não causar uma aumento na virulência da bactéria em camundongos. Em 1997 e 1998, Simmons e cols. utilizaram como vacina em ovinos linhagens mutantes *aroQ* e *pld* atenuados a partir das linhagens parentais C231 e TB521 de *C. pseudotuberculosis*, respectivamente. Impulsionados pelas expectativas introduzidas com o terceiro paradigma vacinal (vacinas de DNA), Chaplin e cols. (1999) construíram e testaram uma versão atenuada de PLD apresentada em vacina de DNA na imunização de ovinos numa tentativa de criar uma vacina de DNA contra LC. O resultado dessa vacina de DNA foi o aumento da resposta humoral, gerando uma proteção parcial contra o desafio experimental de *C. pseudotuberculosis*. O resultado foi similar ao obtido pela vacina de subunidade inativada por formalina contra LC em ovinos (Muckle e Gyles, 1982; Paton e cols., 1994; Stanford e cols., 1998). Coelho (2007) também criou uma vacina de DNA e outra recombinante contra LC utilizando o gene da proteína do choque térmico *hsp60*, porém as cobaias vacinadas não apresentaram proteção alguma contra LC.

Entre o intervalo dos experimentos de Chaplin e Coelho, foram lançadas no mercado duas vacinas comerciais para tentar combater a LC, em 2001 a Biodectin® e em 2004 a Glanvac™3. Ambas utilizam toxoides inativados de clostrídios, sendo que a Biodectin® possui frações inativadas de *C. pseudotuberculosis e* diversos outros antígenos, ao passo que a Glanvac™3 possui como componente adicional a PLD.

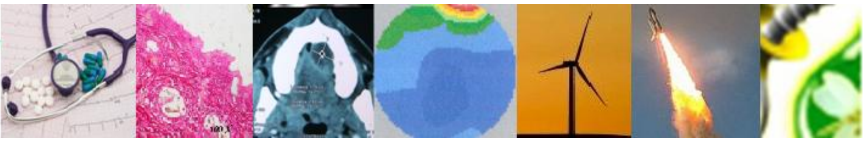
Outro evento na pesquisa contra a linfadenite caseosa, diz respeito ao trabalho de Dorella (2009b). Utilizando mutagênese aleatória com o TnFuZ em 34 linhagens recombinantes de *C. pseudotuberculosis*, linhagem T1, foram identificados 21 loci diferentes. Estes 21 mutantes foram utilizados em ensaios de imunização visando a busca por novas alternativas vacinais contra a LC. Análises da produção de imunoglobulinas e citocinas foram realizadas para identificar o padrão da resposta imune após a imunização. Dentre as 21 linhagens testadas, as linhagens vivas atenuadas denominadas CP13 (proteína secretada ligada ao sistema de transporte de ferro) e CP09 (subunidade fimbrial) mostraram os melhores níveis de proteção sendo de 80 e 60% em camundongos, respectivamente. Camundongos imunizados com o mutante CP09 apresentaram 60% de proteção contra o desafio, o mesmo resultado observado no grupo de camundongos imunizados com a vacina comercial Glanvac™3. Essas linhagens mutantes estão sendo testadas em modelos ovinos (Ribeiro, Tese em andamento).

Três vacinas foram testadas contra LC, porém até a presente data nenhuma proteção efetiva foi alcançada. A escolha de alvos capazes de estimular uma resposta humoral e celular contra LC pode ser o fator para que uma vacina apresente eficácia protetora. Na busca por alvos vacinais, a escolha de genes adequados foi uma tarefa limitada pela ausência de genomas do agente etiológico, a bactéria *C. pseudotuberculosis*. Considerando que poucos antígenos vacinais são conhecidos contra LC, a busca por novos alvos vacinais é importante para combater a doença.

1.2.4 Vacinologia Reversa

Nos últimos anos, a ampla disponibilidade de sequências genômicas completas mudou a maneira de se pensar sobre alvos vacinais. De uma dúzia de alvos potenciais agora pode-se contar com centenas de alvos por organismo. Esta infinidade de candidatos vacinais é extensivamente analisada com base no conceito de Vacinologia Reversa (VR), com atenção especial reservada para as alvos exportados, gerando resultados promissores para vários organismos. No entanto, deve-se ter em mente que ainda não há vacinas eficazes para organismos sequenciados há mais de uma década, um período maior do que o esperado para a produção de uma vacina eficaz por meio da VR. Esta consideração leva à reflexão que, na pesquisa de uma vacina, outras variáveis podem ser tão importantes como a escolha de alvos. Deve-se levar em consideração que o universo de possibilidades para uma vacina eficaz pode seguir uma distribuição matemática exponencial, uma ordem de grandeza de 2^n em que n é o número de variáveis envolvidas. Esta revisão compila os resultados de algumas pesquisas chave usando o conceito de VR e levanta algumas questões potenciais que podem atrapalhar o uso eficiente dessa técnica para o alcance de alvos atraentes e promissores para a pesquisa de vacinas.

Este trabalho, cujo doutorando foi o primeiro autor, fundamentou as ações seguintes do grupo de pesquisa na busca por alvos para vacinas, drogas e diagnóstico contra a linfadenite caseosa. Devido a ausência de alelos de MHC específicos dos hospedeiros mais comuns da bactéria *C. pseudotuberculosis* definiu-se, por exemplo, que uma nova abordagem imunoinformática precisaria ser utilizada contra a linfadenite caseosa.



IIOAB
ISSN: 0976-3104

The IIOAB Journal

REVIEW: IMMUNOMICS AND VACCINOLOGY

THE REVERSE VACCINOLOGY - A CONTEXTUAL OVERVIEW

Anderson Santos[1], Amjad Ali [1], Eudes Barbosa[1], Artur Silva[2], Anderson Miyoshi[1], Debmalya Barh[3], and Vasco Azevedo[1*]

[1]*Biochemistry Departament, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, BRAZIL*

[2] *DNA Polimorfism Laboratory, Universidade Federal do Pará, Campus do Guamá - Belém, PA, BRAZIL*

[3] *Centre for Genomics and Applied Gene Technology, Institute Of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal-721172, INDIA*

Received on: 14th-Nov-2010; **Revised on:** 26th-Dec -2010; **Accepted on:** 8th-Feb-2011; **Published on:** 3rd-May-2011.

[*]**Corresponding author:** Email: vasco@icb.ufmg.br Tel: +55-3134092610; Fax: +55-3134092610

ABSTRACT

In recent years, the wide availability of complete genome sequences has changed the way we think about vaccine targets. From a few dozen potential targets we can now count on hundreds of targets per organism. This candidate vaccine is an extensively scrutinized plethora based on the concept of reverse vaccinology (RV) with special attention reserved for exported targets, generating promising results for various organisms. However it should be borne in mind that we still lack effective vaccines for organisms sequenced within a decade, a period much longer than expected for producing an effective vaccine by RV. This consideration leads to the reflection that, in the research on a vaccine, other variables may be as important as choosing a target exported. Attention is paid to the fact that the universe of possibilities for an effective vaccine can be exponential in the order of 2^n where n is the number of variables. This review compiles results of some key research using the concept of RV and raises some potential issues that may be hindering the efficient use of this technique to attain attractive and promising targets for vaccine research.

.

Keywords: *reverse vaccinology; vaccine variables chances; exported proteins; exponential function; vaccine candidates*

[I] INTRODUCTION

A decade has passed since the term *Reverse Vaccinology (RV)* was first introduced [1]. RV starts from the genomic sequence of a pathogen, which is an expected codified sequence for all the possible genes expressed in the life cycle of the pathogen. All Open Reading Frames (ORF's) derived from the genome sequence can be evaluated with a computer program in order to determine their ability to be vaccine candidates. Special attention is given to exported proteins because they are essential in host pathogen interaction. Examples of this interaction can be cited: (i) adherence to host cells, (ii) the invasion of the cell to which there was compliance, (iii) damage to host tissues, (iv) resistance to environmental stresses from machinery defense of the cell being infected and finally, (v) mechanisms for subversion of host immune response [2-5].

The word 'Reverse' from RV can be explained by the reverse genetics (RG) technique. Before the dawn of genomics, there have been attempts to discover the responsible genes from one phenotype. With Crick's Central Dogma (DNA → RNA → Protein) the research path was reversed. In possession of the likely gene sequence, several techniques were used to identify changes in the phenotype of an organism derived from sequence changes in genes. The principle of the Crick's dogma is also used by RV, in which possession of a gene sequence is searched for the possibility of a probable protein encoded by this sequence to be an antigen capable of stimulating an immune response in a host organism.

Long before the creation of the term RV, a number of approaches had been considered to meet the demand of exported proteins in order to move to the next step of the production of a subunit vaccine [6]. For example, the research using exported proteins was motivated as alternative to subunit vaccines based on polysaccharide capsule of meningococci. Vaccines produced with such antigens have low capacity to induce a satisfactory immune response. This research effort on exported proteins dates back to almost two decades of work searching for a vaccine against meningococcal serogroup B, and now it produces good results. This vaccine currently is the best RV research results in the production of a subunit vaccine for *Neisseria meningitidis* serogroup B. Meningitis caused by serogroup B (Men B) is responsible for approximately half of the worldwide incidence of the disease [6] and this research result for targeted vaccination is commonly used as a reference card for the RV due to its excellent results. Currently a subunit vaccine against Men B created with antigens targeted by RV is selected in clinical trials of phase 2 [7, 8]. The advantages of RV are still attractive, enabling vaccine research for organisms whose cultivation in the laboratory remains difficult or impossible. However, reducing the time of selection by target proteins is feasibly usable in different species or strains at the same time and allows selecting vaccine candidates with possibility of eliciting adaptive immune responses. To achieve these benefits all we need is to have a sequenced genome, a personal computer and core software widely known by the scientific community. These conditions show another advantage of using RV, the low cost. What we agreeably call the core software is a set of tools for identifying well-known motifs such as, for example, SignalP, TMHMM, LipoP, and HMMSEARCH. In the use of core software there is still room for innovation when it determined that the choice can be directed to the identification of vaccine candidates specific to an organism such as in the case of gram-negative (bilayer) or gram positive (monolayer) or also placed according to the type of heuristic for selection of vaccine candidates with specific characteristics. For example, membrane or exported to the extracellular environment [9-12].

The concept of RV was adapted to fit a new reality of widespread availability of genomic data [13]. Instead of doing the research for vaccine targets in a single strain or subspecies of an organism, we can do it simultaneously in dozens of genomes, exploring potential joint antigens or exclusive to multiple genomes [14]. The possibility of having a large number of genomes available to implement RV leads to the emergence of the concept of Pan Genomics RV (PGRV) [8]. PGRV can also apply the concepts of core, extended, and character genomes. The core genome in PGRV is composed of exported genes (genes that transcribes for exported proteins) that are common to all strains, genes that could be candidates for a universal vaccine, while the extended genome consists of genes that are absent in at least one of the strains of the studied species and the character genome consists of genes that are specific to a strain [14]. From the standpoint of vaccine, the core and character genomes would be good candidates to compose a vaccine that is suitable for all strains studied, without losing sight of the particularities of specific genes in each strain.

[II] SYSTEMATICAL ANALYZES OF VARIABLES

Considering the motion that many studies using RV are yet to produce effective vaccines [15], an evidence that the limiting factors of RV still have considerable strength despite the enormous advances in genome sequencing has been created herein. Such limiting factors are insignificant amount of currently known antigens and the RV inability to detect non-protein antigens as polysaccharides and glycolipids [16]. These major drawbacks could be minimized with introduction of glycomics and lipidomics studies combined with genomics, proteomics, and peptidomics approaches in vaccine research that would culminate in knowledge and discovery of a wider range of antigens for *in silico* comparisons as new antigens from a survey of RV. More so, core software could also be created to identify patterns in polysaccharides and glycolipids, increasing the repertoire of antigens of an organism.

A limiting factor to the success of RV is the belief that identifying a set of exported proteins is the solution to the lack of production of an effective subunit vaccine against pathogen. Therefore, there are many possibilities of failure and only one chance of success; raising three hypothetical questions in planning a vaccine: **(A)** "Is the set of antigens suitable?" [17], **(E)** "Are antigens expressed in a critical stage of infection?" [10, 18] and (V) "What is the use of a DNA vaccine?" [19-20]. Supposing that initially, each of these three questions could have a TRUE or FALSE answer. In this case we can relate the questions A, E and V into a set of eight (2^3) possibilities, as shown in Table 1. It is the end result that matters, **(R)** "Will the vaccine be effective? The response is "YES" only if the three questions are answered with an assertive TRUE; otherwise the response will invariably be "NO".

Table-1 shows that there are possibilities, as earlier mentioned, of choosing a set of antigens sufficient to confer immunogenicity. In other words, there are chances of choosing a set of antigens effective in conferring immunogenicity, for example, for only one bacterium strain, or the selection of antigens not expressed in an important stage of infection or even the simple act of trying a subunit vaccine instead of DNA vaccine, though the set of selected antigens are adequate and expressed.

The planning of a hypothetical vaccine as shown in **Table-1** still leaves room for doubts by not taking the type of immune response most appropriate to a certain pathogen into consideration. Supposing, for example, a humoral response is not the most suitable for the pathogen of this hypothetical vaccine. Thus, even though **(A)**, **(E)**, and **(V)** are answered as TRUE, yet the vaccine could not induce protective immunity because the most appropriate response lies in the cellular immunity. So after including a fourth question being a variable in **Table-1**, **(C)** "Does vaccine generate immune response?" [21], a set of 16 possibilities was obtained (2^4) among which there are 15 possibilities of failure and only one possibility that matters the most.

This hypothetical example of planning a vaccine in **Table-1** may explain why only the selection of a set of suitable candidates still, leaving a lot of variables that can lead to failure of a vaccine approach. In planning a hypothetical vaccine for these four questions, even if the question **(A)** holds, there still remain seven other possibilities for failure to be adequately answered.

Table: 1. Possible vaccine results considering only three variables: The result shows seven failure possibilities and a record of just one success which matters the most.

No	(A) "Is the Set of antigens suitable?"	(E) "Are antigens expressed in a critical stage of infection?"	(V) "Use of a DNA vaccine?"	(R) "Will the vaccine be effective?"
1	FALSE	FALSE	FALSE	NOT
2	FALSE	FALSE	TRUE	NOT
3	FALSE	TRUE	FALSE	NOT
4	FALSE	TRUE	TRUE	NOT
5	TRUE	FALSE	FALSE	NOT
6	TRUE	FALSE	TRUE	NOT
7	TRUE	TRUE	FALSE	NOT
8	TRUE	TRUE	TRUE	YES

[III] DISCUSSION

The popularization of new technologies of genome sequencing has led to a substantial increase in the number of complete genomes for use in PGRV [14]. Given the particularities of the operating mode of each of various pathogen results and strategies, these can be used in the search for vaccine targets. Below are some of the pathogens for which RV has been used, starting with initial pathogens in the paper described the concept of RV [1] and as a result, we continue with other pathogens which do not necessarily affect humans.

3.1. Tuberculosis (TB)

Despite the prediction of decline in the world TB cases, its incidence continues to grow with more than 10 million cases reported only in 2010, keeping it among the diseases with the highest incidence worldwide [22]. Also, despite the vast amount of research for vaccine against TB, an efficient vaccine against this global scourge is still a promise. The first *Mycobacterium tuberculosis* genome sequence has been released over a decade [23, 24], but still insufficient to bring about a promising vaccine against TB. Considering the availability of complete M. tuberculosis genome sequences, the global urgency of a final solution against the scourge and the facility to conduct *in silico* research, it is inferable that in the search for vaccines understanding the wide range of research involving TB comes easier. A simple search for the term "tuberculosis" in the last three years using the PubMed database generated over 20 thousands published works that are directly or indirectly related to TB. RV was applied over *M. tuberculosis H37Rv* genome aimed at detecting secreted proteins, generating evidence of seven proteins as exoproteome properties that are possible targets for a vaccine [25]. Three secreted proteins belonging to the cutinase-like protein family (Culp) was tested and the Culp6 eliciting a strong cellular response was found [26]. It is the first cellular response recognized in patients affected by TB. These are examples of studies that fit the question **(A)** "Is the set of antigens suitable?" and the last example also characterizes the question **(C)** "Does vaccine generate immune response?" Although many of these studies did not explicitly cite the term RV, many fit the concept and try to get more information about the functional genome released by special attention to exported proteins. Questions of type **(C)** "Does vaccine generate immune response?" from our hypothetical vaccine shall be answered by researches for more effective antigens. The hypothetical protein Rv2626c was found capable of induction of adaptive and humoral immune responses [27]. However, using the concept of epitope density was to create a list of proteins with "hot spots" of the affinity of MHC class II molecules [28]. A hypothetical protein with high affinity to the promoter of genes fbp (Ag85 complex) was the result from search for over expressed factors in proteins from this antigen complex, a protein that belongs to the protein family of transcriptional regulators Mars [29].

Hypothetical membrane proteins were tested and evidenced that Rv0679c protein is expressed in only three strains of *M. tuberculosis*, although 26 strains of bacteria that possessed the gene for this protein were used [18]. Research like this show attempts to answer questions such as **(E)**, from **Table-1** for the planning of a vaccine, being crucially important as much as the question of identification of a secreted protein. Another variable that could be added to the planning table of vaccines would be **(D)** "How low is genetic diversity of selected antigens?" [30]. Included this variable, our universe of possibilities of failure would increase to 31 (2^5 - 1). For example, it was showed that classical vaccine candidates like genes such as esx, fbpB and Esat-6 would not be affected by genetic diversity in 88 strains of *M. tuberculosis* [31]. In this case the answer is "TRUE", increasing the chances of these candidates in our hypothetical screening of candidates. Under the aspect of PGRV, a set of character genes of *M. tuberculosis H37Rv* characterized as important for invasion and survival of the pathogen in the host was found by Al-Attiyah *et al*. 2010 [21]. Among these genes RD1504 was able to induce a strong immune response T-helper (Th) type 1 and may be an important candidate for vaccine target.

3.2. Group B meningococcus

With rates of 16.9/100,000 for bacterial meningitis and 8.9/100,000 for *Neisseria meningitidis* and high number of fatalities in children, meningococcal disease remains a concern and compounded when considering the short period between infection and death, which can possibly be only one day [32]. The experience gained by *in silico* research for candidate vaccine against Men B [33] was a major factor that led to the creation of the term RV. In this work most of the antigens selected in silico and successfully expressed in *Escherichia coli* were exported proteins, including lipoproteins, OMP's, periplasm and membrane proteins. A list of five selected antigens of this study were tested with the adjuvant aluminum hydroxide, CpG oligonucleotides or MF59, achieving antibodies against more than 90% of 85 strains of meningococci representative of the global population diversity [7]. Research by adjuvants can also be included as a requirement to produce an effective vaccine that could be an additional variable in our planning of a hypothetical vaccine **[Table-1]**. It was showed that the amount of factor H, an important regulator of the complement pathway, is correlated with the level of expression of GNA1870, suggesting the inclusion of this protein in the set of antigens of Men B [34]. This research is useful in trying to answer questions of type **(E)** "Are antigens expressed in a critical stage of infection?" [35]. Although there are two vaccines in development for incorporating the Men B protein named Factor H-binding protein or fHbp [36]; it has not been possible to produce a comprehensive vaccine based on this antigen due to its wide antigenic variety [30, 37]. This variety motivated the establishment of a nomenclature to categorize this diversity [38], a study that answers questions such as **(D)** "How low is genetic diversity of selected antigens?". The discovery that convalescent patients develop long-term protective immunity against *N. meningitidis* motivated the search for antigens capable of eliciting such immune response [15]. Contrary to the RV concept, most of the antigens were found cytoplasmic and were not able to produce a satisfactory immune response in guinea pigs. There is also the protein RplY proven to belong to the cell surface of the pathogen. This result makes it a little more confusing to answer the question **(A)** "Is the set of antigens suitable? ", since most of candidates would not be exported. After a decade of the first results of RV on *N. meningitidis*, suggested antigens continue to be researched. It was discovered recently that GNA2132 known as a protein capable of inducing a bactericidal antibody in mice, is also capable of inducing protective immunity in humans. This protein is recognized by serum of convalescent patients, and has been renamed Neisserial Heparin Binding Antigen (NHBA), which is one of the most promising in the search for a vaccine against the pathogen [39] and helping to answer questions of type (C) "Does vaccine generate cellular immune response?".

3.3. *Staphylococcus aureus*

Staphylococcus aureus, a gram positive bacterium remains one of the major human pathogens and a major cause of nosocomial infections worldwide. Failure of antibiotic therapy to eradicate infection is frequently described in literatures and the rate of resistance to clinically relevant antibiotics, such as methicillin, is increasing. Furthermore, there has been an increase in the number of methicillin-resistant S. aureus community-acquired infections [40]. The high prevalence of infections is confounded by the ability of the pathogen to readily acquire genetic elements that confer resistance to antibiotics [41]. The first *S. aureus* complete genome was available on the Gene Bank databases and the Broad Institute since 2007 [42], followed by other 14 different strains at NCBI. There is need for decoding the sequences of complete genome of *S. aureus* that could offer the possibility for comprehensive screening to identify the targets for effective vaccine development [43]. So, it could be interesting to try the answer **(D)** "How low is genetic diversity of selected antigens?" when considering vaccine candidates. Clinical trials with monovalent traditional vaccines already failed to protect against the disease. Now the need is to shift from monovalent vaccine development towards the potential use of multivalent formulations, therapeutic antibodies, and more systematic and rapid identification of optimal antigens by applying *in silico* tools [44]. The RV concept is most suitable for the *S. aureus* to meet the research needs and there are case studies applying it. For example, there are at least 153 individual antigens characterized with the immunome of *S. aureus* [45], despite their subcelular location, which help to answer the question of type **(C)** "Does vaccine generate immune response?". Antibody responses produced against those antigens are accessible to B cells in vivo, most likely extracellular and cell wall-associated proteins, but also non-protein antigens, such as wall teichoic acids (WTA) lipoteichoic acids (LTA), and peptido glycans (PGN) [45]. Surface protein antigens IsdA, IsdB, SdrD, and SdrE were tested for its vaccine efficacy in a combination and individually, the serum IgG titers of immunized mice were almost the same [46]. Immunodominant antigen B (IsaB) is a surface protein believed to be a virulence factor, although its biological functions are not well defined. Its nucleic acid-binding activity is being observed. IsaB has greater affinity for dsDNA than it has for ssDNA or RNA, there is need to evaluate and understand the role of IsaB and its nucleic acid-binding activity which are important in establishment and/or progression of *S. aureus* infection [47] and to answer questions like **(A)** and **(C)** from our hypothetical vaccine planning. Immune dominance of extracellular and surface-exposed proteins has indeed been observed with an *S. aureus* genomic expression library, ribosome display, and 2D-IB. Also most surface associated genes in the core variable genome as well as a large amount of virulence and resistance factors, and many are encoded on mobile genetic elements [45], helping to answer questions like **(A)** and **(D)** from our hypothetical vaccine planning.

3.4. *Corynebacterium diphtheriae*

Corynebacterium diphtheriae bacteria are responsible for Diphtheria. The pathogen produces a toxin that can harm or destroy body tissues and organs, diphtheria toxin (DT). The disease primarily affects mucous membranes of the respiratory tract (Respiratory Diphtheria), although it can also affect the skin (Cutaneous Diphtheria). The bacteria were identified for the first time in the 1880's. In the 1890's, the first antitoxin were developed, and the first vaccine in 1921 [48]. The vaccine is made of inactive forms of the toxin. According to the World Health Organization (WHO), diphtheria affects people of all ages, but it is more frequent in non-immunized child. In 2004, 5000 deaths owing to the disease were reported worldwide [49]. The most effective treatment consists the administration of diphtheria antitoxin (DAT) associated with the elimination of the microorganism using appropriate antibiotics. Since the beginning of the 20th century, many countries produced their own antitoxin preparation from horses. However many factors led to fall of this traditional stocks. Among this factors it is possible to exemplify the lost of economic viability, once the incidence of the disease has fallen a lot, and public objections to the use of horses as donors [50]. In the 1990's, the lack of DAT and vaccines were the two major causes behind the outbreak of diphtheria in the former Union of Soviet Socialist Republics (USSR) states. Between 1990 and 1998, more than 157,000 cases and 5,000 deaths were registered, which represented 80% of diphtheria reports worldwide [51]. This was the major diphtheria epidemic since the 1950's, when the spread of immunization had began. The first genome annotation of *C. diphtheriae* was released in 2003 [52]. With the genome published, Hansmeier *et al.* mapped and analyzed the extracellular and membrane surface of the C7s(-)tox- lineage. This work identified unambiguously ~32% (85/263) of the protein previously described as being extracellular. There were 107 extracellular proteins and 53 of the cell surface, representing a total of 85 different proteins [53]. The importance of this study is to identify secreted proteins, once they can be involved in important interactions between bacteria and host, helping to plan for a vaccine according to the variables in table 1, more specifically question **(A)** "Is the set of antigens suitable?", characterizing a RV research.

Another possible question is **(P)** "Is the antigen in a Pathogenicity Island (PAI)?" [54]. In order to try answer such question for the model *C. diphtheriae*, a comparative genomic hybridization of different strains was made against the specie reference strain [54]. Now, our hypothetical vaccine planning could become more complicated, reaching a total of 63 (2^6 -1) failure possibilities. C7(-) strain was suggested to lack 11 PAIs among 13, while strain PW8, isolated earlier than the C7(-) strain, were lacking only 3 regions related to PAIs.

Additionally, a large genomic diversity among various *C. diphtheriae* strains and clinical isolates were observed. Although the difference between C7(-) strain was higher than the PW8, the adhesion of C7(-) were comparable to the reference strain, while PW8 showed reduce adherence compared to the other strains [54]. This is some odd result considering that adhesions of *C. diphtheriae* to human epithelial cells followed by internalization are signs of pathogenicity.

Searching the literature we can also find studies about the pilli, a very important structure in the adherence of the bacteria and the host [55, 56]. These structures are often involved in the initial adhesion of the bacteria to host tissues during colonization and so helps answer the question **(A)** "Is the set of antigens suitable?" Besides having a structural importance, there are some reports in the literature stating that it might be an important vaccine target, as it is critical in the invasion process [57, 58]. This also helps answer the question **(E)** "Are antigens expressed in a critical stage of infection?"

Compared to the research of vaccine targets of other pathogens in this review, the research on targets of *C. diphtheriae* is less intensive. This apparent calm may be associated with the impression that diphtheria is controlled and that existing vaccines, mostly based on only one antigen, is sufficient to control the disease. However, a feature that makes the development of a vaccine against the bacteria very interesting is the raise of the nontoxigenic strains. Although they don't release the DT in the organism, they are capable of causing morbidity and death [59]. They were isolated from injection drug users in Switzerland [60], homeless alcoholics in France [61], and from poor populations of Vancouver, Canada [59]. In addition, an increasing proportion of strains isolated in the United Kingdom are nontoxigenic [62]. For such non toxic *C. diphtheriae* strains it is possible that the search for adequate variables also can make the difference between a vaccine success or failure.

[V] CONCLUSION

Despite the extensive use of the initial concept of RV and advanced rated results in the search for vaccines against certain pathogens, in general its best and most practical results are still expected. This conclusion is based on fact that the genome sequences of some of the major human pathogens are known longer than the average time required for RV application and promising vaccine against these pathogens still seems far away. Also despite the limitations of RV, this is a low cost technique, fully feasible of use into the plethora of genomic data being generated. It is justifiable to use it in a broader range of pathogens. It is possible that for some of the major human and animal pathogens we can find an appropriate combination of antigens enabling the creation of effective vaccines capable of improving the people's quality of live directly through prevention of diseases or indirectly by improving the economic conditions dependent on breeding. However, as shown in the hypothetical example of planning a vaccine, the discovery of suitable antigens could be a small part of the problem of producing an effective vaccine, but never the less important.

FINANCIAL DISCLOSURE

This work was carried out with financial support from development agencies CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico), FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), CAPES (Coordenação de aperfeiçoamento de Pessoal de Nível Superior)

ACKNOWLEDGEMENT

The authors thank Dr. Sergio Costa Oliveira for fruitful discussion and comments.

REFERENCES

[1] Rappuoli R. [2000] Reverse vaccinology. *Curr Opin Microbiol* 3:445-450.

[2] Sibbald MJJB, van Dij JML. [2009] Secretome Mapping in Gram-Positive Pathogens. In Karl Wooldridge (ed.), Bacterial Secreted Protein: Secretory Mechanisms and Role in Pathogenesis. *Caister Academic Press* :193-225.

[3] Simeone R, Bottai D, Brosch R, et al. [2009] ESX/type VII secretion systems and their role in host-pathogen interaction. *Curr Opin Microbiol* 12:4-10.

[4] Stavrinides J, McCann HC, Guttman DS, et al. [2008] Host-pathogen interplay and the evolution of bacterial effectors. *Cell Microbiol* 10:285-292.

[5] Bhavsar AP, Guttman JA, Finlay BB, et al. [2007] Manipulation of host-cell pathways by bacterial pathogens. *Nature* 449:827-834.

[6] Diaz Romero J, Outschoorn IM. [1994] Current status of meningococcal group B vaccine candidates: capsular or noncapsular? *Clin Microbiol Rev* 7:559-575.

[7] Giuliani MM, Adu-Bobie J, Comanducci M, et al. [2006] A universal vaccine for serogroup B meningococcus. *Proc Natl Acad Sci U S A* 103:10834-10839.

[8] Bambini S, Rappuoli R. [2009] The use of genomics in microbial vaccine development. *Drug Discov Today* 14:252-260.

[9] Barinov A, Loux V, Hammani A, et al. [2009] Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. *Proteomics* 9:61-73.

[10] Yang H, Zhu Y, Qin J, et al. [2006] In silico and microarray-based genomic approaches to identifying potential vaccine candidates against *Leptospira interrogans*. *BMC Genomics* 7:293.

[11] Taylor PD, Attwood TK, Flower DR, et al. [2006] Combining algorithms to predict bacterial protein sub-cellular location: Parallel versus concurrent implementations. *Bioinformation* 1:285-289.

[12] Taylor PD, Toseland CP, Attwood TK, et al. [2006] TATPred: a Bayesian method for the identification of twin arginine translocation pathway signal sequences. *Bioinformation* 1:184-187.

[13] Rinaudo CD, Telford JL, Rappuoli R, et al. [2009] Vaccinology in the genome era. *J Clin Invest* 119:2515-2525.

[14] Lapierre P, Gogarten JP, [2009] Estimating the size of the bacterial pan-genome. *Trends Genet* 25:107-110.

[15] Mendum TA, Newcombe J, McNeilly CL, et al. [2009] Towards the Immunoproteome of *Neisseria meningitidis PLoS ONE* 4:.

[16] Rappuoli R. [2001] Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine* 19:2688-2691.

[17] Choi G, Eom S, Jung K, et al. [2010] CysA2: A candidate serodiagnostic marker for *Mycobacterium tuberculosis* infection. *Respirology* 15:636-642.

[18] Cifuentes DP, Ocampo M, Curtidor H, et al. [2010] Mycobacterium tuberculosis Rv0679c protein sequences involved in host-cell infection: potential TB vaccine candidate antigen. *BMC Microbiol* 10:109.

[19] Gat O, Grosfeld H, Ariel N, et al. [2006] Search for *Bacillus anthracis* Potential Vaccine Candidates by a Functional Genomic-Serologic Screen. *Infect Immun* 74:3987-4001.

[20] Dunachie SJ, Hill AVS, [2003] Prime-boost strategies for malaria vaccine development. *J Exp Biol* 206:3771-3779.

[21] Al-Attiyah R, Mustafa AS. [2010] Characterization of human cellular immune responses to *Mycobacterium tuberculosis* proteins encoded by genes predicted in RD15 genomic region that is absent in *Mycobacterium bovis* BCG. *FEMS Immunol Med Microbiol* 59:177-187.

[22] Dye C, Williams BG. [2010] The Population Dynamics and Control of Tuberculosis. *Science* 328:856-861.

[23] Cole ST, Brosch R, Parkhill J, et al. [1998] Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* 393:537-544.

[24] Cole ST, Barrell BG. [1998] Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. *Novartis Found Symp* 217:160-72; discussion 172-7.

[25] Vizcaíno C, Restrepo-Montoya D, Rodríguez D, et al. [2010] Computational prediction and experimental assessment of secreted/surface proteins from *Mycobacterium tuberculosis H37Rv*. *PLoS Comput Biol* 6:e1000824.

[26] Shanahan ER, Pinto R, Triccas JA, et al. [2010] Cutinase-like protein-6 of *Mycobacterium tuberculosis* is recognised in tuberculosis patients and protects mice against pulmonary infection as a single and fusion protein vaccine. *Vaccine* 28:1341-1346.

[27] Bashir N, Kounsar F, Mukhopadhyay S, et al. [2010] *Mycobacterium tuberculosis* conserved hypothetical protein rRv2626c modulates macrophage effector functions. *Immunology* 130:34-45.

[28] Gaseitsiwe S, Valentini D, Mahdavifar S, et al. [2010] Peptide microarray-based identification of *Mycobacterium tuberculosis* epitope binding to HLA-DRB1*0101, DRB1*1501, and DRB1*0401. *Clin Vaccine Immunol* 17:168-175.

[29] Romero IC, Mehaffy C, Burchmore RJ, et al. [2010] Identification of promoter-binding proteins of the fbp A and C genes in *Mycobacterium tuberculosis Tuberculosis (Edinb)* 90:25-30.

[30] Nash JH, Findlay WA, Luebbert CC, et al. [2006] Comparative genomics profiling of clinical isolates of *Aeromonas salmonicida* using DNA microarrays. *BMC Genomics* 7:43.

[31] Davila J, Zhang L, Marrs CF, et al. [2010] Assessment of the genetic diversity of *Mycobacterium tuberculosis* esxA, esxH, and fbpB genes among clinical isolates and its implication for the future immunization by new tuberculosis subunit vaccines Ag85B-ESAT-6 and Ag85B-TB10.4. *J Biomed Biotechnol* 2010:208371.

[32] Theodoridou MN, Vasilopoulou VA, Atsali EE, et al. [2007] Meningitis registry of hospitalized cases in children: epidemiological patterns of acute bacterial meningitis throughout a 32-year period. *BMC Infect Dis* 7:101.

[33] Pizza M, Scarlato V, Masignani V, et al. [2000] Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287:1816-1820.

[34] Madico G, Welsch JA, Lewis LA, et al. [2006] The Meningococcal Vaccine Candidate GNA1870 Binds the Complement Regulatory Protein Factor H and Enhances Serum Resistance. *J Immunol* 177:501-510.

[35] Montor WR, Huang J, Hu Y, et al. [2009] Genome-Wide Study of *Pseudomonas aeruginosa* Outer Membrane Protein Immunogenicity Using Self-Assembling Protein Microarrays▿. *Infect Immun* 77:4877-4886.

[36] Beernink PT, Welsch JA, Harrison LH, et al. [2007] Prevalence of Factor H–Binding Protein Variants and NadA among Meningococcal Group B Isolates from the United States: Implications for the Development of a Multicomponent Group B Vaccine. *J Infect Dis* 195:1472-1479.

[37] Lipsitch M, O'Hagan JJ. [2007] Patterns of antigenic diversity and the mechanisms that maintain them. *J R Soc Interface* 4:787-802.

[38] Brehony C, Wilson DJ, Maiden MC, et al. [2009] Variation of the factor H-binding protein of *Neisseria meningitidis*. *Microbiology* 155:4155-4169.

[39] Serruto D, Spadafina T, Ciucchi L, et al. [2010] Neisseria meningitidis GNA2132, a heparin-binding protein that induces protective immunity in humans. *Proc Natl Acad Sci U S A* 107:3770-3775.

[40] Kuklin NA, Clark DJ, Secore S, et al. [2006] A novel Staphylococcus aureus vaccine: iron surface determinant B induces rapid antibody responses in rhesus macaques and specific increased survival in a murine *S. aureus* sepsis model. *Infect Immun* 74:2215-2223.

[41] Ventura CL, Malachowa N, Hammer CH, et al. [2010] Identification of a novel *Staphylococcus aureus* two-component leukotoxin using cell surface proteomics. *PLoS ONE* 5:e11634.

[42] Herron-Olson L, Fitzgerald JR, Musser JM, et al. [2007] Molecular correlates of host specialization in *Staphylococcus aureus*. *PLoS ONE* 2:e1120.

[43] McCarthy AJ, Lindsay JA. [2010] Genetic variation in *Staphylococcus aureus* surface and immune evasion genes is lineage associated: implications for vaccine design and host-pathogen interactions. *BMC Microbiol* 10:173.

[44] Otto M. [2010] Novel targeted immunotherapy approaches for staphylococcal infection. *Expert Opin Biol Ther* 10:1049-1059.

[45] Holtfreter S, Kolata J, Bröker BM, et al. [2010] Towards the immune proteome of *Staphylococcus aureus* - The anti-S. aureus antibody response. *Int J Med Microbiol* 300:176-192.

[46] Stranger-Jones YK, Bae T, Schneewind O, et al. [2006] Vaccine assembly from surface proteins of *Staphylococcus aureus*. *Proc Natl Acad Sci U S A* 103:16942-16947.

[47] Mackey-Lawrence NM, Potter DE, Cerca N, et al. [2009] *Staphylococcus aureus* immunodominant surface antigen B is a cell-surface associated nucleic acid binding protein. *BMC Microbiol* 9:61.

[48] Agnew J. [2010] Medicine in the Old West: A History, 1850-1900. McFarland

[49] World Healt Organization [2010] Diphtheria [http://www.who.int/immunization/topics/diphtheria/en/index.html]

[50] Wagner K, Stickings P, White J, et al. [2009] A review of the international issues surrounding the availability and demand for diphtheria antitoxin for therapeutic use. *Vaccine* 28:14 - 20.

[51] Dittmann S, Wharton M, Vitek C, et al. [2000] Successful control of epidemic diphtheria in the states of the Former Union of Soviet Socialist Republics: lessons learned. *J Infect Dis* 181 Suppl 1:S10-22.

[52] Cerdeño-Tárraga AM, Efstratiou A, Dover LG, et al. [2003] The complete genome sequence and analysis of *Corynebacterium diphtheriae NCTC13129*. *Nucleic Acids Res* 31:6516-6523.

[53] Hansmeier N, Chao T, Kalinowski J, et al. [2006] Mapping and comprehensive analysis of the extracellular and cell surface proteome of the human pathogen *Corynebacterium diphtheriae*. Proteomics 6:2465-2476.

[54] Iwaki M, Komiya T, Yamamoto A, et al. [2010] Genome organization and pathogenicity of *Corynebacterium diphtheriae* C7(-) and PW8 strains. *Infect Immun* 78:3791-3800.

[55] Kang HJ, Paterson NG, Gaspar AH, et al. [2009] The *Corynebacterium diphtheriae* shaft pilin SpaA is built of tandem Ig-like modules with stabilizing isopeptide and disulfide bonds. *Proc Natl Acad Sci U S A* 106:16967-16971.

[56] Mandlik A, Swierczynski A, Das A, et al. [2007] *Corynebacterium diphtheriae* employs specific minor pilins to target human pharyngeal epithelial cells. *Mol Microbiol* 64:111-124.

[57] Telford JL, Barocchi MA, Margarit I, et al. [2006] Pili in gram-positive pathogens. *Nat Rev Microbiol* 4:509-519.

[58] Proft T, Baker EN. [2009] Pili in Gram-negative and Gram-positive bacteria - structure, assembly and their role in disease. *Cell Mol Life Sci* 66:613-635.

[59] Romney MG, Roscoe DL, Bernard K, et al. [2006] Emergence of an invasive clone of nontoxigenic *Corynebacterium diphtheriae* in the urban poor population of Vancouver, Canada. *J Clin Microbiol* 44:1625-1629.

[60] Gubler J, Huber-Schneider C, Gruner E, et al. [1998] An outbreak of nontoxigenic *Corynebacterium diphtheriae* infection: single bacterial clone causing invasive infection among Swiss drug users. *Clin Infect Dis* 27:1295-1298.

[61] Patey O, Bimet F, Riegel P, et al. [1997] Clinical and molecular study of *Corynebacterium diphtheriae* systemic infections in France. Coryne Study Group. *J Clin Microbiol* 35:441-445.

[62] Health Protection Agency. [2006] Diphtheria Notifications and Deaths: England and Wales 1986–2006. [http://www.hpa.org.uk/Topics/InfectiousDiseases/InfectionsAZ/Diphtheria/EpidemiologicalData/]

IIOAB ISSN: 0976-3104

ABOUT AUTHORS

Anderson Rodrigues Santos; *MSc, has a degree in Computer Science from Catholic University of Minas Gerais (1995) and MSc in Computer Science from Universidade Federal de Minas Gerais (1999). He has experience in computer science, with emphasis on Logic Programming. Possibly as a guest professor at several graduate courses as Electrical Engineering, C. Accounting, C. Administrative and Digital Games. Started a PhD in Bioinformatics in March 2008. Outstanding performance in the assembly and annotation of the first fully assembled and annotated genome in the state of Minas Gerais for Genomic Network of Minas Gerais on bacteria Corynebacterium pseudotuberculosis 1002, which infects goats. By LGCM and Genomic Network of Pará participated specifically in automatic and manual annotation of six other strains of the same bacteria (C231, I19, PAT10, 162, 258 and CIP5297) through the use of databases management systems and compilers*

Amjad Ali; *BS (Hons) MPhil in Biotechnology/Genetics (2008), PhD student in Genetics with special focus on Genetics of Microorganisms: assembly and annotation of genomes, pan genomics, comparative genomics and pathogenomics analysis of different Corynebacterium psuedotuberculosis strains, for the development of Vaccines against caseous Lymphadenitis (CLA) in sheep and Goats.*

Eudes Guilherme Vieira Barbosa; *is in the final year of Biology in Federal University of Minas Gerais and is an undergraduate research in the Genetic Department of the Instituto de Ciências Biológicas. Has experience in Genetics, focusing on Molecular Genetics and of Microorganisms. Currently is working in the assembly and annotation of Corynebacterium pseudotuberculosis strains.*

Prof. Artur Luiz da Costa da Silva*; MSc, PhD, has a master and PH.D decrees in Genetics and Molecular Biology from the Federal University of Pará (UFPA). Since 1996 he is a professor in UFPA and he one of the coordinators of the Genomic and Proteomic Web of Pará. He is a level 2 CNPq researcher and is affiliated to the Brazilian Academy of Science.*

Prof. Anderson Miyoshi; *M.Sc, PhD, is adjunt professor of the General Biology Departmant in Federal University of Minas Gerais. His research interests are in Molecular Genetics and of Microorganisms. Currently is working in the development of new genomic expression systems. His total research publications are 35 and 5 book chapters.*

Prof. Vasco Azevedo; *DVM, M.Sc, PhD, FESC, is full Professor of Federal University of Minas Gerais. Professor Azevedo is a pioneer of genetics of Lactic Acid Bacteria and Corynebacterium pseudotuberculosis in Brazil. He has specialized in bacterial genetics, genome, transcriptome, proteome, developement of new vaccines and diagnostic against infectious diseases. He is the Associate Editor of Genetics and Molecular Research, and member of editorial board of Open Veterinary Science Journal and Internacional Journal of Microbiology. His total number of research publications is 103 and has authored 11 book chapters.*

1.3 Vacinologia reversa pangenômica

Um pangenoma caracteriza genes comuns a diversas linhagens de uma espécie ou a diversas espécies (Bambini e Rappuoli, 2009). Esse conjunto de genes comum é conhecido como genoma central. Genes presentes em mais de um genoma, mas não em todos os genomas sob estudo, compõem o genoma acessório, enquanto genes exclusivos de um genoma constituem o genoma único ou linhagem específica. Tettelin e cols. (2005) deram ênfase na diversidade intra-espécies por um estudo baseado na comparação do genoma de diferentes linhagens de *Streptococcus agalactiae* (*Group B Streptococcus* ou *GBS*) representante da diversidade genética das espécies. Os resultados mostraram que um pangenoma consistindo dos genomas central e acessório pode ser maior do que o genoma de uma única linhagem, dado que o número de genes únicos de cada linhagem foi maior do que o esperado.

Novos genes continuam a ser adicionados ao repertório genético das espécies toda vez que uma nova linhagem é sequenciada. Até Fevereiro de 2012 existiam mais de 1.860 genomas bacterianos completos publicados no banco de dados do *National Center for Biotechnology Information* (NCBI) conforme mostrado no sítio dessa instituição (http://www.ncbi.nlm.nih.gov/sites/genome).

O sítio *Genomes On-line Database* (GOLD) v3.0 (http://www.genomesonline.org), é referência mundial para armazenamento de genomas completos ou em fase de rascunho. Armazena também dados que caracterizam os genomas, também chamados de metadados, permitindo buscas por genomas com base em características documentadas. Desde seu surgimento em 1999 o GOLD tem recebido um número crescente de projetos genoma e atualmente a maior quantidade de projetos submetidos é de organismos bacterianos. Até outubro de 2011, conforme exibido na Figura 5, existiam 4812 projetos de genomas bacterianos cadastrados no GOLD, sendo que dentre esses aproximadamente 1200 eram de genomas bacterianos finalizados.

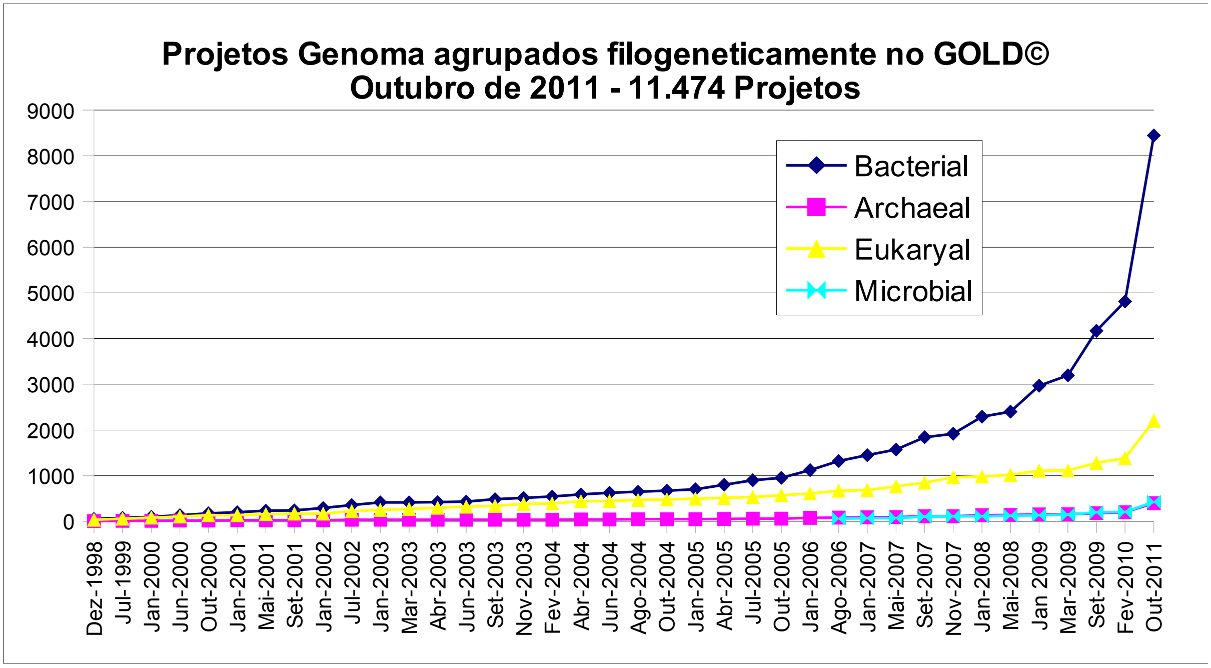


Figura 5: Projetos genoma cadastrados no GOLD.

(Fonte: http://www.genomesonline.org/)

Desse total de 8448 projetos de genomas bacterianos, 11% ou ~930 são de genomas de actinobactérias, conforme exibido pela Figura 6. Entre exemplos de bactérias pertencentes a esse filo podem ser citadas a *M. tuberculosis* agente etiológico da tuberculose (TB) em humanos, a *C. diphtheriae* causadora da difteria também em humanos e a *C. pseudotuberculosis*, causadora da linfadenite caseosa (LC) em caprinos e ovinos.

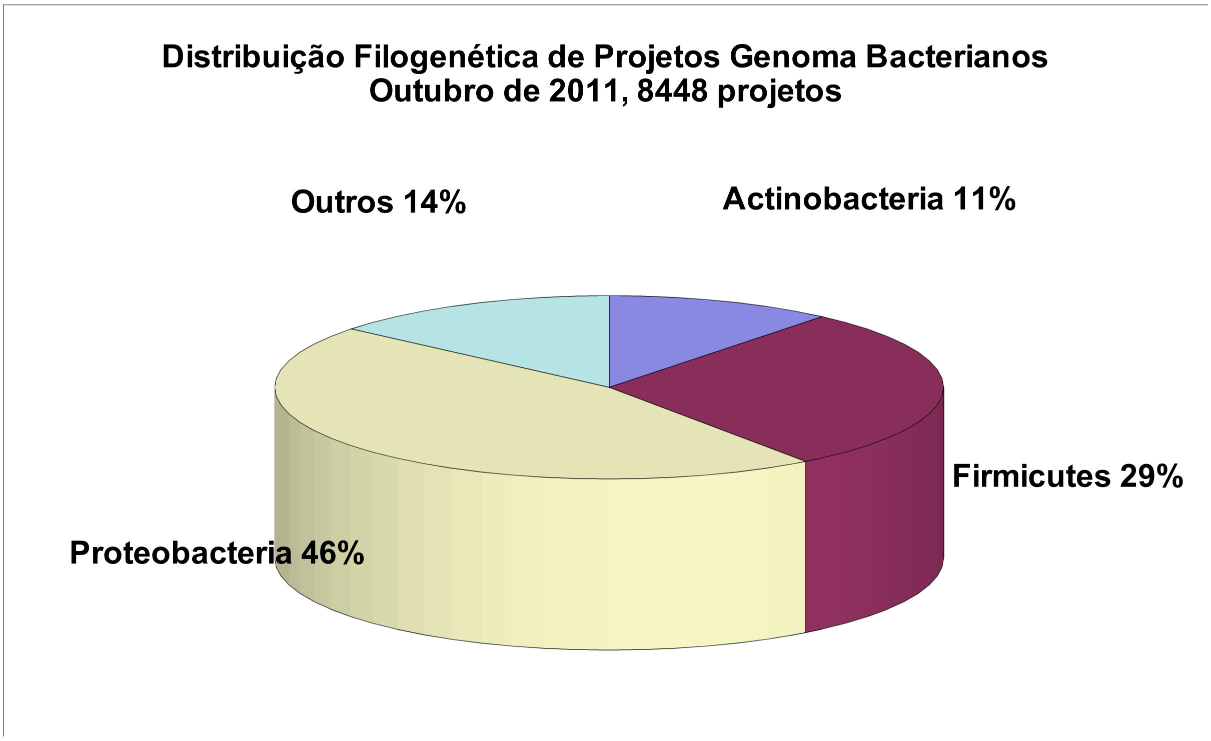


Figura 6: Projetos de filogenética de genomas bacterianos no GOLD.

(Fonte: http://www.genomesonline.org/)

Binnewies e cols. (2006) mostraram em análises realizadas sobre centenas de genomas bacterianos que existe grande diversidade genotípica entre todas as espécies, diversidade esta que é encontrada até mesmo em genomas de uma mesma espécie de bactéria quando várias linhagens são comparadas. Esta diversidade é gerada por uma variedade de mecanismos como, por exemplo, a transferência horizontal de genes, incluindo os elementos genéticos móveis como transposons e bacteriófagos (Binnewies e cols., 2006), mecanismos também abordados na seção Erro: Origem da referência não encontrada. O conceito de que a diversidade genética intra-espécies pode ser tão significativa quanto entre espécies foi revelado por estudos de hibridação subtrativa e hibridação comparativa do genoma (CGH), em que isolados pertencentes à mesma espécie bacteriana foram analisados por estudos de microarranjos utilizando uma linhagem sequenciada como referência, por exemplo, *Campylobacter jejuni* (Dorrell e cols., 2001), *Escherichia coli* e *Shigella flexneri* (SF), *S. boydii* (SB) e *S. sonnei* (SS) (Fukiya e cols., 2004). Esses estudos mostraram que há genes que não estão conservados em todas as linhagens da mesma espécie e que há uma ampla diversidade genética. Experimentos de hibridação comparativa do genoma, no entanto, não são capazes de identificar os genes ausentes no genoma de referência (Earl e cols., 2007).

O rápido desenvolvimento das tecnologias de sequenciamento, fornecendo sequências genômicas completas de diferentes isolados disponíveis para análises comparativas, tem impulsionado o uso da genômica para investigação da variabilidade dentro de uma única espécie. Enquanto no início da era dos genomas as sequências disponíveis pertenciam principalmente, às representantes de linhagens patogênicas de espécies diferentes. A disponibilidade de uma quantidade enorme de genomas sequenciados estimularam o desenvolvimento da genômica comparativa, que realiza comparações de espécies diferentes e de linhagens diferentes de uma mesma espécie fornecendo novos pontos de vista sobre a biologia preditiva (Gay e cols., 2007). Pesquisas de genômica comparativa e pós-genômica, baseadas em ferramentas de bioinformática, tecnologia de microarranjo e transcriptoma utilizando tecnologia de sequenciamento de próxima geração, aproveitam a comparação das sequências do genoma para identificarem determinantes de virulência, drogas e alvos vacinais contra doenças infecciosas. Por exemplo, a comparação de sequências do genoma de bactérias intimamente relacionadas à virulência e bactérias não virulentas pode ajudar na identificação de padrões genéticos (Ruiz e cols., 2011).

Além de estudos de diversidade genotípica, uma possível aplicação do pangenoma é a identificação de novas vacinas e alvos antimicrobianos (Vacinologia Reversa Pangenômica). A primeira aplicação do pangenoma em vacinas foi feita por Maione e colegas usando a bactéria estreptococo do grupo B (Maione e cols., 2005). O foco foi sobre 589 proteínas da superfície celular, alvos obtidos por abordagem computacional, das quais 396 eram pertencentes ao genoma central e o restante eram genes ausentes em pelo menos uma linhagem. Possíveis antígenos, na forma de proteínas recombinantes, foram expressos, purificados e testados como vacinas de subunidades, e quatro conferiram imunidade protetora em um modelo animal. Entre estes antígenos, apenas um era parte do genoma central, porém não foi capaz de conferir proteção global, daí a formulação da vacina final necessitou incluir uma combinação de outros três antígenos.

O exemplo do estreptococo do grupo B tem demonstrado que um genoma com múltiplas sequências de cada espécie ou linhagem é importante para abranger a diversidade genômica de muitos patógenos. No caso das espécies altamente diferenciadas, a variabilidade genotípica dos patógenos pode ser um problema para o desenvolvimento de vacinas baseadas em proteínas, o que deve ser concebida para cobrir um amplo painel de linhagens (vacinologia populacional). Para este propósito, pode ser fundamental a contribuição dos estudos de epidemiologia molecular visando selecionar linhagens representativas da variabilidade genotípica mundial de um determinado microrganismo.

1.4 Imunoinformática

Uma lista de genes obtidos por meio do uso da VR, em um genoma bacteriano, pode facilmente alcançar o montante de algumas centenas de possíveis candidatos a alvos vacinais. Essa quantia considerável de candidatos para serem validados *in vitro* demanda muitos recursos financeiros e experimentais para validação. Uma solução para minimizar a quantidade de alvos vacinais da etapa experimental está na imunoinformática. Trata-se de mais um filtro computacional para tentar identificar candidatos vacinais que seriam mais propensos a induzir uma resposta imune e protetora no organismo de um hospedeiro de um certo agente etiológico (Lundegaard e cols., 2008; Larsen e cols., 2007). A imunoinformática tem como foco converter dados oriundos das novas tecnologias de sequenciamento em problemas imunológicos que possam ser analisados *in silico*; resolver estes problemas com abordagens matemáticas e converter os resultados em interpretações com significado imunológico (Kleinstein, 2008). Dentre esses problemas está a busca de peptídeos que seriam ligantes às células apresentadoras de antígenos, mais especificamente às moléculas *Major Histocompatibility Complex* (MHC) classes I e II. Os peptídeos ligantes a essas classes se diferenciam basicamente pelo tamanho em aminoácidos e por se ligarem com conformação tridimensional ou linear. Peptídeos ligantes ao MHC I variam em tamanho de 8 a 11 aminoácidos e não precisam assumir uma conformação tridimensional para se ligarem ao sítio de ligação da molécula apresentadora de antígenos. Por esse motivo peptídeos ligantes ao MHC I também são denominados lineares. Peptídeos ligantes ao MHC II denominados não lineares ou conformacionais assumem uma conformação tridimensional para se ligarem à célula apresentadora de antígenos. São maiores, estando geralmente em uma faixa de 13 até 17 aminoácidos. Essas duas diferenças básicas entre peptídeos lineares ligantes aos MHC de classe I e não lineares ligantes ao MHC de classe II definem o nível de dificuldade de mapeá-los por meios inteiramente computacionais, sendo mais fácil e preciso identificar peptídeos lineares (Lundegaard e cols., 2008).

Os peptídeos ligantes ao MHC são denominados epitopos sendo complementares ao sítio de ligação da célula apresentadora denominado paratopo. Epitopos ligantes ao MCH I são indicativos da possibilidade de uma resposta celular, conferindo imunidade adaptativa ao organismo infectado por um agente etiológico por meio da ativação de linfócitos T CD8+, enquanto epitopos ligantes ao MHC II são indicativos da possibilidade de uma resposta humoral por meio da ativação de linfócitos T CD4+ ou “*helper*” (Larsen e cols., 2007).

Dessa forma, uma abordagem utilizada muito comum na imunoinformática é a busca por proteínas que possuem em sua composição sequências lineares ou não lineares de peptídeos para os quais existem melhores chances em criar uma resposta imune.

2. Justificativa e objetivos

2.1 Justificativa

C. pseudotuberculosis é um patógeno Gram-positivo, intracelular facultativo e causador da doença conhecida como linfadenite caseosa (LC). Esta doença ocorre principalmente em ovinos e caprinos, no entanto, o patógeno tem a capacidade de infectar outros hospedeiros, como camelídeos, equinos, bovinos, bubalinos e, em casos raros, os humanos. A distribuição mundial desta doença, seu impacto no agronegócio, a ausência de um tratamento eficaz e de vacinas, levam a estudos mais aprofundados desse organismo.

O presente trabalho faz parte do projeto de Patogenômica da *C. pseudotuberculosis* que está sendo desenvolvido pelo nosso grupo de pesquisa. A Patogenômica é definida como a análise ao nível genômico dos processos envolvidos na patogênese bacteriana causada pela interação dos micróbios patogênicos e seus hospedeiros. Para realização de análises ao nível genômico, surgiu a necessidade de montar, anotar e analisar resultados *in silico* de diversos genomas de *C. pseudotuberculosis.* Os resultados dessas análises podem, por exemplo, propiciar a criação de listas de possíveis candidatos para vacinais, drogas e diagnósticos. A busca por alvos para vacinas, diagnóstico e drogas foi motivada pelos resultados obtidos com o organismo *Neisseria meningitidis* do sorogrupo B (MenB) (Maione e cols., 2005). Essa pesquisa obteve ótimos resultados, apenas cinco proteínas conferiram imunidade protetora contra 22 linhagens do patógeno. Entretanto a lista inicial de candidatos *in silico* era composta de quase quatrocentos genes, situação que gerou um grande esforço de validação experimental. No caso específico do nosso grupo de pesquisa, as publicações referentes ao exoproteoma (Pacheco e cols., 2011; Silva e cols., 2012) foram utilizados como mais um instrumento associado à vacinologia reversa.

2.2 Objetivo geral

Sequenciar, montar, anotar e analisar os genomas de *C. pseudotuberculosis* isolados de diferentes hospedeiros para identificar proteínas com potencial para o desenvolvimento de diagnósticos, vacinas e drogas contra LC.

2.3 Objetivos específicos

1) Auxiliar na montagem, anotação e homogeneização dos genomas de 15 linhagens da *C. pseudotuberculosis*;

2) Criar um banco de dados relacional, programas de entrada de dados e relatórios para análises dos genomas da *C. pseudotuberculosis*, denominado CpDB;

3) Documentar e automatizar o processo de transferência automática de anotação entre genomas homólogos, por meio do CpDB;

4) Realizar análises comparativas, predição de proteínas exportadas e imunoinformática nos genomas da *C. pseudotuberculosis*;

5) Combinar predições gerando um conjunto de candidatos para vacinas, diagnóstico e drogas;

6) Desenvolver um *pipeline* computacional automatizando as análises para geração dessas listas de alvos proteicos.

3. Resultados e discussões

3.1 Genômica Estrutural

Os resultados apresentados na seção de genômica estrutural começam pelo primeiro genoma inteiramente montado, anotado e depositado pelo estado de Minas Gerais. Esse projeto, iniciado com a linhagem 1002 da *C. pseudotuberculosis,* foi o primeiro passo do nosso grupo de pesquisa para iniciar o projeto do pan genoma desse organismo que inclui 15 genomas (seção 3.1.6). Na sequência, são apresentadas as publicações dos genomas de três linhagens de *C. pseudotuberculosis*: PAT10, I19 e CIP 52.97, que serão utilizados em análises pangenomicas. O grupo de pesquisa especializou-se na anotação e reanotação de genomas para viabilizar o pan genoma da *C. pseudotuberculosis*. Um dos produtos dessa especialização é a reanotação do genoma da *C. diphtheriae*, cujos resultados mostram a importância de refazer a predição gênica periodicamente utilizando novas ferramentas de software que enriqueçam os dados sobre um genoma. Por último, é apresentado como resultado um banco de dados relacional utilizado em todas as etapas do pan genoma de *C. pseudotuberculosis*. Esse banco de dados é o repositório oficial de todos os genomas anotados pelo nosso grupo de pesquisa e também é utilizado para a realização de análises comparativas, identificação de erros de montagem e anotação.

Nesta seção, o doutorando participou ativamente de equipes para garantir uma boa qualidade dos genomas publicados por meio da curadoria manual, bem como análises de localização subcelular das proteínas preditas (subseções 3.1.1 até 3.1.6). Entretanto, sua maior contribuição está na subseção 3.1.7 que descreve um banco de dados para anotação e análise de genomas bacterianos. O doutorando foi responsável pela conceituação, implementação e gerenciamento do banco de dados relacional e ferramentas que automatizaram a geração de relatórios que, por vez, agilizaram a velocidade de anotação dos genomas. Esses relatórios também propiciaram métodos de inferir sobre a qualidade de um genoma ainda na fases de montagem.

3.1.1 Redução do genoma e aquisição de fatores de virulência como evidências de evolução em duas linhagens de *C. pseudotuberculosis*

Nesse trabalho, caracterizou-se dois genomas *C. pseudotuberculosis*, a linhagem 1002, isolada de caprino e a linhagem C231, isolado de ovino. A análise desses genomas mostrou alta similaridade em relação à arquitetura, conteúdo e ordem. Comparando-se *C. pseudotuberculosis* com outras espécies do mesmo gênero, tornou-se evidente que esta espécie patogênica perdeu numerosos genes, resultando em um dos menores genomas no gênero *Corynebacterium*. Outra diferença que salienta a adaptação à patogenicidade inclui um menor teor de GC, de cerca de 52%. O genoma da *C. pseudotuberculosis* inclui também sete prováveis ilhas de patogenicidade, com fatores de virulência clássicos, incluindo genes para subunidades fimbriais, fatores de adesão, absorção de ferro e toxinas secretadas. Além disso, todos os fatores de virulência em ilhas têm características que indicam a transferência horizontal.

As características particulares do genoma de *C. pseudotuberculosis*, bem como seus fatores de virulência adquiridos em prováveis ilhas de patogenicidade, fornecem evidências sobre suas vias de patogenicidade e o seu estilo de vida no processo de infecção. Os genomas das linhagens 1002 e C231, citados neste estudo, estão disponíveis no banco de dados *GenBank* do NCBI (*GenBank* http://www.ncbi.nlm.nih.gov/ /) sob os identificadores CP001809 e CP001829, respectivamente.

Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains

Jerônimo C. Ruiz[1], Vívian D'Afonseca[2], Artur Silva[3], Amjad Ali[2], Anne C. Pinto[2], Anderson R. Santos[2], Aryanne A. M. C. Rocha[2], Débora O. Lopes[4], Fernanda A. Dorella[2], Luis G. C. Pacheco[2,20], Marcília P. Costa[5], Meritxell Z. Turk[2], Núbia Seyffert[2], Pablo M. R. O. Moraes[2], Siomar C. Soares[2], Sintia S. Almeida[2], Thiago L. P. Castro[2], Vinicius A. C. Abreu[2], Eva Trost[6], Jan Baumbach[7], Andreas Tauch[6], Maria Paula C. Schneider[3], John McCulloch[3], Louise T. Cerdeira[3], Rommel T. J. Ramos[3], Adhemar Zerlotini[1], Anderson Dominitini[1], Daniela M. Resende[1,8], Elisângela M. Coser[1], Luciana M. Oliveira[9], André L. Pedrosa[8,10], Carlos U. Vieira[11], Cláudia T. Guimarães[12], Daniela C. Bartholomeu[13], Diana M. Oliveira[5], Fabrício R. Santos[2], Élida Mara Rabelo[14], Francisco P. Lobo[13], Glória R. Franco[13], Ana Flávia Costa[2], Ieso M. Castro[15], Sílvia Regina Costa Dias[14], Jesus A. Ferro[16], José Miguel Ortega[13], Luciano V. Paiva[17], Luiz R. Goulart[11], Juliana Franco Almeida[11], Maria Inês T. Ferro[16], Newton P. Carneiro[12], Paula R. K. Falcão[18], Priscila Grynberg[13], Santuza M. R. Teixeira[13], Sérgio Brommonschenkel[19], Sérgio C. Oliveira[13], Roberto Meyer[20], Robert J. Moore[21], Anderson Miyoshi[2], Guilherme C. Oliveira[1,22], Vasco Azevedo[2*]

1 Research Center René Rachou, Oswaldo Cruz Foundation, Belo Horizonte, Minas Gerais, Brazil, **2** Department of General Biology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **3** Department of Genetics, Federal University of Pará, Belém, Pará, Brazil, **4** Health Sciences Center, Federal University of São João Del Rei, Divinópilis, Minas Gerais, Brazil, **5** Department of Veterinary Medicine, State University of Ceará, Fortaleza, Ceará, Brazil, **6** Department of Genetics, University of Bielefeld, CeBiTech, Bielefeld, Nordrhein-Westfale, Germany, **7** Department of Computer Science, Max-Planck-Institut für Informatik, Saarbrücken, Saarlan, Germany, **8** Department of Pharmaceutical Sciences, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil, **9** Department of Phisics, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil, **10** Department of Biological Sciences, Federal University of Triangulo Mineiro, Uberaba, Minas Gerais, Brazil, **11** Department of Genetics and Biochemistry, Federal University of Uberlândia, Uberlândia, Minas Gerais, Brazil, **12** Brazilian Agricultural Research Corporation (EMBRAPA), Sete Lagoas, Minas Gerais, Brazil, **13** Department of Biochemistry and Immunology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **14** Department of Parasitology, Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil, **15** Department of Pharmacy, Federal University of Ouro Preto, Ouro Preto, Minas Gerais, Brazil, **16** Department of Technology, State University of São Paulo, Jaboticabal, São Paulo, Brazil, **17** Department of Chemistry, Federal University of Lavras, Lavras, Minas Gerais, Brazil, **18** Brazilian Agricultural Research Corporation (EMBRAPA), Campinas, São Paulo, Brazil, **19** Department of Plant Pathology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil, **20** Department of Biointeraction Sciences, Federal University of Bahia, Salvador, Bahia, Brazil, **21** CSIRO Livestock Industries, Australia, **22** Center of Excellence in Bioinformatics, National Institute of Science and Technology, Research Center René Rachou, Oswaldo Cruz Foundation, Belo Horizonte, Minas Gerais, Brazil

Abstract

Background: *Corynebacterium pseudotuberculosi*s, a Gram-positive, facultative intracellular pathogen, is the etiologic agent of the disease known as caseous lymphadenitis (CL). CL mainly affects small ruminants, such as goats and sheep; it also causes infections in humans, though rarely. This species is distributed worldwide, but it has the most serious economic impact in Oceania, Africa and South America. Although *C. pseudotuberculosis* causes major health and productivity problems for livestock, little is known about the molecular basis of its pathogenicity.

Methodology and Findings: We characterized two *C. pseudotuberculosis* genomes (Cp1002, isolated from goats; and CpC231, isolated from sheep). Analysis of the predicted genomes showed high similarity in genomic architecture, gene content and genetic order. When *C. pseudotuberculosis* was compared with other *Corynebacterium* species, it became evident that this pathogenic species has lost numerous genes, resulting in one of the smallest genomes in the genus. Other differences that could be part of the adaptation to pathogenicity include a lower GC content, of about 52%, and a reduced gene repertoire. The *C. pseudotuberculosis* genome also includes seven putative pathogenicity islands, which contain several classical virulence factors, including genes for fimbrial subunits, adhesion factors, iron uptake and secreted toxins. Additionally, all of the virulence factors in the islands have characteristics that indicate horizontal transfer.

Conclusions: These particular genome characteristics of *C. pseudotuberculosis*, as well as its acquired virulence factors in pathogenicity islands, provide evidence of its lifestyle and of the pathogenicity pathways used by this pathogen in the infection process. All genomes cited in this study are available in the NCBI Genbank database (http://www.ncbi.nlm.nih.gov/genbank/) under accession numbers CP001809 and CP001829.

Citation: Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, et al. (2011) Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains. PLoS ONE 6(4): e18551. doi:10.1371/journal.pone.0018551

Editor: Igor Mokrousov, St. Petersburg Pasteur Institute, Russian Federation

Received November 29, 2010; **Accepted** March 11, 2011; **Published** April 18, 2011

Copyright: © 2011 Ruiz et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This project received financial support from the following grants: FAPEMIG (Fundação de Amparo à pesquisa do Estado de Minas Gerais) (CBB-1181/0 and REDE-186/08 to Guilherme Oliveira), NIH (National Institutes of Health) - Fogarty (TW007012 to Guilherme Oliveira). Guilherme Oliveira and Vasco Azevedo are CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) fellows. The work also received support from CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vasco@icb.ufmg.br

◑ These authors contributed equally to this work.

Introduction

Corynebacterium pseudotuberculosis is a facultative intracellular pathogen that mainly infects sheep and goats, causing the disease called caseous lymphadenitis (CL). This bacterium can also cause ulcerative lymphangitis in equines; superficial abscesses in bovines, pigs, deer and laboratory animals; arthritis and bursitis in ovines; pectoral abscesses in equines and, more rarely, in camels, caprines and deer [1-3]. In both disease manifestations, its main characteristic is abscessing of the lymph nodes [4]. Rare cases of human infection have also been reported [5,6].

Despite the broad spectrum of hosts, the high incidence of CL reported from various countries, including Australia, New Zealand, South Africa, the United States of America, Canada and Brazil, mainly refers to small ruminants [7-11]. According to the World Animal Health Organization, among 201 countries that reported their sanitary situations, 64 declared the presence of animals with CL within their borders (OIE, 2009). The highest prevalence of CL has been reported in Brazil [12]. Pinheiro and colleagues (2000) reported 66.9% of animals with clinical signs of CL in the state of Ceará. In Minas Gerais state, a prevalence of 75.8% was reported for sheep [13] and 78.9% for goats [14]. In Australia, 61% of sheep flocks showed signs of infection [15]. In the USA, the prevalence ranges up to 43% [16]. Similar levels have been reported from the Canadian province of Quebec, with a prevalence of 21 to 36% [10]. In the United Kingdom, 45% of the producers that were polled reported abscesses in their sheep [9].

The high prevalence of CL in sheep and goats has made studies on ways to detect *C. pseudotuberculosis* in these hosts increasingly important; an efficient means to accomplish this would be a valuable tool for the control of this disease. Currently, there is no sufficiently sensitive and specific diagnostic test for subclinical CL. Diagnosis is currently achieved only by routine bacterial culture of purulent material collected from animals that have external abscesses, with subsequent biochemical identification of the isolates [17]. A few vaccines against CL are currently available, although they have not been licensed for use in many countries. Not all vaccines that have been developed for sheep are effective in goats. It is usually necessary to adjust vaccination programs to each animal host species [18].

Considering the current unfortunate status of CL prevalence in the world, especially in Brazil and Australia, there is a pressing need for more efficient alternatives for disease control that not only cure sick animals but also minimize or even prevent the onset of disease in herds. One of the major efforts to eradicate this disease involves the identification of genes that are related to the *C. pseudotuberculosis* pathogenicity and lifestyle. As an intracellular facultative pathogen, *C. pseudotuberculosis* exhibits several characteristics in its genome, such as gene loss, low GC content and a reduced genome [19] that differ from those of non-pathogenic *Corynebacterium* species. The finding of seven putative pathogenicity islands containing classical virulence elements, including genes for iron uptake, fimbrial subunits, insertional elements and secreted toxins [20], probably mostly acquired through horizontal transfer, contributes to our understanding of how this species causes disease. Comprehensive knowledge of an organism's genome facilitates an exhaustive search for candidates for virulence genes, vaccine and antimicrobial targets, and components that could be used in diagnostic procedures.

The information retrieved from a single genome is insufficient to provide an understanding of all *C. pseudotuberculosis* strains. Comparative genomics can shed light on the molecular attributes of a strain that affect its virulence, host specificity, dissemination potential and resistance to antimicrobial agents [21,22]. Furthermore, comparison of entire genome sequences of strains belonging to the same species, but from different geographic, epidemiological, chronological and clinical backgrounds, as well as affecting different hosts, would be useful for determining the molecular basis of these differences. As part of an effort to provide means to control CL, we examined the genomes of two strains of *C. pseudotuberculosis* isolated from sheep and goats, respectively, and compared them to each other and to the genomes of two other strains already available in a public database [6,23].

Results

Corynebacterium pseudotuberculosis genome

Overviews of the *C. pseudotuberculosis* genomes can be seen in Figure 1. The genomes are available in the NCBI GenBank database under accession numbers Cp1002:CP001809 and CpC231:CP001829.

The two strains are very similar, with an amino acid similarity of at least 95% between their predicted proteins. In their genomic composition, the isolates were found to have the same mean i) GC content, ii) gene length, iii) operon composition and iv) gene density. However, some significant differences were observed in: i) genome size, ii) number of pseudogenes and iii) lineage-specific genes (Table 1).

Gene order in *C. pseudotuberculosis*

To determine whether synteny was maintained between the two *C. pseudotuberculosis* strains, we made a comparative analysis of global gene order. As expected, the two *C. pseudotuberculosis* strains showed high synteny conservation; approximately 97% of their genes were found to be conserved in the comparison between the two strains. Previous studies provide evidence of a high degree of conservation of gene order in four *Corynebacterium* genomes, *C. diphtheriae*, *C. glutamicum*, *C. efficiens* and *C. jeikeium*, showing only 10

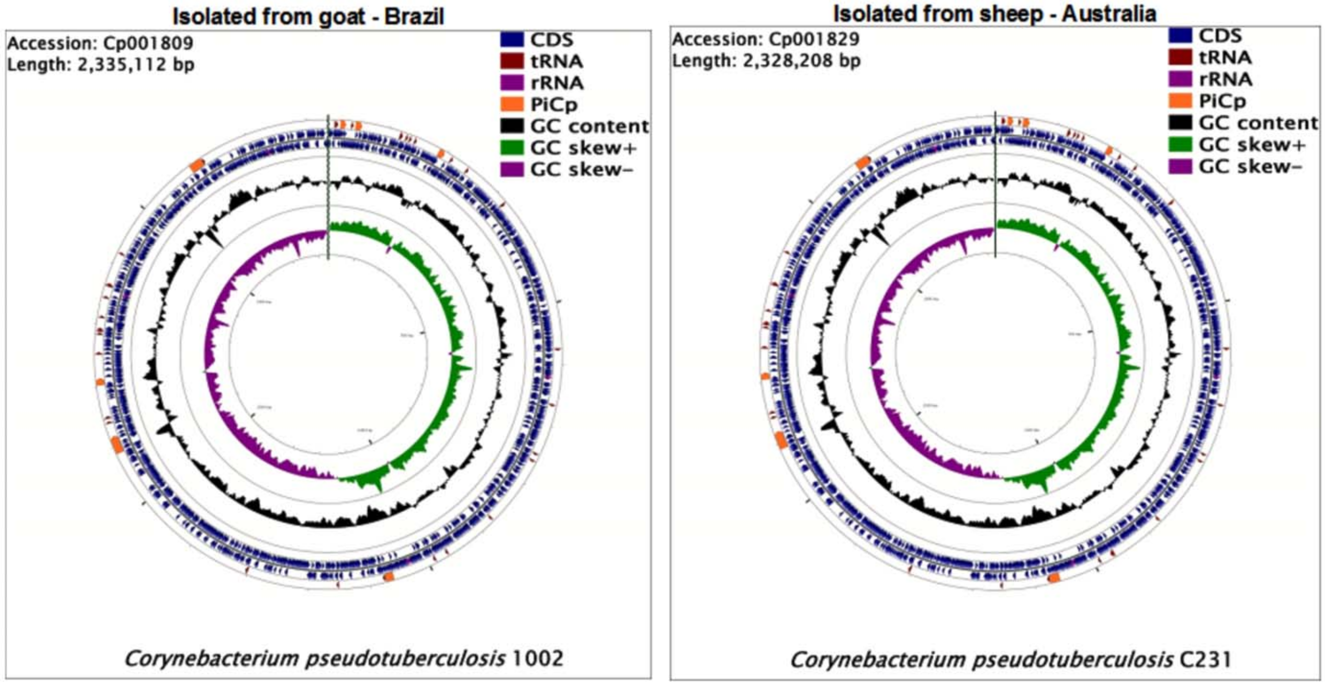


Figure 1. The whole genome of *Corynebacterium pseudotuberculosis.* Cp1002 strain isolated from a goat in Brazil and CpC231 strain isolated from sheep in Australia. Highlighted in yellow are the pathogenicity islands (PiCps) of *C. pseudotubeculosis* and its location in the genomes.
doi:10.1371/journal.pone.0018551.g001

gene-order breakpoints; rearrangement events during evolution in this species appear to be rare [24,25]. We checked the validity of this conclusion by making a comparative analysis of the genomes of the two *C. pseudotuberculosis* strains against *C. diphtheriae*, the *Corynebacterium* species that is most closely related to *C. pseudotuberculosis* [26,27].

Both *C. pseudotuberculosis* genomes showed a high degree of conservation in gene position, when compared to the *C. diphtheriae* genome, with few rearrangement points. This finding supports the hypothesis of a high degree of synteny conservation in this genus [25].

Table 1. General features of the genomes of two *Corynebacterium pseudotuberculosis* strains.

Genome feature	Cp1002	CpC231
Genome size (bp)	2,335,112	2,328,208
Gene number	2111	2103
Operon predicted number	474	468
Pseudogene number	53	50
tRNA number	48	48
rRNA operon	4	4
Gene mean length (bp)	964	968
Gene density (%)	0.88	0.88
Coding percentage	84.9	85.4
GC content (gene) (%)	52.88	52.86
GC content (genome) (%)	52.19	52.19
Lineage-specific genes	52	49

doi:10.1371/journal.pone.0018551.t001

Pathogenicity islands (PAIs)

Pathogenicity islands in bacterial genomes can be characterized by looking for characteristics linked to horizontal gene transfer, such as differences in codon usage, G+C content, dinucleotide frequency, insertion sequences, and tRNA flanking regions, together with transposase coding genes, which are involved in incorporation of DNA by transformation, conjugation or bacteriophage infection [28].

Pathogenicity islands had not been reported for *C. pseudotuberculosis*; to date; we used a multi-pronged approach called PIPS (submitted article) to identify the putative PAIs of *C. pseudotuberculosis*. Seven regions with most or all of the characteristics of horizontally-acquired DNA were found in both strains, Cp1002 and CpC231: i) base composition and/or codon usage deviations, ii) tRNA flanking, and iii) transposase genes. These regions were not found in a non-pathogenic species belonging to the same genus, *C. glutamicum*, and were classified as putative pathogenicity islands in *C. pseudotuberculosis* (PiCp). PiCps encode for proteins involved in the ABC transport system, for glycosil transferase, a two-component system, the *fag* operon and phospholipase D Table 2 provides a list of some genes found in the PAIs, with their respective functions.

Genetic composition of *C. pseudotuberculosis* Pathogenicity Islands

The genetic composition of PAIs can shed light on the lifestyle of pathogenic bacteria, since they include virulence genes that mediate mechanisms of adhesion, invasion, colonization, proliferation into the host and evasion of the immune system [29,30]. In addition, PAIs are characterized as being unstable regions that can be affected by insertions and deletions, influencing bacterial adaptability to new environments and hosts [31]. Here follows descriptions of the most relevant genetic elements found in the *C. pseudotuberculosis* pathogenicity islands. For more information, see

Table 2. Genes and proteins present in pathogenicity islands of the *Corynebacterium pseudotuberculosis* strain genomes.

PAI	Cp1002	CpC231	Protein
	tnp7109-9	tnp7109-9	Transposase for insertion sequence
	pld	pld	Phospholipase D precursor (PLD)
PiCp 1	fag C	fag C	ATP binding cytoplasmic membrane protein - FagC
	fag B	fag B	Iron-enterobactin transporter - FagB
	fag A	fag A	Integral membrane protein - FagA
	fag D	fag D	Iron siderophore binding protein - FagD
	mgtE	mgtE	Mg2+ transporter mgtE
	malL	malL	Oligo-1,6-glucosidase
PiCp 2	tetA	tetA	Putative tetracycline-efflux transporter
	cskE	cskE	Anti-sigma factor
	sigK	sigK	ECF family sigma factor K
	dipZ	dipZ	Integral membrane C-type cytochrome biogenesis protein DipZ
	potG	potG	Putrescine ABC transport system
	afuB	afuB	Putative transport system permease (iron)
PiCp 3	afuA	afuA	Iron (Fe3+) ABC superfamily ATP binding cassette transporter, binding protein
	glpT	glpT	Glycerol-3-phosphate transporter
	phoB	phoB	Two-component regulatory protein
	lcoS	lcoS	Two-component sensor protein, sensor histidine kinase
	ciuA	ciuA	Putative iron transport system binding (secreted) protein
	ciuB	ciuB	Putative iron transport system membrane protein
PiCp 4	ciuC	ciuC	Putative iron transport system membrane protein
	ciuD	ciuD	Putative iron ABC transport system
	ciuE	ciuE	Putative siderophore biosynthesis related protein
	σ70	σ70	Putative RNA polymerase sigma factor 70
	Pseudogene	Pseudogene	Putative chromosome segregation ATPase
PiCp 5	hsdR	hsdR	Putative type III restriction-modification system
	pfoS	pfoS	PfoR superfamily protein
	htaC	htaC	HtaA family protein
	guaB3	guaB3	Inosine 5-monophosphate dehydrogenase
PiCp6	pipA1	pipB	Proline iminopeptidase
	mfsD1	mfsD1	Major facilitator superfamily domain-containing protein 1
	dcd	dcd	Deoxycytidine triphosphate deaminase
	udg	udg	UDP-glucose 6-dehydrogenase
	lysS1	lysS1	Lysyl-tRNA synthetase
	alaT	alaT	Aminotransferase AlaT
	ureA	ureA	Urease gamma subunit
	ureB	ureB	Urease beta subunit
	ureC	ureC	Putative urease subunit alpha
PiCp 7	ureE	ureE	Urease accessory protein
	ureF	ureF	Urease accessory protein
	ureG	ureG	Urease accessory protein
	ureD	ureD	Urease accessory protein
	fepC2	fepC2	ABC superfamily ATP binding cassette transporter
	fecD	fecD1	Iron(III) dicitrate transport system permease fecD
	phuC	phuC	Iron(III) dicitrate transport permease-like protein yusV
	arsR	arsR1	ArsR-family transcription regulator

doi:10.1371/journal.pone.0018551.t002

the list of these orthologous genes in other *Corynebacterium* species in the Table S1 (online supporting information).

PiCp 1. *C. pseudotuberculosis* PiCp 1 harbors key genes involved in virulence and pathogenicity; these include PLD, the major virulence factor of this organism, which plays a role in spreading through the host; the *fag* operon, responsible for extracellular iron acquisition and, consequently, for survival in hostile environments; and a transposase gene, probably responsible for insertion of the island into the *C. pseudotuberculosis* genome. The finding that *C. ulcerans* can produce phospholipase D protein [32] indicates acquisition of PiCp1 by both *C. pseudotuberculosis* and *C. ulcerans*.

PiCp 2. Gene *mgtE* of island 2 has Mg^{2+} influx activity [33]. In prokaryotes, Mg^{2+} has been identified as an important regulatory signal that is essential for virulence, since it is involved in thermal adaptation, protecting bacteria from heat shock caused by fever in warm-blooded mammals [34]. Translation of the *mgtE* gene is regulated by changes in cytosolic Mg^{2+} concentration; loss of MgtE reduces biofilm formation and motility in the pathogenic bacteria *Aeromonas hydrophila* [33].

The protein MalL (*malL*), a maltose-inducible α-glucosidase, hydrolyzes various disaccharides, such as maltose and isomaltose, which can serve as carbon and energy sources [35,36].

The *tetA* gene codes for a tetracycline-efflux transporter protein that extrudes antibiotics from the cell and confers resistance to biofilm cells. The *tetA* gene is often carried by transmissible elements, such as plasmids, transposons, and integrons [37], thus explaining its presence in a PAI.

The *sigK* gene is an extracytoplasmic function sigma factor (sigma ECF) regulated by cskE, an anti-sigma factor. Another sigma ECF, *sigK*, mediates targeted alterations in bacterial transcription via transduction of extracellular signals. In *M. tuberculosis*, *sigK* regulates several genes (*Rv2871*, *mpt83*, dipZ, *mpt70*, *Rv2876*, and *mpt53*). Also, *sigK* mutations produce reduced quantities of the antigens MPT70 and MPT83 in vitro, and only induce strong expression during infection of macrophages [38–40].

PiCp2 also harbors a *dipZ* gene, which is regulated by *sigK* and seems to play a role in macrophage infection by *M. tuberculosis*, although its function is not clearly elucidated. DipZ is found as two separate proteins in most bacteria: CcdA and TlpA-like. Also, a full-length *dipZ* gene, found in the phylum *Actinobacteria*, is present exclusively in pathogenic bacteria (*C. diphtheriae*, *C. jeikeium*, *M. avium*, *M. kansasii*, *M. marinum*, *M. ulcerans* and *M. tuberculosis*) [40].

PiCp 3. *potG* gene, of the *pot*FGHI operon, is a membrane-associated/ATP-binding protein that provides energy for putrescine (polyamine) uptake from the periplasmic space [41]. Although the *potFGHI* operon is a putrescine-specific transport system, *potG* is downregulated by another polyamine (spermine), which is produced only by eukaryotes. Carlson et al. (2009) demonstrated that transcription of the *potG* gene in *Francisella tularensis* decreases with high levels of spermine, while transcription of IS elements ISFtu1 and ISFtu2 increases in response to high levels of spermine in macrophages responding to bacterial infection. Also, many of the upregulated genes of *F. tularensis* (pseudogenes and transposase genes) are located near the IS elements in the chromosome [42].

The gene *glpT* belongs to the organophosphate:phosphate antiporter family of the major facilitator superfamily (MFS); it mediates transport of glycerol 3-phosphate (G3P) across the membrane in bacteria [43].

The PhoPR system regulates expression of various genes involved in metabolic, virulence and resistance processes in several intracellular bacterial pathogens [44]. Based on the information obtained from the complete genome sequence of *C. pseudotuberculosis*, we found that the PhoPR system is constituted of the *phoP* (714 bp) and *phoR* (1506 bp) genes, separated by a small 39-bp sequence, suggesting that these two genes are transcribed by a bicistronic operon. The size and organization of this system in *C. pseudotuberculosis* is similar to those of other Gram-positive bacteria [45]. Live bacteria attenuated via *phoP* inactivation are also promising vaccine candidates against tuberculosis. Several studies have reported the efficacy of attenuated mutant strains of *M. tuberculosis* as vaccines [46,47]. Phylogenetic relationships within the class *Actinobacteria* strongly suggest correlation of the *C. pseudotuberculosis* PhoPR system with virulence mechanisms. The *phoP* gene is an important subject for regulation studies; and is also a probable vaccine candidate against CL.

PiCp4. The operon *ciuABCDE* (*corynebacterium* iron uptake) was described in *C. diphtheriae* as an iron transport and siderophore biosynthesis system. Proteins involved in iron acquisition are recognized as virulence factors, since they help pathogens to obtain iron from a host by using siderophores to strip iron from carrier proteins, such as transferrin, lactoferrin, and hemoglobin-haptoglobin [48,48].

PiCp5. Island 5 harbors a gene (*pfoS*) related to the *pfoR* superfamily. The *pfo*R gene was previously characterized as responsible for positive regulation of production of perfringolysin A (*pfoA*) and other toxins in *Clostridium perfringens* [50]. The virulence factors regulated by *pfoR* have not been totally elucidated. However, it is well known that deactivation of this gene inhibits hemolysis through negative regulation of several *C. perfringens* toxins. *Clostridium perfringens* harbors a phospholipase C gene (*plc*) that serves a function similar to that of phospholipase D [51]. Additionally, PiCp 5 contains a putative sigma 70 factor that is responsible for transporting the transcription machinery to specific promoters. Interestingly, the putative sigma 70 factor presents a nonsense mutation in *C. pseudotuberculosis* strain C231, which could be responsible for differential gene expression.

PiCp6. The *pipA1* gene, which codes for a proline iminopeptidase, may have a role in pathogenesis, since it catalyses the removal of N-terminal proline residues from peptides; it also has a role in energy production [52]. In addition, a PIP-type protein is required for virulence of *Xanthomonas campestris pv. campes*tris [53].

PiCp7. Island 7 harbors a urease operon that is also present in *C. glutamicum*; it is flanked, on both sides, by regions that are absent in the non-pathogenic *C. glutamicum*. This mosaicism is a common feature of pathogenicity islands [54]. The *ure* operon presents a codon usage deviation in *C. glutamicum*, as in *C. pseudotuberculosis*, indicating that this region is a putative genomic island in *C. glutamicum*.

The *ure* operon is responsible for nitrogen acquisition through hydrolysis of urea to carbamate and ammonia. Production of ammonia by uropathogenic and enteropathogenic bacteria causes cellular damage and compromises the action of the host's immune system [55]. Considering this fact, due to the intramacrophagic location of *C. pseudotuberculosis* and the finding of this operon in a non-pathogenic bacterial species, additional studies will be needed to elucidate how *C. pseudotuberculosis* obtains urea from the host and how this operon affects pathogenicity.

PiCp 7 also harbors a lysyl-tRNA synthetase (*lysS*), responsible for lysine incorporation into its respective transfer tRNA. The importance of *lysS* would normally make its location on a PAI inviable, since it is essential for cell metabolism. However, it is the only tRNA synthetase gene that is duplicated in the genome.

Protein classification of *C. pseudotuberculosis* in the biological process

Using the controlled vocabulary of functional terms proposed by the Gene Ontology (GO) Consortium for gene products

Table 3. Subcellular prediction of the protein locations derived from complete genomes of *Corynebacterium* species.

Category/Species	Ce	CgB	CgK	CgR	Cj	Cd	Cu	Cp1002	CpC231	Total
Cytoplasm	2,158	2,11	2,082	2,158	1,49	1,594	1,432	1,399	1,389	15,812
Cytoplasm	504	557	541	561	333	375	332	364	356	3,923
PSE	230	254	249	252	197	204	179	201	201	1,967
Secreted	102	136	121	109	100	99	79	95	107	948
Total	2,994	3,057	2,993	3,08	2,12	2,272	2,02	2,059	2,053	22,648

Ce: *C. efficiens*; CgB: *C. glutamicum B*; CgK: *C. glutamicum K*; CgR: *C. glutamicum R*; Cj: *C. jeikeium;* Cd: *C. diphtheriae;* Cu: *C. urealyticum;* Cp1002: *C. pseudotuberculosis* 1002; CpC231: *C. pseudotuberculosis* C231. PSE: potential surface exposure.

doi:10.1371/journal.pone.0018551.t003

classification [56], the predicted proteomes of the two genomes were analyzed according to the three organizing principles of gene ontology: cellular component, biological process and molecular function. The most abundantly represented categories are linked to metabolic processes in the two strains (cellular metabolic, biosynthetic, primary and macromolecule processes).

The gene products composition characterized using GO terminology suggests that *C. pseudotuberculosis* is a facultative intracellular pathogen. It is commonly found that pathogens specialized for an intracellular lifestyle have a high proportion of proteins linked to the above-mentioned processes. Moreover, the low proportion of proteins linked to the metabolism of secondary metabolites is an indication that *C. pseudotuberculosis* does not possess the metabolic machinery to deal with secondary metabolites, because they are supplied by the host.

Sub-cellular localization of *C. pseudotuberculosis* proteins

Prediction of the sub-cellular localization of *C. pseudotuberculosis* proteins was made by *in silico* analysis, using the SurfG+ tool [57]. Surfg+ is a pipeline for protein sub-cellular prediction, incorporating commonly used software for motif searches, including SignalP, LipoP and TMHMM, along with novel HMMSEARCH profiles to predict protein retention signals. Surfg+ starts by searching for retention signals, lipoproteins, SEC pathway export motifs and transmembrane motifs, roughly in this order. If none of these motifs are found in a protein sequence, then it is characterized as being cytoplasmic. A novel possibility introduced by Surfg+ is the ability to distinguish between integral membrane proteins versus PSE (potentially surface-exposed proteins). This is done by a parameter that determines the expected cell wall thickness, expressed in amino acids. Using published information or electron microscopy, it is possible to estimate cell wall thickness value for procaryotic organisms. *C. pseudotuberculosis* proteins were classified into four different sub-cellular locations: cytoplasmic, membrane, PSE (potentially surface exposed), or secreted. The *C. pseudotuberculosis* genomes were compared to those of other species of the genus, including C. *diphtheriae, C. efficiens, C. glutamicum*, C. *jeikeium* and C. *urealyticum*, also predicted by Surfg+, based on published cell wall thicknesses. Table 3 shows the number of predicted proteins in each sub-cellular location.

Comparison of the frequencies of subcellular occurrence of the *C. pseudotuberculosis* proteins and other *Corynebacterium* proteomes was made with Chi-square tests. The ratio between the four groups (cytoplasmic, membrane anchored, potentially exposed and secreted proteins) was found to be nearly constant among the *Corynebacterium* species. The proportions of the four protein categories cited above were similar to published data [58,59]. Song and colleagues (2009) showed that approximately 30% of proteins secreted in gram-positive bacteria are exported through the Sec pathway. Few proteins (n = 27) were predicted to be secreted by the Tat pathway in Cp1002. About 2% of the proteins predicted to be secreted presented tertiary structures. In terms of proportions of secreted proteins, Cp1002 and CpC231 are at the higher end of the spectrum. They present 4.61 and 5.21%, respectively, predicted secreted proteins (Table 3).

Differences in metabolic pathways in the two strains of *C. pseudotuberculosis*

Automated reconstruction of the *C. pseudotuberculosis* Cp1002 metabolic pathways identified 156 pathways and 744 enzymatic reactions. As expected, quite similar results were encountered for strain CpC231: 154 pathways and 754 reactions (Table 4). Proteins of predicted functions that did not map to pathways, such as transport reactions, enzymes, transporters, and compounds, were also identified. The metabolic pathway database can be accessed online at http://corynecyc.cebio.org. This database enabled us to visualize and compare the metabolism of these two *C. pseudotuberculosis* strains (Figure 2).

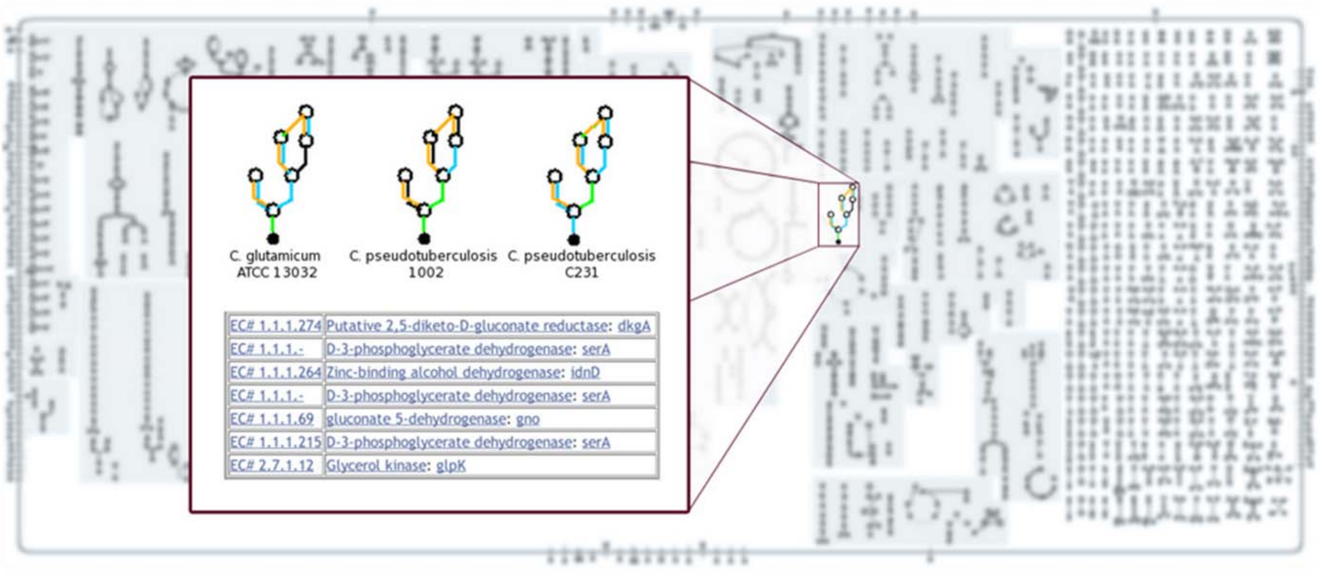
We made a comparative analysis of transport reactions, pathways, compounds and proteins for *C. pseudotuberculosis* strains Cp1002 and CpC231 (Table 5). Despite the high similarity of the metabolic pathways, some differences were observed.

The metabolic pathways in each of the two bacterial strains (Cp1002 and CpC231) were classified into several pathway classes; each pathway class was further broken down to show the distribution of pathways among the next-level subclasses. Analysis of the metabolism database of *C. pseudotuberculosis* strains Cp1002 and CpC231 revealed specific pathway differences between the two strains. Overall, CpC231 had 13 specific metabolic pathways

Table 4. Comparative summary of the *Corynebacterium pseudotuberculosis* strain gene data types.

Data Type	Cp1002	CpC231
Gene products	2,059	2,053
Pathways	156	154
Enzymatic Reactions	744	754
Transport Reactions	8	4
Polypeptides	2,065	2,059
Enzymes	516	506
Transporters	10	10
Compounds	639	651

doi:10.1371/journal.pone.0018551.t004



EC# 1.1.1.274	Putative 2,5-diketo-D-gluconate reductase: dkgA
EC# 1.1.1.-	D-3-phosphoglycerate dehydrogenase: serA
EC# 1.1.1.264	Zinc-binding alcohol dehydrogenase: idnD
EC# 1.1.1.-	D-3-phosphoglycerate dehydrogenase: serA
EC# 1.1.1.69	gluconate 5-dehydrogenase: gno
EC# 1.1.1.215	D-3-phosphoglycerate dehydrogenase: serA
EC# 2.7.1.12	Glycerol kinase: glpK

Figure 2. ***Corynebacterium glutamicum*** **metabolic pathways overview.** *C. glutamicum* reactions are presented in blue and the reactions shared with *C. pseudotuberculosis* C231 and 1002 in red and green, respectively. By clicking on any compound or reaction, a window pops up showing details of each pathway. The fatty acid biosynthesis initiation pathway is the chosen example since computational evidence indicates it is not present only in strain C231.
doi:10.1371/journal.pone.0018551.g002

not found in strain Cp1002, and the latter had 11 metabolic pathways not found in strain CpC231 (Table 6).

Two amine and polyamine biosynthesis pathways, choline degradation I and glycine betaine biosynthesis I (Gram-negative bacteria), were found in strain Cp1002 but not in strain CpC231. Strain CpC231 was found to have an extra amino acid biosynthesis pathway, the citrulline-nitric oxide cycle. Strain Cp1002 was found to have three additional carbohydrate biosynthesis pathways: gluconeogenesis, trehalose biosynthesis II and trehalose biosynthesis III. Strain CpC231 showed three cofactor biosynthesis, prosthetic group and electron carrier pathways, corresponding to adenosylcobalamin biosynthesis from cobyrinate a,c-diamide I, heme biosynthesis from uroporphyrinogen II and siroheme biosynthesis. Strain Cp1002 showed only one unique cofactor biosynthesis pathway, heme biosynthesis from uroporphyrinogen I. Two extra pathways of fatty acid and lipid biosynthesis were found in strain Cp1002, cardiolipin biosynthesis I and fatty acid biosynthesis initiation I. Strain CpC231 showed only the biotin-carboxyl carrier protein. Among metabolic regulator biosynthesis genes, strain CpC231 showed the citrulline-nitric oxide cycle. Strain CpC231 also showed an extra pathway, the canavanine biosynthesis pathway, part of secondary metabolite biosynthesis.

Among degradation/utilization/assimilation pathways, strain Cp1002 showed an extra pathway: glycerol degradation II, for alcohol degradation, as well as choline degradation I for amine and polyamine degradation. Strain CpC231 was found to have two additional pathways, 2-ketoglutarate dehydrogenase complex and citrulline-nitric oxide cycle, for amino acid pathways; strain Cp1002 showed only one extra pathway, valine degradation I. Among carboxylate degradation pathways, involving fatty acid and lipid degradation, strain Cp1002 showed two extra pathways: one corresponding to acetate formation from acetyl-CoA I, and the second linked to triacylglycerol degradation. Two inorganic nutrient metabolism pathways were found in strain CpC231 but not in strain Cp1002: nitrate reduction III (dissimilatory) and nitrate reduction IV (dissimilatory), and a nucleoside and nucleotide degradation and purine deoxyribonucleoside recycling degradation pathway.

Finally, when we analyzed the generation of precursor metabolites and energy, strain CpC231 showed three extra pathways: 2-ketoglutarate dehydrogenase complex, nitrate reduction III (dissimilatory) and nitrate reduction IV (dissimilatory). The differences are presented in Table 6.

Metabolic pathways in *C. pseudotuberculosis* compared to other *Corynebacterium* species

The web interface enabled us to visually compare the metabolic pathways of strains Cp1002 and CpC231 reactions (Figure 2) with those of four other bacteria of the genus *Corynebacterium*: *C. diphtheriae, C. efficiens, C. glutamicum, and C. jeikeium*. Using these diagrams we were able to easily spot reactions present in *C. pseudotuberculosis* and absent in other *Corynebacterium* species.

A comparative analysis of reactions, pathways, compounds and proteins was also done for *C. pseudotuberculosis* and other closely-related bacteria in the same genus. The list of *C. pseudotuberculosis* specific pathways is shown in Table 7.

We found that *C. pseudotuberculosis* has several pathways that are not found in other species of the genus *Corynebacterium*. However, little information is available about these pathways in *Corynebacterium* spp. We found no published information concerning the following pathways: asparagine biosynthesis II, citrulline-nitric oxide cycle (amino acid biosynthesis and degradation), pyrimidine deoxyribonucleotide salvage pathways, methylglyoxal degradation III, reductive monocarboxylic acid cycle, chitobiose degradation, conversion of succinate to propionate, ammonia oxidation I (aerobic), nitrate reduction IV (dissimilatory), D-glucarate degradation, betanidin degradation, D-galactarate degradation, and ammonia oxidation I (aerobic).

Some studies reported five pathways: lysine biosynthesis V, glycerol degradation II, alanine degradation IV, lysine degradation I and phospholipases. However, none of the studies, except for those concerning lysine degradation I and phospholipase

Table 5. Comparative summary of the number of pathways of *Corynebacterium pseudotuberculosis* strains Cp1002 and CpC231.

Pathway Class	Cp1002	CpC231
- Pathway subclass		
Biosynthesis	105	104
- Amine and Polyamine Biosynthesis	5	3
- Amino acid Biosynthesis	25	26
- Aminoacyl-tRNA Charging	1	1
- Aromatic Compound Biosynthesis	1	1
- Carbohydrate Biosynthesis	10	7
- Cell structure Biosynthesis	4	4
- Cofactor, Prosthetic Group, Electron Carrier Biosynthesis	27	29
- Fatty Acid and Lipid Biosynthesis	8	7
- Metabolic Regulator Biosynthesis	1	2
- Nucleoside and Nucleotide Biosynthesis	12	12
- Other Biosynthesis	1	1
- Secondary Metabolites Biosynthesis	1	2
Degradation/Utilization/Assimilation	53	54
- Alcohol Degradation	2	1
- Aldehyde Degradation	1	1
- Amine and Polyamine Degradation	5	4
- Amino Acid Degradation	11	12
- C1 Compound Utilization and Assimilation	4	4
- Carbohydrate Degradation	7	7
- Carboxylate Degradation	5	4
- Degradation/Utilization/Assimilation - Other	5	5
- Fatty Acid and Lipid Degradation	3	2
- Inorganic Nutrient Metabolism	4	6
- Nucleoside and Nucleotide Degradation and Recycling	2	3
- Secondary Metabolite Degradation	5	5
Generation of precursor metabolites and energy	16	19
Total	163	164

doi:10.1371/journal.pone.0018551.t005

pathways, involved *C. pseudotuberculosis*. Most of these studies were carried out with *C. glutamicum*.

Four papers concerning *C. glutamicum* were found for the lysine degradation I pathway [60–63]. Studies have focused on: acetohydroxyacid synthase, a novel target for improvement of L-lysine production [62], improvement of L-lysine formation by expression of the *Escherichia coli* pntAB genes [61], genetic and functional analysis of soluble oxaloacetate decarboxylase [63], and modeling and experimental design for metabolic flux analysis of lysine-producing *Corynebacteria* by mass spectrometry [64].

Six studies were found concerning the glycerol degradation II pathway, one performed with C. *diphtheria* [65] and four with C. *glutamicum* [66–69]. In the sixth study, made with *C. glutamicum*, we found information on the alanine degradation IV pathway [64].

Approximately 140 studies, of which 107 were made with *C. glutamicum* alone, dealt with the lysine degradation I pathway, in which cadaverine is biosynthesized from L-lysine. Cadaverine is reported to be essential for the integrity of the cell envelope and for normal growth of the organism, as well as for inhibiting porin-mediated outer membrane permeability, thereby protecting cells from acid stress [70,71].

All studies of specific phospholipase pathways were carried out with *C. pseudotuberculosis*. Phospholipases hydrolyze phospholipids and are ubiquitous in all organisms. Several types of phospholipases were reported; phospholipase D is the best studied and has been considered a major virulence factor for *C. pseudotuberculosis* [72,73]. In our analyses, none of the five bacteria of the genus *Corynebacterium* were found to have pathways belonging to the following subclasses: siderophore biosynthesis; chlorinated compound degradation; cofactor, prosthetic group, electron carrier, and hormone degradation. Clearly more biochemical studies are needed. Our current study brings new insight to relevant biochemical pathways that can be further explored experimentally.

We made a comparative summary of the metabolic pathways of *C. pseudotuberculosis* strains Cp1002 and CpC231 and *C. glutamicum* (Table 8). *C. glutamicum* has several metabolic pathways not found in *C. pseudotuberculosis* Cp1002 and/or in *C. pseudotuberculosis* CpC231. Overall, *C. glutamicum* has approximately 40 additional metabolic pathways.

Among biosynthesis pathways, *C. glutamicum* showed around 30 extra pathways when compared to the two strains of *C. pseudotuberculosis*. These involve pathways of amino acid biosynthesis, aminoacyl-tRNA charging, cofactors, prosthetic groups, electron carrier biosynthesis, fatty acid and lipid biosynthesis and secondary metabolite biosynthesis. However, the two strains of *C. pseudotuberculosis* also have specific pathways that were not found in *C. glutamicum*, these being the pathways of amine and polyamine biosynthesis, carbohydrate biosynthesis and nucleoside and nucleotide biosynthesis.

Among the degradation/utilization/assimilation pathways, *C. glutamicum* presented around 20 extra pathways, when compared to *C. pseudotuberculosis* Cp 1002 and *C. pseudotuberculosis* CpC231. These specific pathways of *C. glutamicum* correspond to pathways of amine and polyamine degradation, amino acid degradation, aromatic compound degradation, carbohydrate degradation, carboxylate degradation, chlorinated compound degradation and the metabolism of inorganic nutrients. Again, the two strains of *C. pseudotuberculosis* also had specific pathways involving degradation/utilization/assimilation, fatty acid and lipid degradation and secondary metabolite degradation that were not found in *C. glutamicum*.

We found 25 pathways involving generation of precursor metabolites and energy in *C. glutamicum*, while *C. pseudotuberculosis* Cp1002 had only 16 and *C. pseudotuberculosis* CpC231 had 19.

Discussion

General aspects of the *C. pseudotuberculosis* genome

The *C. pseudotuberculosis* genome has proven to be one of the smallest genomes of the *Corynebacterium* genus sequenced so far, with Cp1002 being the smallest and Cp231 the fourth smallest, larger only than Cp1002, *C. lipophiloflavum* DSM 44291 (2,293,743 bp) and *C. genitalium* ATCC 33030 (2,319,774 bp); the latter two are both human pathogens. *Corynebacterium pseudotuberculosis* has a very small genetic repertoire, with considerable gene loss when compared to non-pathogenic species such as *C. glutamicum* and *C. efficiens*. When predicted proteomes were compared, *C. pseudotuberculosis* showed a loss of approximately 1,220 genes, in comparison with *C. glutamicum*. Classification of these proteins using GO terminology showed that the majority are linked to metabolic processes, such as cellular, primary, biosyn-

Table 6. Table listing the *Corynebacterium pseudotuberculosis* strain-specific pathways.

Pathway Class	**Cp1002**	**CpC231**
Pathway Name		
Biosynthesis - Amines and Polyamines Biosynthesis		
choline degradation I	present	absent
glycine betaine biosynthesis I (Gram-negative bacteria)	present	absent
Biosynthesis - Amino acid Biosynthesis		
citrulline-nitric oxide cycle	absent	present
Carbohydrates Biosynthesis		
gluconeogenesis	present	absent
trehalose biosynthesis II	present	absent
trehalose biosynthesis III	present	absent
Biosynthesis - Cofactor, Prosthetic Group, and Electron Carrier Biosynthesis		
adenosylcobalamin biosynthesis from cobyrinate a,c-diamide I	absent	present
heme biosynthesis from uroporphyrinogen I	present	absent
heme biosynthesis from uroporphyrinogen II	absent	present
siroheme biosynthesis	absent	present
Biosynthesis - Fatty Acid and Lipid Biosynthesis		
biotin-carboxyl carrier protein	absent	present
cardiolipin biosynthesis I	present	absent
fatty acid biosynthesis initiation I	present	absent
Secondary Metabolite Biosynthesis		
canavanine biosynthesis	absent	present
Biosynthesis - Metabolic Regulators Biosynthesis		
citrulline-nitric oxide cycle	absent	present
Degradation - Alcohols Degradation		
glycerol degradation II	present	absent
Degradation - Aldehyde Degradation		
methylglyoxal degradation I	absent	present
methylglyoxal degradation III	present	absent
Degradation - Amine and Polyamine Degradation		
choline degradation I	present	absent
Degradation - Amino Acid Degradation		
2-ketoglutarate dehydrogenase complex	absent	present
citrulline-nitric oxide cycle	absent	present
valine degradation I	present	absent
Degradation - Carboxylate Degradation		
acetate formation from acetyl-CoA I	present	absent
Degradation - Fatty Acid and Lipids Degradation		
triacylglycerol degradation	present	absent
Inorganic Nutrients Metabolism		
nitrate reduction III (dissimilatory)	absent	present
nitrate reduction IV (dissimilatory)	absent	present
Degradation - Nucleoside and Nucleotide Degradation and Recycling		
purine deoxyribonucleoside degradation	absent	present
Generation of precursor metabolites and energy		
2-ketoglutarate dehydrogenase complex	absent	present
nitrate reduction III (dissimilatory)	absent	present
nitrate reduction IV (dissimilatory)	absent	present

doi:10.1371/journal.pone.0018551.t006

Table 7. List of *Corynebacterium pseudotuberculosis* specific metabolic pathways that were compared to those of closely-related bacteria, including *C. diphtheriae*, *C. glutamicum*, *C. efficiens*, and *C. jeikeium*.

Pathway Class
Pathway Name
Biosynthesis - Amino acid Biosynthesis
Asparagine biosynthesis II
Lysine biosynthesis V
Biosynthesis - Metabolic Regulators Biosynthesis
Citrulline-nitric oxide cycle
Biosynthesis - Nucleoside and Nucleotide Biosynthesis
Salvage pathways of pyrimidine deoxyribonucleotides
Degradation - Alcohol Degradation
Glycerol degradation II
Degradation - Aldehyde Degradation
Methylglyoxal degradation III
Degradation - Amino Acid Degradation
Alanine degradation IV
Citrulline-nitric oxide cycle
Lysine degradation I
Degradation - C1 Compound Utilization and Assimilation
Reductive monocarboxylic acid cycle
Degradation - Carbohydrate Degradation
Chitobiose degradation
Degradation - Carboxylate Degradation
Conversion of succinate to propionate
Degradation - Fatty Acid and Lipid Degradation
Phospholipases
Inorganic Nutrients Metabolism
Ammonia oxidation I (aerobic)
Nitrate reduction IV (dissimilatory)
Degradation - Secondary Metabolite Degradation
D-glucarate degradation
Betanidin degradation
D-galactarate degradation
Generation of precursor metabolites and energy
Ammonia oxidation I (aerobic)

doi:10.1371/journal.pone.0018551.t007

thetic, macromolecule, nitrogen compound and oxidation reduction processes.

Other characteristics of the *C. pseudotuberculosis* genome include the lowest GC content in the *Corynebacterium* genus, this being 52% in both the goat and sheep strains, followed by *C. diphtheriae* with a GC content of 53%. This contrasts with *C. urealyticum*, which has a GC content of 64%. Furthermore, *C. pseudotuberculosis* has a higher number of predicted pseudogenes and a lower number of tRNAs, when compared to other species of the *Corynebacterium* genus for which genome sequences are available.

Merjeh et al. (2009) made a comparative analysis of 317 genomes of bacteria with different lifestyles (free-living, facultative intracellular and obligate intracellular). They found evidence that peculiar characteristics in bacterial genomes can drive the organisms to certain lifestyles. All characteristics cited in their work were identified in the *C. pseudotuberculosis* genomes. Lower GC content generally can occur due to gene loss, which is a means to contract the genome in response to a specialized environment. Moreover, presence of a higher number of pseudogenes could be evidence of bacterial mechanisms to generate non-functional genes and subsequent gene loss [19]. In addition, the high proportion of proteins linked to primary metabolism, and the small proportion of proteins related to secondary metabolism, is usually seen in facultative intracellular organisms. Taking these aspects of the genomic architecture of *C. pseudotuberculosis* into account, it can be affirmed that *C. pseudotuberculosis* has a facultative intracellular lifestyle.

High similarity in the genome architecture

Usually, pseudogenes are characterized as genes that have lost their function in the genome, due either to changes in the reading frame (frameshifts) or to a premature stop codon. Pseudogenes are common in prokaryotes; most have been linked to a sudden change in the environment of the pathogen, with simultaneous loss of metabolic and respiratory activities [74].

The high number of pseudogenes in these two strains of *C. pseudotuberculosis* (52 in Cp1002 and 50 pseudogenes in CpC231) suggest an evolutionary process involving a contracting genome in this species. An example of this is also seen in *Mycobacterium leprae*, which has a large number of pseudogenes (around 1,000). When we compare *M. leprae* to *M. tuberculosis*, the latter has both considerably fewer genes and a higher number of pseudogenes that can drive this gene loss.

Virulence factors acquired

Identification of pathogenicity islands (PAIs) in pathogenic bacteria is highly relevant for understanding the reasons behind different responses to vaccines and the biological mechanisms leading to genome plasticity. The biovars *equi* and *ovis* of *C. pseudotuberculosis* cause distinct diseases in their hosts; assessment of virulence genes could help identify genes involved in these host-specific differences.

Virulence genes, which are central to distinguishing pathogenic from non-pathogenic species, are present in PAIs in large numbers. Additionally, the fact that PAIs are a consequence of horizontal transfer events indicates that the virulence factors they contain can help increase the adaptability of strains to different host environments. This increase in adaptability is demonstrated by the finding of genes with functions associated with uptake of iron (*fag* operon), carbon (*malL*) and Mg^{2+} from the host, since this uptake improves survival under stress conditions, such as iron depletion, starvation and heat shock. Furthermore, PAIs of *C. pseudotuberculosis* present genes that respond to a macrophagic environment (*potG*, *sigK* and *dipZ*), which sheds new light on the mechanisms responsible for the intramacrophagic lifestyle of this organism.

Gene Sharing among *C. pseudotuberculosis* strains

Considering the four available genomes of *C. pseudotuberculosis* strains (Cp1002, CpC231, and CpI19 pFRC41), we identified 1,851 whole genes shared among them (Figure 3).

This repertoire of genes is vast for this specie, since, among the four isolates the maximum number of genes is 2,377 (called the pangenome of the species). When we compare the number of genes shared by these four *C. pseudotuberculosis* strains with a study of 17 strains of the bacterium *E. coli* [75], we conclude that *C. pseudotuberculosis* has a greater proportion of shared genes. In isolates of *E. coli*, 2,220 genes constituted the core genome, less

Table 8. Comparative summary of *Corynebacterium pseudotuberculosis* strains Cp1002 and CpC231 and *C. glutamicum* pathways.

Pathway Class **- Pathway subclass**	**Cp1002**	**CpC231**	***C. glutamicum***
Biosynthesis	105	104	131
- Amine and Polyamine Biosynthesis	5	3	3
- Amino acid Biosynthesis	25	26	29
- Aminoacyl-tRNA Charging	1	1	3
- Aromatic Compound Biosynthesis	1	1	1
- Carbohydrate Biosynthesis	10	7	9
- Cell structure Biosynthesis	4	4	4
- Cofactor, Prosthetic Group, Electron Carrier Biosynthesis	27	29	38
- Fatty Acid and Lipids Biosynthesis	8	7	14
- Metabolic Regulator Biosynthesis	1	2	1
- Nucleoside and Nucleotide Biosynthesis	12	12	10
- Other Biosynthesis	1	1	1
- Secondary Metabolite Biosynthesis	1	2	6
Degradation/Utilization/Assimilation	53	54	72
- Alcohols Degradation	2	1	2
- Aldehyde Degradation	1	1	1
- Amine and Polyamine Degradation	5	4	6
- Amino Acid Degradation	11	12	15
- Aromatic Compound Degradation	0	0	9
- C1 Compound Utilization and Assimilation	4	4	2
- Carbohydrate Degradation	7	7	10
- Carboxylate Degradation	5	4	6
- Chlorinated Compound Degradation	0	0	4
- Degradation/Utilization/Assimilation - Other	5	5	2
- Fatty Acid and Lipid Degradation	3	2	2
- Inorganic Nutrient Metabolism	4	6	9
- Nucleoside and Nucleotide Degradation and Recycling	2	3	1
- Secondary Metabolite Degradation	5	5	4
Generation of precursor metabolites and energy	16	19	25
Total	163	164	206

doi:10.1371/journal.pone.0018551.t008

than half of the genes in this species, with a mean of 5,000 genes in each genome [75]. Other significant information that emerges from this data is that the *C. pseudotuberculosis* genomes are extremely similar, since we found no significant change in the composition of the repertoire of genes for this species after adding the two new strains (Figure 3).

Gene Sharing between *C. pseudotuberculosis* and other *Corynebacterium* species

Previous comparative studies of sequences of the rpoB gene of *C. pseudotuberculosis* and *C. diphtheriae* have suggested a close relationship between them [27,76]. In our current study, we confirmed this close relationship with several types of evidence: i) a similar codon bias, ii) high similarity at the amino acid level and iii) conserved synteny. Synteny analysis of the genomes of the two *C. pseudotuberculosis* strains compared to *C. diphtheriae* indicates that these genomes are highly conserved; the gene position is conserved within the species. This observation reinforces the conclusions of previous research claiming conserved synteny in this genus, which indicated that few rearrangement events occurred during evolution [25].

Corynebacterium pseudotuberculosis shares more orthologous genes with *C. glutamicum* (1,345 genes), *C. efficiens* (1,330), *C. diphtheriae* (1,263 genes) and *C. auricumucosum* (1,273 genes); it shares only 1,030 genes with *C. jeikeium* and *C. kroppenstedtii*.

The larger number of genes shared between *C. pseudotuberculosis*, *C. glutamicum* and *C. efficiens* (72%), compared to other species (pathogenic species, 60%), may be a result not only of their close relationships, but also because a comparison is made among species with a larger gene repertoire, such as *C. glutamicum* and *C. efficiens*, which are non-pathogenic microorganisms, thus increasing the possibility of sharing genes.

Lineage-specific genes in *C. pseudotuberculosis*

Most of the lineage-specific genes are involved in processes of virulence, pathogenicity, drug resistance and response to certain types of stress. These factors can increase the adaptability of microorganisms to the niches they inhabit, but they are not

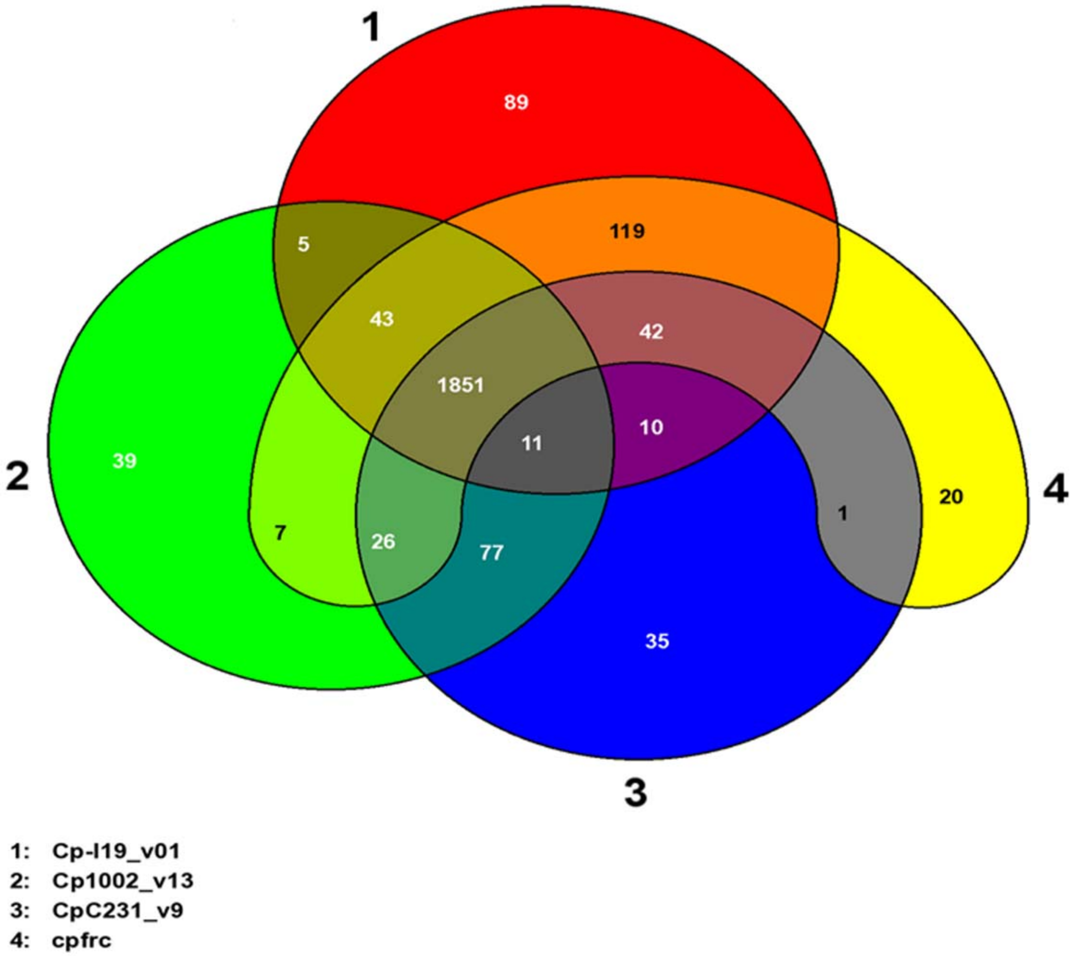


Figure 3. Venn diagram illustrating the three genomic categories of four *Corynebacterium pseudotuberculosis* strains: core, accessory and extended genome. Data obtained from the comparison of the predicted proteomes of four *C. pseudotuberculosis* speices in the EDGAR program (Blom et al., 2009). In red: Cp-I19; green: Cp1002; blue: CpC231 and yellow: CpFRC41. The remaining colors illustrate the shared genes among strains. The numbers within the forms indicate the number of shared genes.
doi:10.1371/journal.pone.0018551.g003

indispensable to the survival of pathogens. Moreover, some copies of these genes can be acquired by horizontal transfer. These genes are not ORFans; they already have been characterized in other species. The terminology 'lineage-specific' portrays only some genes found among the four strains in our study; the same genes may be found in other species.

We found 49 lineage-specific genes in CpC231 and 52 in Cp1002. For most of them, we did not have a descriptive characterization of their products, and they were classified as hypothetical proteins. In addition, many of these identified genes, in both strains, encode membrane and secreted proteins and pseudogenes. On the other hand, some well-characterized proteins were found in the genome. One example is found in CpC231, which has the gene called *pth*A; this gene encodes an effector system of type III secretion and is related to bacterial growth and host cell lesions, as found in *Xanthomonas campestris* [77]. This gene may be a good target for understanding the development of *C. pseudotuberculosis* CpC231 inside the host and the necrosis seen in CL abscesses, where it plays the same role in this pathogen.

In Cp1002, a very interesting gene was found, *tat*A, which encodes a membrane protein translocase, involved in the secretion of proteins in their final conformation, through the inner membrane to the extracellular environment. This gene is interesting because it is independent of the Sec secretion system and is a unique copy among the strains, suggesting that Cp1002 may have other routes for secretion. Regarding the large number of hypothetical proteins found in this strain, it may harbor genes that came from horizontal transfer, including some from phylogenetically-distant organisms, for which genomic molecular characterization has not been made.

Finally, lineage-specific genes may be good tools for understanding the host-pathogen interaction and may be good targets for the development of computational tools for differentiation between these strains, for molecular epidemiology.

Biochemical properties of *C. pseudotuberculosis*

In the latest review of the biochemical properties of *C. pseudotuberculosis* [76], Dorella and colleagues gathered information concerning its metabolism, virulence and pathogenesis. They reported that the peptidoglycan in the cell wall is based on meso-DAP acid, and that arabinose and galactose are major cell-wall sugars. Our analyses predicted all of the reactions of the peptidoglycan biosynthesis II pathway; the meso-DAP acid compound was found as a product/substrate of the reaction catalyzed by UDP-N-acetylmuramyl tripeptide synthase (6.3.2.13). The complete pathway of UDP-galactose biosynthesis was also found; although there was no evidence of biosynthesis of arabinose, we detected a membrane transporter, known as arabinose efflux permease.

We also found short-chain mycolic acids; 10 variations of acids of this type were encountered, including 6-O-cis-keto-mycolyl-trehalose-6-phosphate, and 6-O-mycolyl-trehalose-6-phosphate. The two strains of *C. pseudotuberculosis* showed considerable fermentation ability, with several fermentation pathways, including glycolysis III, mixed acid fermentation and pyruvate fermentation to acetate IV, ethanol I and lactate.

Several sugar degradation pathways were also found in the two strains of *C. pseudotuberculosis*, including galactose, lactose, sucrose and L-and D-arabinose degradation. We confirmed that, as reported by Dorella et al. (2006), all these pathways produce acids and no gasses, generating large amounts of energy.

It was also previously reported that *C. pseudotuberculosis* is phospholipase D and catalase positive. Our analysis showed that both phospholipase D and catalase are involved in important processes. The main molecular functions of phospholipase D are phospholipase D activity, magnesium ion binding, NAPE-specific phospholipase D activity and sphingomyelin phosphodiesterase D activity. Catalase, which is produced by the *cat* gene, is involved in response to oxidative stress and oxidation reduction. Although two enzymes of the denitrification pathway (nitrate reduction I) were found, absence of the remaining enzymes is probably the determining factor for the inability of these strains to reduce nitrate to N^2, as reported by Dorella et al. (2006).

We also detected iron acquisition genes (*fag*) A, B, C and D in both strains of *C. pseudotuberculosis* [78]. Genes *fagA* and *fagB* produce the integral membrane proteins FagA, an iron-enterobactin transporter, and FagBy; both have important roles, including ion, transmembrane, organic acid and protein transport. The ATP binding cytoplasmic membrane protein, FagC, produced by gene *fagC*, has two main molecular functions: ATP binding and ATPase activity. Finally, gene *fagD* produces the iron siderophore binding protein, FeAcquisition gene D, which has a role in iron ion transmembrane transport activity.

Computational reconstruction of the *C. pseudotuberculosis* pathways in our database not only allowed us to better visualize the metabolism of this bacterium, but also to compare it to closely related species. The main purpose of this analysis was to describe *C. pseudotuberculosis* metabolism by computational means, providing a predictive tool for "wet-lab" research.

Methods

Bacterial strains and growth conditions

Corynebacterium pseudotuberculosis 1002 biovar ovis (herein referred to as Cp1002) is a wild strain, isolated from a caprine host in Brazil. *Corynebacterium pseudotuberculosis* C231 biovar ovis (herein referred to as CpC231) is also a wild strain, isolated from an ovine host in Australia. Both strains were confirmed to be *C. pseudotuberculosis* by routine biochemical tests (API CORYNE, Biomerieux, Marcy l'Etoile, France). These strains were maintained in brain-heart-infusion broth (BHI – HiMedia Laboratories Pvt. Ltda, India) at 37°C, under rotation.

Preparation of high molecular weight DNA

Chromosomal DNA extraction was performed as follows: 50 mL of 48–72 h cultures of the two strains were centrifuged at 4°C and 2000 x *g* for 20 min. Cell pellets were re-suspended in 1 mL Tris/EDTA/NaCl [10 mM Tris/HCl (pH 7.0), 10 mM EDTA (pH 8.0), and 300 mM NaCl] and centrifuged again under the same conditions. Supernatants were discarded, and the pellets were re-suspended in 1 mL TE/lysozyme [25 mM Tris/HCl (pH 8.0), 10 mM EDTA (pH 8.0), 10 mM NaCl, and 10 mg lysozyme/mL]. Samples were then incubated at 37°C for 30 min. Thirty milliliters of 30% (w/v) sodium N-lauroyl-sarcosine (Sarcosyl) were added to each sample and the mixtures were incubated for 20 min at 65°C, followed by incubation for 5 min at 4°C. DNA was purified using phenol/chloroform/isoamyl alcohol (25:24:1) and precipitated with ethanol. DNA concentrations were determined spectrophotometrically, and the DNA was visualized in ethidium bromide-stained 0.7% agarose gels.

Construction of *Corynebacterium pseudotuberculosis* genomic libraries and Sanger sequencing

For the shotgun strategy used to sequence *C. pseudotuberculosis* 1002, four small fragment libraries were constructed using the TOPO Shotgun cloning kit and the pCR4 Blunt-TOPO vector (Invitrogen), according to the manufacturer's instructions. Sanger sequencing was carried out using the Minas Gerais Genome Network (http://rgmg.cpqrr.fiocruz.br). A total of 6,144 forward and reverse reads were produced using the DYEnamic Dye Terminator kit and run in a Megabace 1000 automated sequencer (GE Healthcare).

Genome Sequencing

Cp1002 was sequenced using both Sanger and pyrosequencing technologies. Pyrosequencing was carried out using 454 Life Sciences (Branford, CT). A total of 397,147 high quality reads and 86,154,153 high quality bases were obtained, which translates into approximately 31-fold coverage. The average length of the sequences was 253 bases. The sequences were delivered after quality filtering and preassembly with the Newbler assembler (454 Life Sciences).

CpC231 was sequenced with a Roche-454 FLX sequencer at the Australian Animal Health Laboratory, Geelong, Australia. A total of 347,361 reads generated 80,336,550 bases, giving 34-fold coverage of the genome. *De novo* assembly of the filtered sequence data was carried out using the Newbler software. This assembly produced 10 large contigs in four scaffolds. The remaining gaps in the genomic sequence were closed by PCR walking and Sanger sequencing of the resulting fragments.

Treatment and assembly data

The raw Sanger data obtained from sequencing were processed using the Phred-Phrap-Consed package [75]. Possible contaminants (plasmid DNA, sequences with similarity to vectors and other contaminants) were discarded using the Cross_match program (www.phrap.org). The quality value used in the base-calling program was $Q = 40$ (Probability of incorrect base call 1 in 10,000/base call accuracy 99.99%). An assembly using Phrap parameters (Force Level: 40 and Gap Length: 10,000) was carried out.

The 454 data were processed using the Newbler assembler (454 Life Sciences), and the final genomic consensus sequence was obtained using the Phrap algorithm.

Genome annotation

The annotation procedures involved the use of several algorithms in a multi-step process. Structural annotation was performed using the following software: FgenesB: gene predictor (www.softberry.com); RNAmmer: rRNA predictor [79]; tRNA-scan-SE: tRNA predictor [80]; and Tandem Repeat Finder: repetitive DNA predictor (tandem.bu.edu/trf/trf.html). Functional annotation was performed by similarity analyses, using public databases and InterProScan analysis [81]. Manual annotation was performed using Artemis [82].

Identification and confirmation of putative pseudogenes in the genome was carried out using Consed. Manual analysis was performed based on the Phred quality of each base in the frameshift area. This analysis enabled the identification of erroneous insertions or deletions of bases in the genome information produced by the sequencing process, and it avoided identification of false-positive pseudogenes.

Predictions of the cellular locations of *Corynebacterium* proteins were made using the program SurfG Plus (version 1.0), with a minimum protein size of 73 amino acids. Classification of predicted proteins in functional categories was made using the BLAST2GO program (www.blast2go.org). The cutoff value used was 10−6 (http://www.blast2go.org/).

In silico Identification of Pathogenicity Islands

In order to accurately identify and classify putative Pathogenicity Islands (PAIs) in the corynebacterial genomes, we developed a combined computational approach using several in-house scripts to integrate the prediction of diverse algorithms and databases, namely: Colombo-SIGIHMM [83], Artemis [82], tRNAscan-SE [80]; EMBOSS-geecee [84], ACT: the Artemis Comparison Tool [85], and mVIRdb [86].

In silico metabolic pathway construction

The two main data sources used for reconstructing the *C. pseudotuberculosis* metabolic pathways were the genome sequence file in FASTA format and the genome annotation file in GBK format. Metabolic pathways databases for strains 1002 and C231 were created using the Pathway tools 13 software, developed by SRI International [87]. The Pathway tools software contains algorithms that predict metabolic pathways of an organism from its genome by comparison to a reference pathways database known as MetaCyc [88]. Construction of a metabolic pathways database was done using BioCyc [89], in order to compare the different bacteria, *C. diphtheriae* NCTC 13129, *C. efficiens* YS-314, *C. glutamicum* ATCC 13032, and *C. jeikeium* K411, to the deduced *C. pseudotuberculosis* pathways.

Comparative analysis of *Corynebacterium pseudotuberculosis* strains

Comparative analyses were made for the two *C. pseudotuberculosis* strains. Similarity analyses of the two genomes were made using the BLAST - NCBI [90,91] and InterProScan databases. The Mauve algorithm (gel.ahabs.wisc.edu/mauve) and the ACT tool were used to identify whether blocks had undergone gene rearrangements or remained preserved. The Plotter program of the MUMMer 3.22 package (mummer.sourceforge.net) was used for synteny analysis.

Supporting Information

Table S1 Orthologous genes present inside PAIs regions of *C. pseudotuberculosis* and their counterparts in other Corynebacterium species.
(DOC)

Acknowledgments

The authors acknowledge the scientific support of two genomics networks: Rede Paraense de Genômica e Proteômica and Rede Genoma de Minas Gerais. In addition, the authors thank for valuable contribution of all Public Institutions involved and their respective co-authors during the development of the present work.

Author Contributions

Conceived and designed the experiments: JCR AS MPC RJM AM GCO VA. Performed the experiments: FAD SB MITF GCO AM VA VD EMC LMO MCP SRCD AFC JFA. Analyzed the data: JCR AS RJM GCO AM VA VD ARS FAD LGCP MZT NS TLPC JM AZ SCS SSA VACA DMR. Contributed reagents/materials/analysis tools: VA GCO GRF DOL ALP CUV CTG DCB DMO FRS EMR IMC JMO LVP LRG JAF MITF NPC PRKF SMRT SB SCO. Wrote the paper: JCR AS RJM GCO AM VA VD ARS FAD LGCP MZT NS TLPC JM AZ SCS SSA VACA DMR ET JB AT. Obtained permission for use of cell line: RJM RM AM VA. Bioinformatic support: JCR GCO AD FPL PG.

References

1. Ayers JL (1977) Caseous lymphadenitis in goat and sheep: review of diagnosis, pathogenesis, and immunity. JAVMA 171: 1251–1254.
2. Brown CC, Olander HJ, Alves SF (1987) Synergistic hemolysis-inhibition titers associated with caseous lymphadenitis in a slaughterhouse survey of goats and sheep in northeastern Brazil. Can J Vet Res 51: 46–49.
3. Merchant IA, Packer RA (1967) The genus *Corynebacterium*. In: Merchant IA, Packer RA, eds. Veterinary Bacteriology and Virology. USA: The Iowa State University Press. pp 425–440.
4. Piontkowski MD, Shivvers DW (1998) Evaluation of a commercially available vaccine against *Corynebacterium pseudotuberculosis* for use in sheep. J Am Vet Med Assoc 212: 1765–1768.
5. Join-Lambert OF, Ouache M, Canioni D, Beretti JL, Blanche S, et al. (2006) *Corynebacterium pseudotuberculosis* necrotizing lymphadenitis in a twelve-year old patient. Pediatr Infect Dis J 25(9): 848–851.
6. Trost E, Ott L, Schneider J, Schroder J, Jaenicke S, et al. (2010) The complete genome sequence of *Corynebacterium pseudotuberculosis FRC41* isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. BMC Genomics 11(1): 728–745.
7. Connor KM, Quirie MM, Baird G, Donachie W (2000) Characterization of united kingdom isolates of *Corynebacterium pseudotuberculosis* using pulsed-field gel electrophoresis. J Clin.Microbiol 38: 2633–2637.
8. Ben Saïd MS, Ben Maitigue H, Benzarti M, Messadi L, Rejeb A, et al. (2002) Epidemiological and clinical studies of ovine caseous lymphadenitis. Arch Inst Pasteur Tunis 79: 51–57.
9. Binns SH, Bailey M, Green LE (2002) Postal survey of ovine caseous lymphadenitis in the United Kingdom between 1990 and 1999. Vet Rec 150: 263–268.
10. Arsenault J, Girard C, Dubreuil P, Daignault D, Galarneau JR, et al. (2003) Prevalence of and carcass condemnation from maedi-visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. Prev Vet Med 59: 67–81.
11. Paton MW, Walker SB, Rose IR, Watt GF (2003) Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. Aust Vet J 81: 91–95.
12. Pinheiro RR, Gouveia AMG, Alves FSF, Haddad JP (2000) Aspectos epidemiológicos da caprinocultura cearense. Arquivo Brasileiro de Medicina Veterinária e Zootecnia 52: 534–543.
13. Guimarães AS, Seyffert N, Portela RWD, Meyer R, Carmo FB, et al. (2009) Caseous lymphadenitis in sheep flocks of the state of Minas Gerais, Brazil: prevalence and management surveys. Small Ruminants Research 87(1): 86–91.
14. Seyffert N, Guimarães AS, Pacheco LGC, Portela RW, Bastos BL, et al. (2010) High seroprevalence of caseous lymphadenitis in brazilian goat herds revealed by *Corynebacterium pseudotuberculosis* secreted proteins-based ELISA. Res Vet Sci 88: 50–55.
15. Eggleton DG, Middleton HD, Doidge CV, Minty DW (1991) Immunisation against ovine caseous lymphadenitis: comparison of *Corynebacterium pseudotuberculosis* vaccines with and without bacterial cells. Aust Vet J 68: 317–319.
16. Stoops SG, Renshaw HW, Thilsted JP (1984) Ovine caseous lymphadenitis: disease prevalence, lesion distribution, and thoracic manifestations in a population of mature culled sheep from western United States. Am J Vet Res 45(3): 557–61.
17. Ribeiro MG, Júnior JGD, Paes AC, Barbosa PG, Júnior GN, et al. (2001) Punção aspirativa com agulha fina no diagnóstico de *Corynebacterium pseudotuberculosis* na linfadenite caseosa caprina. Arq Inst Biol 68: 23–28.
18. Dorella FA, Estevam EM, Pacheco LGC, Guimarães CT, Lana UGP, et al. (2006) In vivo insertional mutagenesis in *Corynebacterium pseudotuberculosis*: an efficient means to identify DNA sequences encoding exported proteins. Appl Environ Microbiol 72: 7368–7372.

19. Merhej V, Royer-Carenzi M, Pontarotti P, Raoult D (2009) Massive comparative genomic analysis reveals convergent evolution of specialized bacteria. Biol Direct 4: 13–37.
20. Webb SAR, Karleh CM (2008) Bench-to-bedside review: Bacterial virulence and subversion of host defences. Critical Care 12: 234–241.
21. Dobrindt U, Hentschel U, Kaper JB, Hacker J (2002) Genome plasticity in pathogenic and nonpathogenic enterobacteria. Curr Top Microbiol Immunol 264: 157–175.
22. Hall BG, Ehrlich GD, Hu FZ (2010) Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156(4): 1060–8.
23. Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, et al. (2011) Complete genome sequence of *Corynebacterium pseudotuberculosis* I19, a strain isolated from a cow in Israel with bovine mastitis. J Bacteriol 193(1): 323–324.
24. Nakamura Y, Nishio Y, Ikeo K, Gojobori T (2003) The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. Gene 317: 149–155.
25. Tauch A, Kaiser O, Hain T, Goesmann A, Weisshaar B, et al. (2005) Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. J Bacteriol 187: 4671–4682.
26. Khamis A, Raoult D, La Scola B (2004) rpoB gene sequencing for identification of *Corynebacterium* species. J Clin Microbiol 42(9): 3925–31.
27. Khamis A, Raoult D, La Scola B (2005) Comparison between *rpoB* and 16S rRNA gene sequencing for molecular identification of 168 clinical isolates of *Corynebacterium*. J Clin Microbiol 43: 1934–1936.
28. Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2: 414–424.
29. Karaolis DK, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, et al. (1998) A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. Proc Natl Acad Sci U.S.A 95: 3134–3139.
30. Schumann W (2007) Thermosensors in eubacteria: role and evolution. J Biosci 32: 549–557.
31. Hentschel U, Hacker J (2001) Pathogenicity islands: the tip of the iceberg. Microbes Infect 3: 545–548.
32. McNamara PJ, Cuevas WA, Songer JG (1995) Toxic phospholipases D of *Corynebacterium pseudotuberculosis*, *C. ulcerans* and *Arcanobacterium haemolyticum*: cloning and sequence homology. Gene 156: 113–118.
33. Moomaw AS, Maguire ME (2008) The unique nature of Mg2+ channels. Physiology (Bethesda) 23: 275–285.
34. O'Connor K, Fletcher SA, Csonka LN (2009) Increased expression of Mg(2+) transport proteins enhances the survival of *Salmonella enterica* at high temperature. Proc Natl Acad Sci U.S.A 106: 17522–17527.
35. Schönert S, Buder T, Dahl MK (1998) Identification and enzymatic characterization of the maltose-inducible alpha-glucosidase mall (sucrase-isomaltase-maltase) of *Bacillus subtilis*. J Bacteriol 180: 2574–2578.
36. Yamamoto H, Serizawa M, Thompson J, Sekiguchi J (2001) Regulation of the glv operon in *Bacillus subtilis*: yfia (*glvR*) is a positive regulator of the operon that is repressed through *ccpA* and *cre*. Bacteriol 183: 5110–5121.
37. May T, Ito A, Okabe S (2009) Induction of multidrug resistance mechanism in *Escherichia coli* biofilms by interplay between tetracycline and ampicillin resistance genes. Antimicrob Agents Chemother 53: 4628–4639.
38. Smith I (2003) *Mycobacterium tuberculosis* pathogenesis and molecular determinants of virulence. Clin Microbiol Rev 16: 463–496.
39. Saïd-Salim B, Mostowy S, Kristof AS, Behr MA (2006) Mutations in *Mycobacterium tuberculosis* RV0444c, the gene encoding anti-sigK, explain high level expression of *mpb*70 and *mpb*83 in *Mycobacterium bovis*. Mol Microbiol 62: 1251–1263.
40. Veyrier F, Saïd-Salim B, Behr MA (2008) Evolution of the mycobacterial sigK regulon. J Bacteriol 190: 1891–1899.
41. Vassylyev DG, Tomitori H, Kashiwagi K, Morikawa K, Igarashi K (1998) Crystal structure and mutational analysis of the *Escherichia coli* putrescine receptor: structural basis for substrate specificity. J Biol Chem 273: 17604–17609.
42. Carlson PEJ, Horzempa J, O'Dee DM, Robinson CM, Neophytou P, et al. (2009) Global transcriptional response to spermine, a component of the intramacrophage environment, reveals regulation of *Francisella* gene expression through insertion sequence elements. J Bacteriol 191: 6855–6864.
43. Enkavi G, Tajkhorshid E (2010) Simulation of spontaneous substrate binding revealing the binding pathway and mechanism and initial conformational response of *glp*T. Biochemistry 49: 1105–1114.
44. Pérez E, Samper S, Bordas Y, Guilhot C, Gicquel B, et al. (2001) An essential role for *pho*P in *Mycobacterium tuberculosis* virulence. Mol Microbiol 41(1): 179–87.
45. Soto CY, Menéndez MC, Pérez E, Samper S, Gómez AB, et al. (2004) IS*6110* Mediates Increased Transcription of the *phoP* Virulence Gene in a Multidrug-Resistant Clinical Isolate Responsible for Tuberculosis Outbreaks. J Clin Microb 42(1): 212–219.
46. Aguilar D, Infante E, Martin C, Gormley E, Gicquel G, Pando RH (2006) Immunological responses and protective immunity against tuberculosis conferred by vaccination of Balb/C mice with the attenuated *Mycobacterium tuberculosis* (*phoP*) SO2 strain. Clin Exper Immunol 147: 330–338.
47. Gonzalo-Asensio J, Mostowy S, Harders-Westerveen J, Huygen K, Hernández-Pando R, et al. (2008) PhoP: A Missing Piece in the Intricate Puzzle of *Mycobacterium tuberculosis* Virulence. PLoS ONE 3(10): 1–11.
48. Carson SD, Klebba PE, Newton SM, Sparling PF (1999) Ferric enterobactin binding and utilization by *Neisseria gonorrhoeae*. J Bacteriol 181: 2895–2901.
49. Kunkle CA, Schmitt MP (2005) Analysis of a *dtxR*-regulated iron transport and siderophore biosynthesis gene cluster in *Corynebacterium diphtheriae*. J Bacteriol 187: 422–433.
50. Shimizu T, Okabe A, Minami J, Hayashi H (1991) An upstream regulatory sequence stimulates expression of the perfringolysin o gene of *Clostridium perfringens*. Infect Immun 59: 137–142.
51. Urbina P, Flores-Díaz M, Alape-Girón A, Alonso A, Goni FM (2009) Phospholipase C and sphingomyelinase activities of the *Clostridium perfringens* alpha-toxin. Chem Phys Lipids 159: 51–57.
52. Selby T, Allaker RP, Dymock D (2003) Characterization and expression of adjacent proline iminopeptidase and aspartase genes from *Eikenella corrodens*. Oral Microbiol Immunol 18: 256–259.
53. Zhang L, Jia Y, Wang L, Fang R (2007) A proline iminopeptidase gene upregulated in planta by a *lux*R homologue is essential for pathogenicity of *Xanthomonas campestris* pv. campestris. Mol Microbiol 65: 121–136.
54. Böltner D, MacMahon C, Pembroke JT, Strike P, Osborn AM (2002) R391: a conjugative integrating mosaic comprised of phage, plasmid, and transposon elements. J Bacteriol 184: 5158–5169.
55. Burne RA, Chen YY (2000) Bacterial ureases in infectious diseases. Microbes Infect 2: 533–542.
56. Huntley RP, Binns D, Dimmer E, Barrell D, O'Donavan C, et al. (2009) QuickGO: a user tutorial for the web-based Gene Ontology browser. Database 10: 1–19.
57. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, et al. (2009) Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other gram-positive bacteria. Proteomics 9: 61–73.
58. Song C, Kumar A, Saleh M (2009) Bioinformatic comparison of bacterial secretomes. *Genomics Proteomics* Bioinformatics 7: 37–46.
59. Wooldridge L, Lissina A, Cole DK, van den Berg HA, Price DA, Sewell AK (2009) Tricks with tetramers: how to get the most from multimeric peptide-MHC. Immunology 126: 147–164.
60. Wittmann C, Kiefer P, Zelder O (2004) Metabolic Fluxes in *Corynebacterium glutamicum* during Lysine Production with Sucrose as Carbon Source. Applied and Environmental Microbiology 70: 7277–7287.
61. Kabus A (2007) Expression of the *Escherichia coli pntAB* genes encoding a membrane-bound transhydrogenase in *Corynebacterium glutamicum* improves l-lysine formation. Appl Microbiol Biotechnol 75: 47–53.
62. Blombach B, Arndt A, Auchter M, Eikmanns BJ (2009) L-Valine production during growth of pyruvate dehydrogenase complex-deficient *Corynebacterium glutamicum* in the presence of ethanol or by inactivation of the transcriptional regulator SugR. Appl Environ Microbiol 75: 1197–1200.
63. Klaffl S, Eikmanns BJ (2010) Genetic and Functional Analysis of the Soluble Oxaloacetate Decarboxylase from *Corynebacterium glutamicum*. Journal of Bacteriology 192: 2604–2612.
64. Wittmann C, Heinzle E (2001) Modeling and experimental design for metabolic flux analysis of lysine-producing *Corynebacteria* by mass spectrometry. Metab Eng 3(2): 173–91.
65. Parche S, Thomae AW, Schlicht M, Titgemeyer F (2001) *Corynebacterium diphtheriae*: a PTS View to the Genome. J Mol Microbiol Biotechnol 3(3): 415–422.
66. Rübenhagen R, Rönsch H, Jung H, Krämer R, Morbach S (2000) Osmosensor and osmoregulator properties of the betaine carrier *betP* from *Corynebacterium glutamicum* in proteoliposomes. J Biol Chem 275: 735–741.
67. Rittmann D, Schaffer S, Wendisch VF, Sahm H (2003) Fructose-1,6-bisphosphatase from *Corynebacterium glutamicum*: expression and deletion of the *fbp* gene and biochemical characterization of the enzyme. Arch Microbiol 180: 285–292.
68. Kiefer P, Heinzle E, Zelder O, Wittmann Z (2004) Comparative Metabolic Flux Analysis of Lysine-Producing *Corynebacterium glutamicum* Cultured on Glucose or Fructose. Applied and Environmental Microbiology 70: 229–239.
69. Rumbold K, Buijsen HJJ, Overkamp KM, Groenestijn JW, Punt PJ, Werf MJ (2009) Microbial production host selection for converting second-generation feedstocks into bioproducts. Microbial Cell Factories 8: 1–11.
70. Casalino M, Prosseda G, Barbagallo M, Lacobino A, Ceccarini P, et al. (2010) Interference of the *cadC* regulator in the arginine-dependent acid resistance system of *Shigella* and enteroinvasive *E. coli*. Int J Med Microbiol 300(5): 289–95.
71. Alvarez-Ordóñez A, Fernández A, Bernardo A, López M (2010) Arginine and lysine decarboxylases and the acid tolerance response of *Salmonella typhimurium*. Int J Food Microbiol 136: 278–282.
72. Hodgson AL, Carter K, Tachedjian M, Krywult J, Corner LA, et al. (1999) Efficacy of an ovine caseous lymphadenitis vaccine formulated using a genetically inactive form of the *Corynebacterium pseudotuberculosis* Phospholipase D. Vaccine 17: 802–808.
73. D'Afonseca V, Moraes PM, Dorella FA, Pacheco LGC, Meyer R, et al. (2008) A description of genes of *Corynebacterium pseudotuberculosis* useful in diagnostics and vaccine applications. Genet Mol Res 7: 252–260.
74. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409: 1007–1011.
75. Rasko DA, Rosovitz MJ, Garry SA, Emmanuel FM, Fricke WF, et al. (2008) The pan-genome structure of *Escherichia coli:* comparative genomic analysis of *E. coli* commensal and pathogenic isolates. Journal of Bacteriology 190(20): 6881–6893.

76. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V (2006) *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res 37: 201–218.
77. Shiotani H, Yoshioka T, Yamamoto M, Matsumoto R (2008) Susceptibility to citrus canker caused by *Xanthomonas axonopodis* pv. citri depends on the nuclear genome of the host plant. J Gen Plant Pathol 74: 133–137.
78. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8: 186–194.
79. Lagesen K, Hallin P, Rødland EA, Staerfeldt H, Rognes T, et al. (2007) Rnammer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res 35: 3100–3108.
80. Lowe TM, Eddy SR (1997) Trnascan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955–964.
81. Zdobnov EM, Apweiler R (2001) Interproscan--an integration platform for the signature-recognition methods in INTERPRO. Bioinformatics 17: 847–848.
82. Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. (2000) Artemis: sequence visualization and annotation. Bioinformatics 16: 944–945.
83. Waack S, Keller O, Asper R, Brodag T, Damm C, et al. (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models. BMC Bioinformatics 7: 142.
84. Rice P, Longden I, Bleasby A (2000) Emboss: the European molecular biology open software suite. Trends Genet 16: 276–277.
85. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG, et al. (2005) ACT: the Artemis Comparison Tool. Bioinformatics 21: 3422–3423.
86. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, et al. (2007) Mvirdb--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. Nucleic Acids Res 35: D391–394.
87. Karp PD, Paley S, Romero P (2002) The pathway tools software. Bioinformatics 18(1): S225–32.
88. Caspi R, Foerster H, Fulcher CA, Kaipa P, Krummenacker M, et al. (2008) The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res 36: D623–31.
89. Caspi R, Karp PD (2007) Using the metacyc pathway database and the biocyc database collection. Curr Protoc Bioinformatics Chapter 1: Unit1.17.
90. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
91. Krauthammer M, Rzhetsky A, Morozov P, Friedman C (2000) Using BLAST for identifying gene and protein names in journal articles. Gene 259: 245–252.

3.1.2 Sequência completa do genoma de *C. pseudotuberculosis* I19, uma linhagem isolada a partir de um bovino em Israel com mastite bovina

Este trabalho relata a montagem e anotação do genoma de *C. pseudotuberculosis* I19, isolado a partir de um bovino de um plantel em Israel com mastite clínica grave. Para produzirmos a sequência completa do genoma, foi utilizada a abordagem de montagem *de novo* sobre 33 milhões de leituras curtas pareadas (25 pb) geradas pela sequenciador SOLiD.

Esse trabalho mostrou pela primeira vez que é possível realizar uma montagem com uma abordagem híbrida: *de novo* e por referência, utilizando apenas leituras curtas pareadas oriundas do sequenciador SOLiD, versão 2. Esse genoma representou um grande desafio para o nosso grupo de pesquisas porque não haviam trabalhos científicos publicados mostrando como superar os desafios do processo de montagem, como resolução de regiões repetitivas, impostos pelas leituras curtas. Nosso grupo de pesquisa desenvolveu e publicou o método que utiliza a montagem das leituras geradas a partir de diferentes programas, em múltiplas etapas de alinhamento a fim de eliminar *gaps* (Cerdeira e cols., 2011).

Journal of Bacteriology

Complete Genome Sequence of *Corynebacterium pseudotuberculosis* I19, a Strain Isolated from a Cow in Israel with Bovine Mastitis

Artur Silva, Maria Paula C. Schneider, Louise Cerdeira, Maria Silvanira Barbosa, Rommel Thiago J. Ramos, Adriana R. Carneiro, Rodrigo Santos, Marília Lima, Vivian D'Afonseca, Sintia S. Almeida, Anderson R. Santos, Siomar C. Soares, Anne C. Pinto, Amjad Ali, Fernanda A. Dorella, Flavia Rocha, Vinicius Augusto Carvalho de Abreu, Eva Trost, Andreas Tauch, Nahum Shpigel, Anderson Miyoshi and Vasco Azevedo

J. Bacteriol. 2011, 193(1):323. DOI: 10.1128/JB.01211-10.
Published Ahead of Print 29 October 2010.

Updated information and services can be found at:
http://jb.asm.org/content/193/1/323

These include:

REFERENCES

This article cites 11 articles, 5 of which can be accessed free at:
http://jb.asm.org/content/193/1/323#ref-list-1

CONTENT ALERTS

Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://jb.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on November 25, 2011 by guest

Journal of Bacteriology, Jan. 2011, p. 323–324 Vol. 193, No. 1
0021-9193/11/$12.00 doi:10.1128/JB.01211-10
Copyright © 2011, American Society for Microbiology. All Rights Reserved.

Complete Genome Sequence of *Corynebacterium pseudotuberculosis* I19, a Strain Isolated from a Cow in Israel with Bovine Mastitis▽

Artur Silva,[1] Maria Paula C. Schneider,[1] Louise Cerdeira,[1] Maria Silvanira Barbosa,[1] Rommel Thiago J. Ramos,[1] Adriana R. Carneiro,[1] Rodrigo Santos,[1] Marília Lima,[1] Vivian D'Afonseca,[2] Sintia S. Almeida,[2] Anderson R. Santos,[2] Siomar C. Soares,[2] Anne C. Pinto,[2] Amjad Ali,[2] Fernanda A. Dorella,[2] Flavia Rocha,[2] Vinicius Augusto Carvalho de Abreu,[2] Eva Trost,[3] Andreas Tauch,[3] Nahum Shpigel,[4] Anderson Miyoshi,[2] and Vasco Azevedo[2]*

Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, PA, Brazil[1]; Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil[2]; CeBiTec, Universität Bielefeld, 33594 Bielefeld, Germany[3]; and the Koret School of Veterinary Medicine, Hebrew University of Jerusalem, P.O. Box 12, Rehovot 76100, Israel[4]

Received 8 October 2010/Accepted 19 October 2010

This work reports the completion and annotation of the genome sequence of *Corynebacterium pseudotuberculosis* I19, isolated from an Israeli dairy cow with severe clinical mastitis. To present the whole-genome sequence, a *de novo* assembly approach using 33 million short (25-bp) mate-paired SOLiD reads only was applied. Furthermore, the automatic, functional, and manual annotations were attained with the use of several algorithms in a multistep process.

Corynebacterium pseudotuberculosis is the etiology of common disease conditions in sheep, goats, South American camelids, and horses; however, infections in cattle and humans are sporadic and rare.

Based on nitrate reduction, *C. pseudotuberculosis* has two biovars: *C. pseudotuberculosis* bv. *equi*, infecting mainly bovines and equines, and *C. pseudotuberculosis* bv. *ovis*, infecting sheep and goats (1, 2, 8). The widespread occurrence and the economic importance of infection with this pathogen have prompted investigation of its pathogenesis. The use of whole-genome sequence analysis helps to understand the molecular and genetic bases of this bacterium's virulence. Genome sequencing of strains isolated from a human being, a goat, and a sheep was carried out by our team.

Israel is probably the only place in the world to experience large-scale outbreaks of bovine *C. pseudotuberculosis* infection. These outbreaks are also associated with cases of mastitis (9, 10). Strain I19 was isolated from a dairy cow with severe clinical mastitis in two quarters; milk samples from both quarters were positive for *C. pseudotuberculosis*. The cow was culled on the day of milk sampling. In the present research, the SOLiD system was used in sequencing the entire genome of *C. pseudotuberculosis* I19. The sequencing generated 33,368,273 mate-paired 25-nucleotide-long short reads, which is tantamount to 834,206,825 nucleotides of information, rendering a mean genome coverage depth of 321-fold given an expected genome size of 2.6 Mb. The *de novo* assembly strategy for the assembly of short reads in this work combines De Bruijn graph and overlap-layout-consensus methods with the use of a reference genome as a basis for orientation and ordering of the *de novo*-generated contigs (6). This strategy allowed closure of all gaps and an effective coverage of 35-fold.

The genome of *C. pseudotuberculosis* strain I19 consists of a 2,337,730-bp circular chromosome. The average G+C content of the chromosome is 52.84%. The annotation procedure involved the use of several algorithms in a multistep process. For structural annotation, the following software programs were employed: FgenesB, a gene predictor (http://www.softberry.com); RNAmmer, an rRNA predictor (4); tRNAscan-SE, a tRNA predictor (5); and Tandem Repeats Finder, a repetitive-DNA predictor (http://tandem.bu.edu/trf/trf.html). Functional annotation was performed by similarity analyses using public databases and by InterProScan analysis (11). Manual annotation was performed using Artemis (7). Identification and confirmation of putative pseudogenes in the genome were carried out using Consed. Manual analysis was performed based on the Phred quality of each base in the frameshift area (3). This analysis enabled the identification of erroneous insertions or deletions of bases in the genome information produced by the sequencing process and prevented identification of false-positive pseudogenes. The genome of *C. pseudotuberculosis* strain I19 was predicted to contain 2,124 coding sequences (CDSs), 4 rRNA operons, and 50 tRNAs, and 55 pseudogenes were found.

More detailed analysis of this genome and comparative analysis with other sequenced genomes of members of the genus and the same species will provide further insight for understanding virulence and may be useful for the development of new diagnostic methods and vaccines, contributing to the control of the different diseases caused by this pathogen.

* Corresponding author. Mailing address: Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos 6627, Pampulha, CEP 31270-901, Belo Horizonte, MG, Brazil. Phone and fax: 55 31 3409 2610. E-mail: vasco@icb.ufmg.br.

▽ Published ahead of print on 29 October 2010.

Downloaded from http://jb.asm.org/ on November 25, 2011 by guest

Nucleotide sequence accession numbers. The genome and annotation data for strain I19 have been deposited in the NCBI GenBank database (http://www.ncbi.nlm.nih.gov/GenBank/) under accession no. CP002251. Genome sequences of the strains isolated from a human being, a goat, and a sheep have been deposited in the GenBank database under accession no. CP002097.1, CP001809.1, and CP001829.1, respectively.

This research work is the result of collaboration of reputable organizations, with the support of long-standing institutions which include the Rede Paraense de Genômica e Proteômica, supported by the Fundação de Amparo a Pesquisa do Estado do Pará, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and the Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). M.P.C.S., V.A., and A.S. were supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq). We also acknolwedge support from the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

1. **Baird, G. J., and M. C. Fontainey.** 2007. *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis. J. Comp. Pathol. **137:**179–210.
2. **Dorella, F. A., L. G. C. Pacheco, S. C. Oliveira, A. Miyoshi, and V. Azevedo.** 2006. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet. Res. **37:**201–218.
3. **Ewing, B., and P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:**186–194.
4. **Lagesen, K., P. Hallin, E. A. Rødland, H. H. Staerfeldt, T. Rognes, and D. W. Ussery.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108.
5. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.
6. **Miller, J. R., S. Koren, and G. Sutton.** 2010. Assembly algorithms for next-generation sequencing data. Genomics **95:**315–317.
7. **Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16**(10)**:**944–945.
8. **Williamson, L. H.** 2001. Caseous lymphadenitis in small ruminants. Vet. Clin. North Am. Food Anim. Pract. **17**(2)**:**359–371.
9. **Yeruham, I., D. Elad, M. Van Ham, N. Y. Shpigel, and S. Perl.** 1997. *Corynebacterium pseudotuberculosis* infection in Israeli cattle: clinical and epidemiological studies. Vet. Rec. **140:**423–427.
10. **Yeruham, I., D. Elad, S. Friedman, and S. Perl.** 2003. *Corynebacterium pseudotuberculosis* infection in Israeli dairy cattle. Epidemiol. Infect. **131:** 947–955.
11. **Zdobnov, E. M., and R. Apweiler.** 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**(9)**:**847–848.

Downloaded from http://jb.asm.org/ on November 25, 2011 by guest

3.1.3 Sequência completa do genoma da *C. pseudotuberculosis,* linhagem PAT10 isolada de uma ovelha na Patagônia, Argentina

Neste trabalho, relatou-se a sequência completa do genoma da *C. pseudotuberculosis*, linhagem PAT10, isolada a partir de um abscesso pulmonar de uma ovelha de um plantel na Patagônia (Argentina), como resultado de uma investigação sobre a patogênese.

Este foi o segundo genoma completo montado e finalizado apenas com leituras curtas pareadas (25 pb) geradas pela sequenciador SOLiD, versão 2 (Cerdeira e cols., 2011).

Journal of Bacteriology

Whole-Genome Sequence of Corynebacterium pseudotuberculosis PAT10 Strain Isolated from Sheep in Patagonia, Argentina

Louise Teixeira Cerdeira, Anne Cybelle Pinto, Maria Paula Cruz Schneider, Sintia Silva de Almeida, Anderson Rodrigues dos Santos, Eudes Guilherme Vieira Barbosa, Amjad Ali, Maria Silvanira Barbosa, Adriana Ribeiro Carneiro, Rommel Thiago Jucá Ramos, Rodrigo Santos de Oliveira, Debmalya Barh, Neha Barve, Vasudeo Zambare, Silvia Estevão Belchior, Luis Carlos Guimarães, Siomar de Castro Soares, Fernanda Alves Dorella, Flavia Souza Rocha, Vinicius Augusto Carvalho de Abreu, Andreas Tauch, Eva Trost, Anderson Miyoshi, Vasco Azevedo and Artur Silva

J. Bacteriol. **2011, 193(22):6420. DOI: 10.1128/JB.06044-11.**

Updated information and services can be found at:
http://jb.asm.org/content/193/22/6420

These include:

REFERENCES

This article cites 8 articles, 3 of which can be accessed free at:
http://jb.asm.org/content/193/22/6420#ref-list-1

CONTENT ALERTS

Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://jb.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on November 25, 2011 by guest

Journal of Bacteriology, Nov. 2011, p. 6420–6421 Vol. 193, No. 22
0021-9193/11/$12.00 doi:10.1128/JB.06044-11
Copyright © 2011, American Society for Microbiology. All Rights Reserved.

Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* PAT10 Strain Isolated from Sheep in Patagonia, Argentina

Louise Teixeira Cerdeira,[1] Anne Cybelle Pinto,[2] Maria Paula Cruz Schneider,[1] Sintia Silva de Almeida,[2] Anderson Rodrigues dos Santos,[2] Eudes Guilherme Vieira Barbosa,[2] Amjad Ali,[2] Maria Silvanira Barbosa,[1] Adriana Ribeiro Carneiro,[1] Rommel Thiago Jucá Ramos,[1] Rodrigo Santos de Oliveira,[1] Debmalya Barh,[4] Neha Barve,[4] Vasudeo Zambare,[4,5] Silvia Estevão Belchior,[3] Luis Carlos Guimarães,[2] Siomar de Castro Soares,[2] Fernanda Alves Dorella,[2] Flavia Souza Rocha,[2] Vinicius Augusto Carvalho de Abreu,[2] Andreas Tauch,[6] Eva Trost,[6] Anderson Miyoshi,[2] Vasco Azevedo,[2] and Artur Silva[1]*

Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil[1]; Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil[2]; Universidad Nacional de la Patagonia San Juan Bosco, Chubut, Argentina[3]; Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India[4]; Center for Bioprocessing Research and Development (CBRD), South Dakota School of Mines and Technology, Rapid City, South Dakota[5]; and CeBiTec, University of Bielefeld, 33594 Bielefeld, Germany[6]

Received 20 August 2011/Accepted 26 August 2011

In this work, we report the complete genome sequence of a *Corynebacterium pseudotuberculosis* PAT10 isolate, collected from a lung abscess in an Argentine sheep in Patagonia, whose pathogen also required an investigation of its pathogenesis. Thus, the analysis of the genome sequence offers a means to better understanding of the molecular and genetic basis of virulence of this bacterium.

The incidence of caseous lymphadenitis (CLA) is high in many regions of the world, resulting in huge and significant economic losses in agribusinesses, as it is responsible for a decrease in wool production and carcass quality (3). The disease is endemic in flocks in the provinces of Chubut and Santa Cruz in the Southern Patagonia region of Argentina, thereby leading to an outrageous prevalence rate of about 70% within individual flocks in Patagonia (4). Diseases caused by *Corynebacterium pseudotuberculosis* present in various clinical forms, and sheep and goats are affected by CLA (1). Analysis of the genome sequence improves our understanding of the molecular and genetic basis of the virulence of the bacterium. We hereby report the whole-genome sequence of *C. pseudotuberculosis* PAT10 as determined using the SOLiD platform. In total, we generated 27,858,221 mate-paired short reads (25 bp) of usable sequences (296-fold coverage). Furthermore, a hybrid *de novo* assembly approach was applied using 16,885,903 short (25-bp) mate-paired SOLiD filtered reads; that strategy allowed close gaps without a bench work time cost (2). The automatic and manual annotations were done using several algorithms in a multistep process.

For structural annotation, the following software was used: FgenesB (gene predictor); RNAmmer (rRNA predictor) (5); tRNA-scan-SE (tRNA predictor) (6); and Tandem Repeat Finder (repetitive DNA predictor) (http://tandem.bu.edu/trf/trf.html). Functional annotation was performed using similarity analyses and public databases and by InterProScan analysis (8). Manual annotation was performed using Artemis software (7). Identification and confirmation of pseudogenes in the genome were carried out using CLCBio Workbench 4.0.2 software. Manual analysis was performed based on the Phred quality of each base combined with analysis of the depth of coverage of the frameshift region. That analysis allowed the identification of false-positive pseudogene results. The genome of the PAT10 strain consists of a 2,335,323-bp circular chromosome, and the average G+C content of the chromosome is 52.19%. The genome was predicted to contain 2,079 coding sequences (CDS), four rRNA operons, 49 tRNAs, and 61 pseudogenes.

The characterization of the PAT10 genome should identify and unravel the mechanisms of virulence of this pathogen through comparative analyses performed with other sequenced genomes of the genus and the same species, thereby allowing the development of new diagnostics kits and vaccines.

Nucleotide sequence accession number. The genome sequence obtained in this study has been deposited in the GenBank database under accession number CP002924 (chromosome).

This work is part of the Rede Paraense de Genômica e Proteômica supported by FAPESPA CNPq, CAPES, and FAPEMIG.

REFERENCES

1. **Barakat, A. A., S. A. Sekim, A. Atef, M. S. Saber, and E. K. Nafie.** 1984. Two serotypes of *Corynebacterium pseudotuberculosis* isolated from different animal species. Rev. Sci. Tech. Off. Int. Epiz.: **3:**151–163.
2. **Cerdeira, L. T., et al.** 2011. Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. J. Microbiol. Methods **86:**218–223.
3. **Dorella, F. A., L. G. C. Pacheco, S. C. Oliveira, A. Miyoshi, and V. Azevedo.** 2006. Corynebacterium pseudotuberculosis: microbiology, biochemical

* Corresponding author. Mailing address: Instituto de Ciências Biológicas, Universidade Federal do Pará, Av. Augusto Corrêa 01, Guamá, CEP 66075-110, Belém, PA, Brazil. Phone and fax: 55 91 3201-8426. E-mail: asilva@ufpa.br.

Downloaded from http://jb.asm.org/ on November 25, 2011 by guest

properties, pathogenesis and molecular studies of virulence. Vet. Res. **37:**201–218.

4. **Estevao-Belchior, S., A. Gallardo, A. Abalos, N. Jodor, and D. Jensen.** 2006. Actualización sobre linfoadenitis caseosa: el agente etiológico y la enfermedad. Ve. Argent. **23:**258–278.
5. **Lagesen, K., et al.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108.
6. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.
7. **Rutherford, K., et al.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945.
8. **Zdobnov, E. M., and R. Apweiler.** 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:**847–848.

Downloaded from http://jb.asm.org/ on November 25, 2011 by guest

3.1.4 Sequência completa do genoma de *C. pseudotuberculosis* linhagem CIP 52.97, isolada de um cavalo no Quênia

Neste trabalho, relatou-se a sequência completa do genoma de *C. pseudotuberculosis*, biovar equi, linhagem CIP 52.97 (Coleção do Instituto Pasteur), isolado em 1952 a partir de um caso de linfangite ulcerativa em um cavalo queniano.

Este foi o primeiro genoma do biovar equi (tipo II) depositado em bancos de dados públicos, e esta é a primeira publicação de um genoma desse biovar. Forneceu pistas a respeito das diferenças entre os biovares no tocante à capacidade de redução do nitrito em nitrato, bem como pistas sobre a preferência do patógeno em relação ao hospedeiro e o desenvolvimento da doença.

O genoma da linhagem CIP 52.97 foi o terceiro montado apenas com leituras curtas pareadas (25 pb) geradas pela sequenciador SOLiD, versão 2 (Cerdeira e cols., 2011).

Journal of Bacteriology

Complete Genome Sequence of Corynebacterium pseudotuberculosis Strain CIP 52.97, Isolated from a Horse in Kenya

Louise Teixeira Cerdeira, Maria Paula Cruz Schneider, Anne Cybelle Pinto, Sintia Silva de Almeida, Anderson Rodrigues dos Santos, Eudes Guilherme Vieira Barbosa, Amjad Ali, Flávia Figueira Aburjaile, Vinicius Augusto Carvalho de Abreu, Luis Carlos Guimarães, Siomar de Castro Soares, Fernanda Alves Dorella, Flávia Souza Rocha, Erick Bol, Pablo Henrique Caracciolo Gomes de Sá, Thiago Souza Lopes, Maria Silvanira Barbosa, Adriana Ribeiro Carneiro, Rommel Thiago Jucá Ramos, Nilson Antônio da Rocha Coimbra, Alex Ranieri Jerônimo Lima, Debmalya Barh, Neha Jain, Sandeep Tiwari, Rathiram Raja, Vasudeo Zambare, Preetam Ghosh, Eva Trost, Andreas Tauch, Anderson Miyoshi, Vasco Azevedo and Artur Silva

J. Bacteriol. **2011, 193(24):7025. DOI: 10.1128/JB.06293-11.**

Updated information and services can be found at:
http://jb.asm.org/content/193/24/7025

These include:

REFERENCES

This article cites 8 articles, 3 of which can be accessed free at:
http://jb.asm.org/content/193/24/7025#ref-list-1

CONTENT ALERTS

Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: **http://jb.asm.org/site/misc/reprints.xhtml**
To subscribe to to another ASM Journal go to: **http://journals.asm.org/site/subscriptions/**

Journals.ASM.org

Downloaded from http://jb.asm.org/ on November 28, 2011 by guest

Journal of Bacteriology, Dec. 2011, p. 7025–7026 Vol. 193, No. 24
0021-9193/11/$12.00 doi:10.1128/JB.06293-11
Copyright © 2011, American Society for Microbiology. All Rights Reserved.

Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain CIP 52.97, Isolated from a Horse in Kenya

Louise Teixeira Cerdeira,[1] Maria Paula Cruz Schneider,[1] Anne Cybelle Pinto,[2] Sintia Silva de Almeida,[2] Anderson Rodrigues dos Santos,[2] Eudes Guilherme Vieira Barbosa,[2] Amjad Ali,[2] Flávia Figueira Aburjaile,[2] Vinicius Augusto Carvalho de Abreu,[2] Luis Carlos Guimarães,[2] Siomar de Castro Soares,[2] Fernanda Alves Dorella,[2] Flávia Souza Rocha,[2] Erick Bol,[1] Pablo Henrique Caracciolo Gomes de Sá,[1] Thiago Souza Lopes,[1] Maria Silvanira Barbosa,[1] Adriana Ribeiro Carneiro,[1] Rommel Thiago Jucá Ramos,[1] Nilson Antônio da Rocha Coimbra,[1] Alex Ranieri Jerônimo Lima,[1] Debmalya Barh,[3] Neha Jain,[3] Sandeep Tiwari,[3] Rathiram Raja,[3] Vasudeo Zambare,[3,4] Preetam Ghosh,[3,6] Eva Trost,[5] Andreas Tauch,[5] Anderson Miyoshi,[2] Vasco Azevedo,[2] and Artur Silva[1]*

Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém, Brazil[1]; Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil[2]; Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur, West Bengal, India[3]; Center for Bioprocessing Research and Development (CBRD), South Dakota School of Mines and Technology, South Dakota[4]; Institute for Genome Research and Systems Biology, Center for Biotechnology, Germany Institute for Genome Research, Bielefeld University, Bielefeld, Germany[5]; and Department of Computer Science and Center for the Study of Biological Complexity, Virginia Commonwealth University, 401 West Main Street, Room E4234, P.O. Box 843019, Richmond, Virginia 23284-3019[6]

Received 30 September 2011/Accepted 10 October 2011

In this work, we report the whole-genome sequence of *Corynebacterium pseudotuberculosis* bv. equi strain CIP 52.97 (Collection Institut Pasteur), isolated in 1952 from a case of ulcerative lymphangitis in a Kenyan horse, which has evidently caused significant losses to agribusiness. Therefore, obtaining this genome will allow the detection of important targets for postgenomic studies, with the aim of minimizing problems caused by this microorganism.

Corynebacterium pseudotuberculosis is an intracellular pathogen which causes significant losses in the goat, sheep, horse, and cattle breeding industries, since infected animals demonstrate wounds on the skin and internal organs, causing damage to the pelt and even the flesh (4). This bacterium is classified into two biovars, ovis (type I) and equi (type II). The biochemical difference between the types is the ability to reduce nitrate. *C. pseudotuberculosis* bv. ovis is negative for nitrate reduction, while *C. pseudotuberculosis* bv. equi is positive. CIP 52.97, the strain described in this paper, belongs to biovar type II. Ulcerative lymphangitis (also known as chest abscess, pigeon breast, and pigeon fever) is one of the most common and economically threatening infectious diseases of young-adult horses of all breeds and both sexes in California (3). In equine ulcerative lymphangitis, there are two forms of the disease, one characterized by external abscesses and one which affects the internal organs (1). To best understand the molecular and genetic basis of virulence of this bacterium, it was necessary to perform sequencing and genome analysis by using the SOLiD platform. We generated 60,342,023 mate-paired short reads (25 bp), with 580-fold coverage. The assembly process was based on the strategy of Cerdeira et al. (2), which allowed us to close gaps without the bench work time. The structural annotation was done automatically by a multipronged approach using the following programs: for gene prediction, FgenesB (http://www.softberry.com); for rRNA prediction, RNAmmer (5); for tRNA prediction, tRNA-scan-SE (6); and for repetitive DNA prediction, Tandem Repeats Finder (http://tandem.bu.edu/trf/trf.html). Protein domains and motifs were determined by InterProScan analysis (8). Manual curation was achieved using Artemis (7). Identification and confirmation of pseudogene in the genome were carried out using CLCBio Workbench 4.0.2. The genome of CIP 52.97 consists of a 2,320,595-bp circular chromosome, and the average G+C content of the chromosome is 52,14%. The genome was predicted to contain 2,057 coding sequences, four rRNA operons, 47 tRNAs, and 78 pseudogenes.

Nucleotide sequence accession number. The genome sequence obtained in this study has been deposited in the GenBank database under the accession number CP003061 (chromosome).

* Corresponding author. Mailing address: Instituto de Ciências Biológicas. Universidade Federal do Pará. Av. Augusto Corrêa 01, Guamá, CEP 66075-110. Belém, PA, Brazil. Phone and fax: 55 91 3201-8426. E-mail: asilva@ufpa.br.

Downloaded from http://jb.asm.org/ on November 28, 2011 by guest

This work was part of the Rede Paraense de Genômica e Proteômica, supported by FAPESPA, Núcleo Amazônico de Excelência em Genômica de Microorganismos—Pronex/CNPq/FAPESPA, CAPES, and FAPEMIG.

REFERENCES

1. **Aleman, M., S. J. Spier, W. D. Wilson, and M. Doherr.** 1996. Corynebacterium pseudotuberculosis infection in horses: 538 cases (1982–1993). J. Am. Vet. Med. Assoc. **209:**804.
2. **Cerdeira, L. T., et al.** 2011. Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. J. Microbiol. Methods **86:**218–223.
3. **Doherr, M. G., T. E. Carpenter, K. M. Hanson, W. D. Wilson, and I. A. Gardner.** 1998. Risk factors associated with Corynebacterium pseudotuberculosis infection in California horses. Prev. Vet. Med. **35:**229–239.
4. **Dorella, F. A., L. G. C. Pacheco, S. C. Oliveira, A. Miyoshi, and V. Azevedo.** 2006. Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet. Res. **37:**201–218.
5. **Lagesen, K., et al.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. **35:**3100–3108.
6. **Lowe, T. M., and S. R. Eddy.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:**955–964.
7. **Rutherford, K., et al.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945.
8. **Zdobnov, E. M., and R. Apweiler.** 2001. InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:**847–848.

Downloaded from http://jb.asm.org/ on November 28, 2011 by guest

3.1.5 Reanotação do genoma da *Corynebacterium diphteriae* NCTC13129 como uma nova abordagem para o estudo de genes ligados à virulência e patogenicidade

A reanotação de genomas é uma abordagem para a descoberta de novos elementos genéticos. Também contribui para tornar os genomas mais descritivos e atuais, com características relevantes a respeito do fenótipo de um organismo. O genoma da *C. diphtheriae* foi depositado e publicado no ano de 2003 (Cerdeño-Tárraga e cols., 2003). O presente estudo teve por objetivo a reanotação do genoma da *C. diphtheriae*, patógeno humano Gram-positivo e causador da difteria. O depósito de aproximadamente 15 genomas do gênero *Corynebacterium* facilitou a atualização do genoma deste microrganismo. Além disso, o surgimento da doença invasiva causada por linhagens de *C. diphtheriae* atóxicas e o ressurgimento da difteria em populações parcialmente vacinadas, impulsionaram estudos a respeito de seu genoma estrutural e funcional.

Em relação à genômica estrutural, 23 regiões codificantes foram excluídas e 71 novos genes foram adicionados à anotação. No entanto, todos os pseudogenes foram validados e dez pseudogenes novos foram propostos. No tocante à anotação funcional, cerca de 57% da anotação gênica foi atualizada e tornou-se mais informativa. A descrição de 41% dos produtos (973 proteínas) foi atualizada, incluindo 370 produtos que anteriormente eram anotadas como "proteínas hipotéticas". A partir da anotação atualizada, a plasticidade do genoma tornou-se evidente, mostrando melhorias quanto a informação de 13 ilhas de patogenicidade descritas na literatura. Além disso, o grande número de transposases e a presença de genes estruturais de bacteriófagos também reforçaram essa plasticidade. Contrastando esta realidade, permitiu-se o esclarecimento de mecanismos utilizados pela *C. diphtheriae* para interromper a invasão de bacteriófagos.

A reanotação do genoma da *C. diphtheriae* melhorou o entendimento desse genoma sob o aspecto da plasticidade, caracterizando prováveis fatores de virulência. Esse trabalho será utilizado no projeto do pangenoma da *C. diphtheriae* que está sendo desenvolvido pelo nosso grupo de pesquisa e o grupo do professor Andreas Tauch em Bielefield, Alemanha. Além disso, o protocolo utilizado aqui pode ser estendido a outros patógenos, melhorando a informação de genomas depositados em bases de dados públicos e minimizando a propagação de erros de anotação.

Open Access Bioinformatics

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Reannotation of the *Corynebacterium diphtheriae* NCTC13129 genome as a new approach to studying gene targets connected to virulence and pathogenicity in diphtheria

This article was published in the following Dove Press journal:
Open Access Bioinformatics
2 February 2012
Number of times this article has been viewed

Vívian D'Afonseca[1,*]
Siomar C Soares[1,*]
Amjad Ali[1]
Anderson R Santos[1]
Anne C Pinto[1]
Aryane AC Magalhães[1]
Cássio de Jesus Faria[1]
Eudes Barbosa[1]
Luis C Guimarães[1]
Marcus Eslabão[2]
Sintia S Almeida[1]
Vinicius AC Abreu[1]
Adhemar Zerlotini[3,4]
Adriana R Carneiro[5]
Louise T Cerdeira[5]
Rommel TJ Ramos[5]
Raphael Hirata Jr[6]
Ana L Mattos-Guaraldi[6]
Eva Trost[7]
Andreas Tauch[7]
Artur Silva[5]
Maria P Schneider[5]
Anderson Miyoshi[1]
Vasco Azevedo[1]

[1]Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil; [2]Federal University of Pelotas, Pelotas, Rio Grande do Sul, Brazil; [3]FIOCRUZ - CEBIO, Belo Horizonte, Minas Gerais, Brazil; [4]EMBRAPA - CNPTIA, Campinas, São Paulo, Brazil; [5]Federal University of Pará, Belém, Pará, Brazil; [6]Rio de Janeiro State University, Rio de Janeiro, Brazil; [7]Center for Biotechnology, Bielefeld University, Bielefeld, Germany

*These authors contributed equally to this work

Correspondence: Vasco Azevedo
Federal University of Minas Gerais,
Belo Horizonte, 31907-270,
Minas Gerais, Brazil
Tel +55 31 3409 2610
Fax +55 31 3409 2610
Email vasco@icb.ufmg.br

Background: The reannotation of genomes already on file is a new approach to discovering new genetic elements and to make the genomes more descriptive and current with relevant features regarding the organism's lifestyle. Within this approach, the present study aimed to reannotate the genome of the Gram-positive human pathogen *Corynebacterium diphtheriae*, which causes diphtheria. The deposit of massive amounts of information linked to other species of the genus *Corynebacterium* has facilitated the updating of the genomic interpretation of this microorganism. Additionally, the emergence of invasive disease by nontoxigenic strains of *C. diphtheriae* and the reemergence of diphtheria in partially immunized populations have given impetus to new studies in relation to its structural and functional genome.

Results: In relation to structural genomics, 23 coding regions (coding sequences) were deleted and 71 new genes were added to the genome annotation. Nevertheless, all the pseudogenes were validated and ten new pseudogenes were created. In relation to functional genomics, about 57% of the genome annotation was updated and became functionally more informative. The product descriptions of 41% (973 proteins) were updated. Among them, 370 that were previously annotated as "hypothetical proteins," now have more informative descriptions. With the new annotation, the plasticity of the genome became evident, which shows improvements in the annotation of 13 pathogenicity islands already described in the literature. In addition, the large number of transposases and the presence of structural genes of bacteriophages make their genomic versatility evident. Contrasting with this reality, it also allowed the clarification of some aspects concerned with mechanisms used by *C. diphtheriae* to stop the invasion of the genome by bacteriophages, mediated by the clustered regularly interspaced short palindromic repeats region.

Conclusion: The reannotation of the *C. diphtheriae* genome provided an improvement in annotation of the *C. diphtheriae* genome in several aspects, such as virulence characteristics and plasticity events. Moreover, the protocol used here can be extended to various other pathogens in order to improve the genomic information already on file in public databases and to minimize propagating errors. The reannotated archive and updated archive are available at: http://lgcm.icb.ufmg.br/pub/C_diphtheriae_reannotation.embl.

Keywords: *Corynebacterium diphtheriae*, diphtheria, reannotation, CRISPR, pathogenicity islands, genome

Background

In recent years, genomics has regained its foothold in the areas of science that are in full development. With the advent of new sequencing platforms, known as the next generation, the amount of genomic data available in public databases has increased exponentially.[1]

submit your manuscript | www.dovepress.com
Dovepress
http://dx.doi.org/10.2147/OAB.S25500

© 2012 D'Afonseca et al, publisher and licensee Dove Medical Press Ltd. This is an Open Access article which permits unrestricted noncommercial use, provided the original work is properly cited.

This is due to the fact that, currently, the acquisition of genomic data happens in a rapid, efficient, accurate, and low-cost manner. Research groups may then begin projects with their favorite organisms.[2] The reflection of this expansion may be seen in the database Genomes OnLine Database v 3 (http://www.genomesonline.org). There are currently around 10,420 genome projects in progress, and approximately 1700 genomes have been completed and published.

Meanwhile, on the one side, the massive generation of genomic data is good for science on the whole; on the other side, it has brought about the propagation of errors from the annotation where genomes, annotated automatically and without physical oversight, are deposited daily in public domain databases. Connected to this, many genome annotations deposited years ago were not updated, thus worsening this scenario.[3,4] As one approach to improving this panorama and minimizing the propagation of errors, some groups are already undertaking the process called reannotation, in which a genome already deposited passes through a new process of the prediction of genes and other structural elements of the genome; afterwards, they are reviewed manually by a specialist on the organism, and every open reading frame (ORF) has its product reannotated with the aim of improving its description.[2]

Few organisms were reannotated until today.[4–8] However, the reported improvement and the description of new genomic elements have motivated the practice of this new approach. In *Escherichia coli* CFT073, the update allowed the identification of 299 new ORFs, including various classical elements of virulence present in pathogenicity islands (PAIs), which were not predicted in the first version of genome annotation.[5] In the *Campylobacter jejuni* NCTC11168 pathogen, various pseudogenes have been identified, and around 97.8% of the previous genome experienced some type of update, including a change in the description of the gene product, the gene symbol, and new description for hypothetical proteins.[6] Both of the latter cited studies undertook the updating of the genome annotations 7 years after they were published for the first time.

Based on this approach, the present study attempted to reannotate the genome of *Corynebacterium diphtheriae* biotype *gravis*, strain NCTC13129. The sequencing and subsequent availability of the *C. diphtheriae* NCTC13129 genome in public domain databases occurred in 2003, under accession number NC_002935, contributing to the understanding of the pathogenicity, virulence, and lifestyle of this pathogen.[9]

C. diphtheriae is a Gram-positive human pathogen and has a high guanine-cytosine content.[9] This pathogen has the ability to colonize the human respiratory tract and, through the action of its exotoxin, diphtheria toxin, forms a membranous exudate over the tonsils, pharynx, and/or nasal cavity.[10] Diphtheria was under control for decades, but it is currently among the reemerging diseases. More than 150,000 cases and 5000 deaths were reported from diphtheria from 1990–1999 in the Eastern European region. This number is in great contrast to reports from the previous decade, which did not surpass 600 cases, according to the World Health Organization. In addition to the European continent, other continents such as Asia, Africa, and South America have also reported significant numbers of diphtheria cases. After a huge effort from the health organizations to contain the disease, the immunization coverage has reached approximately 82% of the world population in 2009, decreasing the number of reported cases to 857 in the same year (World Health Organization). However, the identification of nontoxigenic *C. diphtheriae* strains, ie, strains unable to produce the diphtheria toxin, which have caused invasive diseases, such as endocarditis,[11,12] has to be treated as a new potential problem in public health.

Therefore, it has become highly relevant to improve and update the already existing data about this pathogen, with the goal of increasing genetic knowledge linked to its genome and to propose new approaches and precise diagnostics that will prevent or minimize the effects of its resurgence. The reannotated file can be downloaded freely through the authors' server, at: http://lgcm.icb.ufmg.br/pub/C_diphtheriae_reannotation.embl. Alternatively, it can also be downloaded through a public server: http://www.bioinformatics.org/groups/?group_id=1103.

Methods

Genome reannotation

The reannotation procedures involved the use of several algorithms, in a multi-step process. Structural annotation was performed using the following software: FGENESB: bacterial operon and gene predictor (http://www.softberry.com; Softberry, Inc, Mount Kisco, NY); RNAmmer: ribosomal ribonucleic acid (RNA) predictor (Center for Biological Sequence Analysis, Lyngby, Denmark);[13] tRNA-scan-SE: transfer RNA predictor (Lowe Lab, Biomolecular Engineering, University of California, Santa Cruz, CA);[14] and Tandem Repeat Finder: repetitive deoxyribonucleic acid (DNA) predictor (Boston University, Boston, MA).[15]

Functional annotation was performed by similarity analyses, using Basic Local Alignment Search Tool (protein) – National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov; Bethesda, MD) with a cutoff of 10^{-6} against a nonredundant database of proteins, InterProScan (European Bioinformatics Institute, Hinxton, Cambridgeshire, UK) and SignalP (Center for Biological Sequence Analysis) analysis.[16] Manual annotation was performed using Artemis (Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, UK).[17]

Criteria for manual curation

To improve the annotation, all coding sequences (CDSs) were manually curated. The correction of the initial methionine was guided by the presence of a signal peptide, and matched with homologous proteins of related organisms. The hits generated in a similarities search, with a minimum identity of 60% and the presence of the same results in almost all hits, were used to update the predicted products.

To improve the annotation of either hypothetical proteins or proteins without available product descriptions, large predicted conserved domains linked to CDS were used, when available.

Subcellular location of predicted proteins and gene targets for vaccine development

Predictions of the cellular locations of *Corynebacterium* proteins were made using the "subcellular localization" option of the software Vaxign. Classification of predicted proteins was done using "Dynamic Vaxign Analysis" of the Vaxign software (University of Michigan Medical School, Ann Arbor, MI),[18] in secreted and cell wall categories of subcellular location. Additionally, the software searched for MHC classes I and II binding proteins, transmembrane helices, and adhesion probability.

In silico identification of PAIs

In order to identify and classify accurately the putative PAIs in the corynebacterial genomes, the authors developed a combined computational approach using several in-house scripts to integrate the prediction of diverse algorithms and databases (http://www.genoma.ufpa.br/lgcm/pips). The algorithms and databases were: Colombo – SIGI-HMM (Institute of Computer Science, University of Göttingen, Göttingen, Germany),[19] Artemis,[17] tRNAscan-SE,[14] HMMER (v 3.0; Howard Hughes Medical Institute, Chevy Chase, MD),[20] Artemis Comparison Tool (Wellcome Trust Sanger Institute),[21] and mVIRdb (Lawrence Livermore National Laboratory, Livermore, CA).[22]

In silico metabolic pathway construction

The two main data sources used for reconstructing the *C. diphtheriae* metabolic pathways were the genome sequence file in FASTA format, and the genome annotation file in GenBank format. Metabolic pathway databases for strains 1002 and C231 were created using the Pathway Tools 13 software, developed by SRI International (Menlo Park, CA).[23] The Pathway Tools software contains algorithms that predict the metabolic pathways of an organism from its genome, by comparison to a reference pathways database known as MetaCyc.[24]

Construction of a metabolic pathways database was done, using BioCyc,[25] in order to compare the pathways of the bacteria *C. pseudotuberculosis* I19, *C. efficiens* YS-314, *C. glutamicum* ATCC 13032, and *C. jeikeium* K411 to the deduced *C. diphtheriae* pathways.

Results

Improving of the *C. diphtheriae* NCTC13129 genome annotation

C. diphtheriae NCTC13129 genome remained annotated without changes for 7 years. Today, the vast genomic information present in the databases allows this scenario to be altered.

Presently, the complete reannotation and updating of the *C. diphtheriae* genome annotation allowed its modification, in its various structural and functional aspects, of which approximately 57% of the prior genome annotation has undergone alteration, making it more descriptive. Additionally, this process assisted in the discovery of new genetic elements, which can provide us with new understanding about the virulence and plasticity of the microorganism. Based on this information, the updated genome annotation shows 2368 genes in contrast to the previous version which had 2320. Within this new tally, 23 CDSs of *C. diphtheria*e were deleted and 71 new CDSs were predicted and validated, as shown in Table 1. The new CDSs and pseudogenes, along with all the predicted PAIs, are represented in Figure 1. For more information about the similarity between the new CDSs and pseudogenes with proteins of the nonredundant database of proteins from NCBI, please refer to Additional File 1.

The criterion for the deletion of the CDSs was their use in the formation of new pseudogenes or the absence of biological evidence. In addition, in three cases (DIP0404, DIP0700, and DIP1975), the new prediction of CDS was done in the DNA strand opposite the predicted CDS in the genome deposited in NCBI Reference Sequence. These new predictions presented biological evidence with strong similarity to other species of the genus *Corynebacterium*, in contrast to the three deleted CDSs that were ORFans or showed meaningless matches in the protein databases. Figure 2 illustrates the reannotated

Table 1 Coding sequences deleted and/or modified from the previous version of the genome annotation (Below are the new coding sequences of the updated *Corynebacterium diphtheriae* genome)

Gene ID Cd* RefSeq	Product	New prediction	Begin	End	Strand	Product	Status
DIP0017	Hypothetical protein	DIP0016A	19709	20107	R	Hypothetical membrane protein	Match with *Corynebacterium* species
DIP0018/DIP0019/DIP0020	Hypothetical protein	DIP0020	20341	21185	R	Putative glycosylase (pseudogene)	Match with *Corynebacterium* species
DIP0039	Hypothetical protein	–	–	–	–	–	Overlap with CRISPR region
DIP0040	Hypothetical protein	–	–	–	–	–	Overlap with CRISPR region
DIP0142/DIP0143	Hypothetical protein	DIP0143	118306	119988	F	Transposase (pseudogene)	Pseudogene increased
DIP0239/DIP0240/DIP0241	Hypothetical protein	DIP0239	205490	206485	R	Putative surface-anchored protein (pseudogene)	Pseudogene increased
DIP0404	Hypothetical protein	DIP0403a	369731	370213	F	Putative membrane protein	Match with *Corynebacterium* species
DIP0700	Hypothetical protein	DIP0699a	676852	677295	R	Conserved hypothetical protein	Match with *Corynebacterium* species
DIP0734/DIP0735	Putative membrane protein	DIP0734	711318	712185	R	Putative sodium/glutamate symporter	Pseudogene increased
DIP0757/DIP0757A	IS element transposase (partial)	DIP0757	736001	736487	R	IS element transposase (partial)	Pseudogene increased
DIP0898/DIP0899	Hypothetical protein	DIP0899	868809	869972	F	HNH endonuclease domain protein	Pseudogene increased
DIP1106/DIP1110	Conserved hypothetical protein (pseudogene)	DIP1106	1093117	1094859	F	Putative signal-transduction protein containing cAMP-binding	Pseudogene increased
DIP1654/DIP1655	Conserved hypothetical protein	DIP1654	1689384	1690688	R	LGFP repeat superfamily protein	Pseudogene increased
DIP1820	Putative membrane protein	DIP1819A	1867528	1867842	R	Hypothetical protein	Match with *Corynebacterium* species
DIP1975	Hypothetical protein	DIP1974A	2022366	2022947	F	Lipoprotein LpqE	Match with *Corynebacterium* species
DIP2023/DIP2024	Hypothetical protein	DIP2023	2076614	2077463	F	Filamentation induced by cAMP protein	Match with *Corynebacterium* species
DIP2033/DIP2034	Putative transposase	DIP2034	2085350	2086593	R	Transposase, mutator family	Pseudogene increased
DIP2149/DIP2152	Putative transposase	DIP2149	2212812	2214416	R	Transposase for insertion sequence	Pseudogene increased
DIP2222	Putative exported protein	DIP2220A/ DIP2220B	2310566/ 2310715	2310790/ 2310927	F	Hypothetical protein	Match with *Corynebacterium* species
DIP2309/DIP2310	Putative DNA-binding protein	DIP2309	2405722	2407315	R	Divergent AAA domain protein	Create new pseudogene with CDS

New CDS with function

Gene ID Cd* RefSeq	Product	Status	Begin	End	Strand
DIP0020	glycosylase	New pseudogene	20341	21185	R
DIP0201	gp1, terminase	Pseudogene unmerged	166488	166832	F
DIP0201A	gp2, terminase	Pseudogene unmerged	166822	168423	F
DIP0493A	Putative molybdopterin converting factor	New CDS	459868	460140	F
DIP1267	Putative short chain dehydrogenase	New CDS	1275251	1275829	F
DIP1974A	Lipoprotein LpqE	New CDS	2022366	2022947	F
DIP2156A	Possible amidohydrolase	New CDS	2217532	2217855	F

New CDSs with unknown functions

Amount	Product
50	Hypothetical protein
9	Hypothetical secreted protein
4	Transposases
1	Hypothetical membrane protein

Notes: **C. diphtheriae*. RefSeq is the National Center for Biotechnology Information Reference Sequence database.

Abbreviations: cAMP, cyclic adenosine monophosphate; CDS, coding sequence; CRISPR, clustered regularly interspaced palindromic repeats; DNA, deoxyribonucleic acid; ID, identification; IS, insertion sequence.

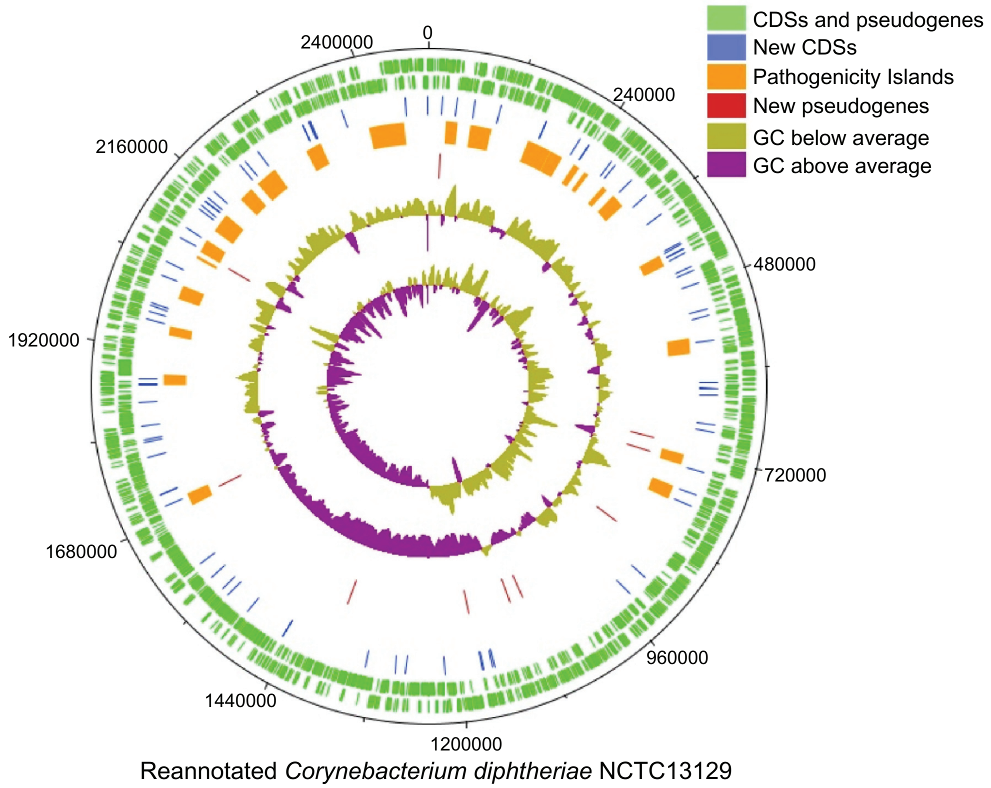


Figure 1 Genomic map showing the new coding sequences (CDSs) and pseudogenes along with all pathogenicity islands.
Notes: Rings from outside to inside: first and second rings (green), CDSs and pseudogenes which have not underpassed through modifications in length; third ring (blue), new CDSs; fourth ring (orange), pathogenicity islands; fifth ring (red), new pseudogenes; sixth ring (purple and brown), guanine-cytosine (GC) plot; seventh ring (purple and brown), GC skew.

region, in which the CDS DIP1975 was previously predicted in the NCBI Reference Sequence.

The product of the new CDS (DIP1974a) had strong similarity with the "lipoprotein – LpqE" protein, with various matches to homologous proteins from other species of *Corynebacterium*. In contrast, CDS DIP1975 did not show any hits with any phylogenetically related organism and for this reason was removed.

In spite of the sparse description of Gram-positive pathogens, it is known that lipoproteins, structural components of the membrane of various bacteria, are important factors linked to the stimulation of an immune response, especially in humans.[26] Hence, the new gene DIP1975 may be intimately related to the virulence of the *C. diphtheriae* and may be the target of studies for the development of new therapies.

Functional reannotation: new annotation reveals genetic elements of *C. diphtheriae* acting against foreign DNA

Various fields may be altered, based on searches for similarity and protein domains conserved in the new annotation. As shown in Figure 3, 43% of the genome annotation remains unaltered,

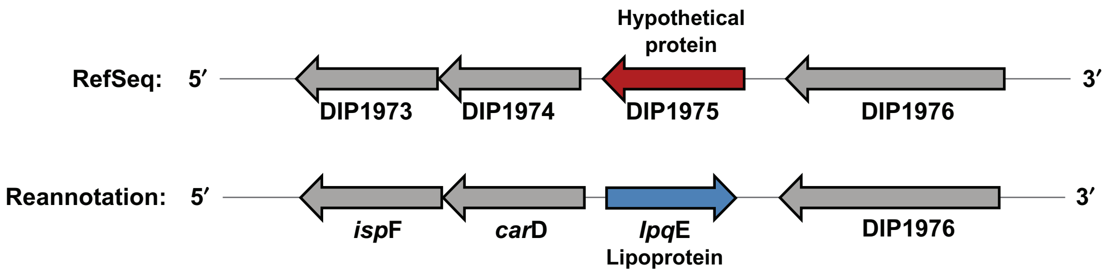


Figure 2 Illustrative schematic of the correction of open reading frames, for the correct orientation of the genome, based on protein similarity.
Notes: Open reading frame DIP1975 is shown in red, in the wrong orientation from the first annotation of the *Corynebacterium diphtheriae* NCTC13129 genome. The corrected open reading frame (DIP_1976) is illustrated in blue with its probable genetic product, which was predicted based on searches for protein similarity (Basic Local Alignment Search Tool [protein]) against the nonredundant protein database (cutoff: 10^{-6}). RefSeq is the National Center for Biotechnology Information Reference Sequence database.

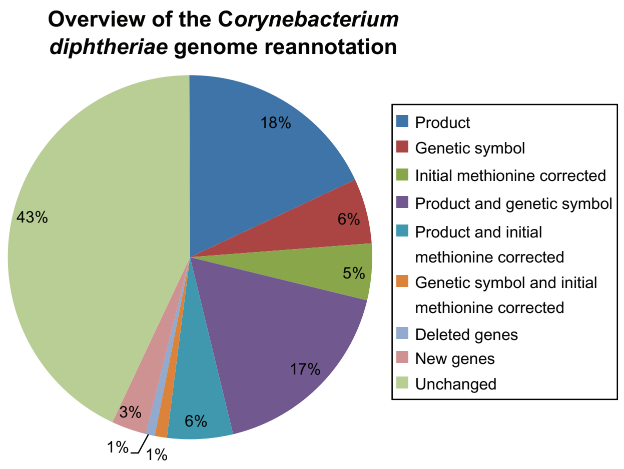


Figure 3 Overview of the changes that occurred in the *Corynebacterium diphtheriae* NCTC13129 genome after the process of reannotation, in the principal categories of change.

while 57% underwent alterations in various aspects. The field most frequently altered was the "product" of the genes (41%): that is, 973 proteins. Among them, 370 proteins ceased to be hypothetical proteins, conserved hypothetical proteins, and/or hypothetical membrane proteins. This alteration provided significant genetic knowledge of the various genes previously annotated as hypothetical proteins, and today, their functions are known, including specific genes encoding virulence factors and pathogenicity. Several acetyltransferases, receptors for iron-binding, fimbrial subunits, transposases, and proteins possibly linked to bacterial virulence were identified.

A common feature of those virulence factors is their location on PAIs, large regions acquired through horizontal gene transfer, which play important roles in the evolution of pathogenic bacteria. *C. diphtheriae* is known to harbor 13 putative PAIs (PiCds 1–13) with genes coding for putative iron transport genes, exported proteins, two component system proteins, insertion sequence transposases, and the diphtheria toxin coding gene (*tox*), which is located in a corynephage-acquired region.[9] Through the use of the software PIPS (http://www.genoma.ufpa.br/lgcm/pips), 11 additional PAIs were identified in the genome sequence of *C. diphtheriae* (Additional File 2) in the original genome annotation (PiCds 14–24). After the reannotation of the genome sequence, the same PiCds 1–24 were identified. However, this new finding deserves special attention.

Two clustered regularly interspaced short palindromic repeats (CRISPR) elements which were initially annotated as "hypothetical proteins" were found to be located in two regions identified as the fourteenth and thirteenth putative PAIs of *C. diphtheriae* (PiCd 14 and PiCd 13), respectively. Table 2 shows these ORFs with their new description. The existence of these regions and of gene families (*cas*) associated with CRISPR regions is known to play an important role against infection by bacteriophages and other mobile genetic elements.[27,28] Such sites showed various repetitions and, in these or some studies, were interpreted as an immune response mechanism against bacterial invasion.[29]

Proteins associated with these regions recognize foreign DNA and use it in a mechanism to silence DNA, similar to RNA interference.[28,30] The DNA is fragmented by *cas*-type proteins in segments of approximately 30 base pairs; the fragments are then inserted into the repetitive regions of CRISPR, which are expressed constitutively.[31,32] These expressed RNAs become guides for other *cas* proteins to process the foreign DNA, as occurs in the RNA interference mechanism.[28,30]

The function of CRISPR in immunity against mobile elements is clearly shown in *Enterococcus faecalis*, where the antibiotic-sensitive strain OG1RF, which possesses two CRISPR arrays, lacks most of the antibiotic resistance genes that are harbored by the hospital-adapted strain V583.[33] Interestingly, Palmer and Gilmore showed that five hybrid

Table 2 Clustered regularly interspaced short palindromic repeats (CRISPR) elements and associated genes described in the new annotation of the *Corynebacterium diphtheriae* genome NCTC13129

Gene ID RefSeq	New ID	Previous annotation RefSeq	Reannotation of *C. diphtheriae*	CRISPR Region
DIP0036	DIP0036	Conserved hypothetical protein	CRISPR-associated protein, Csn1 family	1
DIP0037	DIP0037	Conserved hypothetical protein	CRISPR-associated protein, Cas1 family	1
DIP0038	DIP0038	Conserved hypothetical protein	CRISPR-associated protein, Cas2 family	1
DIP2208	DIP2208	Conserved hypothetical protein	CRISPR-associated protein, Cas5 family	2
DIP2209	DIP2209	Conserved hypothetical protein	CRISPR-associated protein, Cas4 family	2
DIP2210	DIP2210	Conserved hypothetical protein	CRISPR-associated protein, Cse2 family	2
DIP2212	DIP2212	Conserved hypothetical protein	CRISPR-associated protein, Cse3 family	2
DIP2213	DIP2213	Putative helicase	CRISPR-associated helicase Cas3 family	2
DIP2214	DIP2214	Conserved hypothetical protein	CRISPR-associated protein, Cas1 family	2
DIP2215	DIP2215	Conserved hypothetical protein	CRISPR-associated protein, Cas2 family	2

Note: RefSeq is the National Center for Biotechnology Information Reference Sequence database.
Abbreviation: ID, identification.

strains, originating by the acquisition of a resistance island of the strain V583 by OG1RF, are deficient in one CRISPR array possibly due to displacement of this region during DNA incorporation.[33] Moreover, they speculated that modern antibiotic therapy may facilitate the increase in plasticity through the disruption of the balance between the two opposing forces, the acquisition of foreign DNA and degradation of this DNA by self-defense mechanisms.[33]

Following the reannotation of the *C. diphtheriae* genome, the existence of two major operons became clear, denoted as CRISPR 1 region and CRISPR 2 region, which could assume that role in this pathogen. This information has already been presented in studies performed by Mokrousov,[34] but the name of genes and their products remain unchanged in the currently available *C. diphtheriae* genome file.

As shown in Table 2, the CRISPR 1 region is composed of three genes (DIP0036, DIP0037, and DIP0038), *cns1*, *cas1*, and *cas2*, respectively, which participate in the cascade of recognition and silencing of foreign DNA. The CRISPR 2 site of *C. diphtheriae* has seven genes (DIP2208–DIP2210 and DIP2212–DIP2215), containing the *casD*, *casC*, *casB*, *casF*, DIP2211, *casG*, and *casF* genes. Genes superimposed in these two regions were deleted from the annotation.

In spite of a vast genetic repertoire, containing genes that resist the invasion of bacteriophages and mobile genetic elements, the *C. diphtheriae* genome shows another reality. The new annotation showed a large number of transposases and structural proteins originating from bacteriophages. The presence of the CRISPR regions may be an indication that the genome could be even more plastic in their absence.

Pseudogenes, transposases, and other phase-variable elements

The reannotation validated all the pseudogenes of the *C. diphtheriae* genome. In the previous version there were 48 pseudogenes, and 51 pseudogenes were included in the new version. Of the existing pseudogenes, 31 remained the same between the two annotations, and five were no longer pseudogenes and led to eight normal CDSs. These are: DIP0201 (DIP0201 and DIP0201A), DIP0269 (DIP0269), DIP1267 (DIP1267), DIP1523 (DIP1522A and DIP1522B), and DIP2026 (DIP2026). The descriptions of their products are found in Table 3. Furthermore, ten new pseudogenes were detected, which are also shown in Table 3.

It is worth noting that a large part of the features called pseudogenes are probably transposases. In all, there are 56 transposases encoded along the genome. This number, in fact, is a characteristic seen in many species of the *Corynebacterium* genus. The locations of many annotated transposases were in areas flanked by probable PAIs,

Table 3 Coding sequences that ceased to be pseudogenes and new pseudogenes not described in the previous version of the *C. diphtheriae* NCTC13129 genome

New ID	RefSeq ID	Begin	End	Strand	CDS that are no longer pseudogenes	
					Product	**Status**
DIP0201	DIP0201	166488	166832	F	gp1, terminase	No longer a pseudogene
DIP0201A	DIP0201	166822	168423	F	gp2, terminase	No longer a pseudogene
DIP0269	DIP0269	233365	233853	R	Hypothetical protein	No longer a pseudogene
DIP1267	DIP1267	1275251	1275829	F	Putative short chain dehydrogenase	No longer a pseudogene
DIP1522A	DIP1523	1546421	1546561	F	Hypothetical protein	No longer a pseudogene
DIP1522B	DIP1523	1546837	1546971	F	Hypothetical protein	No longer a pseudogene
DIP2026	DIP2026	2079750	2079992	R	Putative transposase for insertion sequence element	No longer a pseudogene
					CDS – new pseudogenes	
DIP0020	DIP0018/DIP0019/DIP0020	20341	21185	R	Glycosidases	New pseudogene
DIP0734	DIP0734/DIP0735	711318	712185	R	Putative sodium/glutamate symporter	New pseudogene
DIP0757	DIP0757/DIP0757A	736001	736487	R	IS element transposase (partial)	New pseudogene
DIP0899	DIP0898/DIP0899	868809	869972	F	HNH endonuclease domain protein	New pseudogene
DIP1095	DIP1095	1077807	1078934	F	Conserved hypothetical integral membrane protein	New pseudogene
DIP1118	DIP1118	1101135	1103803	F	Integral membrane protein, MmpL family	New pseudogene
DIP1177	DIP1177	1175409	1176715	F	Conserved hypothetical protein	New pseudogene
DIP1367	DIP1367	1384058	1384607	R	Transposase of insertion sequence	New pseudogene
DIP1654	DIP1654/DIP1655	1689384	1690688	R	LGFP repeat superfamily protein	New pseudogene
DIP2023	DIP2023/DIP2024	2076614	2077463	F	Filamentation induced by cAMP protein	New pseudogene

Note: RefSeq is the National Center for Biotechnology Information Reference Sequence database.
Abbreviations: cAMP, cyclic adenosine monophosphate; CDS, coding sequence; ID, identification; IS, insertion sequence.

reinforcing the idea of the acquisition of islands by lateral transfer. This is because most of the transposases seen in prokaryotic organisms are of exogenous origin.[35] Nevertheless, the transposases perform an important role in the diversification and evolution of bacterial genomes.[36] In the reannotation of *C. diphtheriae*, the insertion of transposases into genes may be noted, interrupting their reading phases, as in the *hsdM* gene, shown in Figure 4. Today, the high plasticity of the *C. diphtheriae* genome is known;[34] perhaps the large number of transposases present have an important role in diversification and could also confirm the increase of such atypical and more virulent strains. Furthermore, these genes are identified inside PAIs of *C. diphtheriae*.

The interruption of the gene cited above is an interesting finding. The *hsd*(*R*, *S*, and *M*) gene encodes type I restriction enzymes, generally with three subunits, involved in the methylation of adenine residues. The subunit encoded by the *hsdS* gene identifies the DNA region to be methylated, and the *hsdM* gene performs the methyltransferase activity. Finally, the *hsdR* gene translocates the *hsdS*-*M* complex to the target region, even though they are kilobases apart. This mechanism is seen as a preventive action taken by the cell against invasion by foreign DNA, principally by bacteriophages.[37,38]

In *C. diphtheriae*, there is a complete operon with the three genes present: *hsdR*-*S*-*M* (DIP2312, DIP2313, and DIP2314, respectively), and the interrupted gene *hsdM* (DIP2081) in another region of the genome. In many populations of *Mycoplasma pulmonis*, the presence of these enzymes was not detected, and even with intact genes, the bacterium is susceptible to infection by bacteriophages. Additionally, analyzing the operon structurally in *C. diphtheriae*, it appeared functional. However, one of its genes, although extra, was interrupted by the insertion of a transposase (Figure 4).

The presence of CRISPR arrays and restriction enzymes inside PAIs raises the question about what extent genes with functions related to immunity against mobile elements may be incorporated from infecting phages or acquired plasmids to avoid coinfection and/or cotransformation by other incoming DNAs, maintaining the DNA balance.

Discussion

Improvement of annotations of specific genes encoding virulence factors

An important approach currently used in prokaryote genomes is data mining to search for genes that may be linked to virulence and pathogenicity pathways and activities.[39] Many characteristics are taken into consideration in this search, such as immunogenicity of the likely products, proteins with adhesive properties, host-pathogen interaction, bacterial dissemination through the host tissues, and proteins without homology to the host. Therefore, following the reannotation of the *C. diphtheriae* genome, a search was made for these gene targets using the Vaxign software.[40] This program is currently used mainly in the search for new candidate genes in the development of vaccines, but the purpose of the present work was to identify the gene targets connected to virulence, pathogenicity, and immunogenic properties. Furthermore, these results can help us understand the reemergence of diphtheria in the world.

The cell wall and extracellular as well as subcellular locations of the proteins were used. The choice was based on the characteristics of these proteins, as they are the first factors to come in contact with the host to promote the dissemination of the microorganism and are frequently highly immunogenic adhesion molecules.[18] The focus of this study was directed at those proteins whose description showed little detail, or even lacked information in the previous version of the genome. Now, such reannotated proteins provide a better description of their function or their activity. A search through the entire genome revealed 23 good extracellular gene targets, as shown in Table 4. Among the secreted proteins, nine were formerly considered as hypothetical proteins, and they now have a better description. Moreover, this mining of the genome for cell wall proteins resulted in 14 suggested proteins and their probable functions, shown in Table 4.

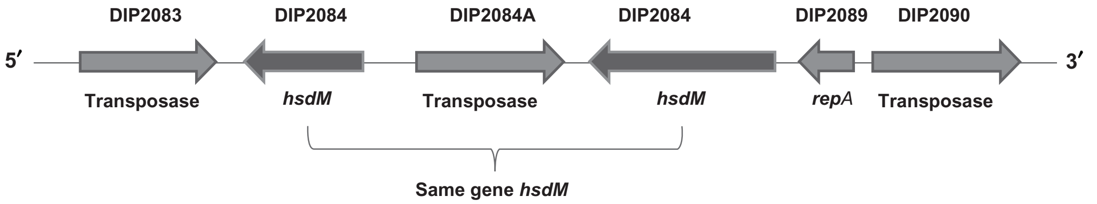


Figure 4 Illustrative schematic of the *hsdM* (DIP_2081) gene, interrupted by the insertion of a transposase.
Notes: Highlighted in dark gray is the *hsdM* gene. The gene was interrupted by the insertion of a transposase (light gray). In addition, the gene is flanked by two more probable transposases (DIP_2080 and DIP_2084). It is possibly a "hotspot" region for the insertion of mobile genetic elements. The interruption of this gene occurs by the addition of the DIP_2082 transposase.

Table 4 New description for *Corynebacterium diphtheriae* genes which have immunological properties and virulence activity (University of Michigan Medical School, Ann Arbor, MI)[27]

Gene ID RefSeq	Product RefSeq	New product	Adhesion probability
DIP0225	Putative secreted polysaccharide deacetylase	Polysaccharide deacetylase	0.125
DIP0298	Putative penicillin-binding secreted protein	Penicillin-binding protein 1B, secreted protein – Pbp1B	0.560
DIP0365	Surface layer protein A	Surface layer protein A	0.577
DIP0543	Putative sialidase precursor	Neuraminidase (sialidase) – NanH	0.200
DIP0554	Putative subtilisin-like cell wall associated serine protease (mycosin)	Subtilisin-like serine protease (mycosin)	0.375
DIP0559	ESAT-6-like protein	ESAT-6-like protein – EsxT	0.521
DIP0640	Hypothetical protein DIP0640	NPL/P60-family secreted protein	0.631
DIP0793	Hypothetical protein DIP0793	Putative twin-arginine translocation pathway signal protein	0.513
DIP0836	Hypothetical protein DIP0836	Putative secreted metalloendopeptidase – MepA	0.825
DIP1097	Putative low molecular weight protein antigen 6	Putative low molecular weight protein antigen 6	0.146
DIP1281	Putative invasion protein	Resuscitation-promoting factor interacting protein – RpfI	0.527
DIP1621	Hypothetical protein DIP1621	NlpC/P60 family protein	0.465
DIP1622	Hypothetical protein DIP1622	NlpC/P60 family protein	0.556
DIP1701	Putative ribonuclease	Guanyl-specific ribonuclease Sa3	0.577
DIP2034	Putative transposase	Transposase, mutator family	0.000
DIP2193	Putative secreted antigen	Trehalose corynomycolyl transferase C – CmtC	0.398
DIP2194	Putative secreted antigen	Trehalose corynomycolyl transferase B – CmtB	0.424
DIP2294	Putative penicillin-binding protein	Penicillin-binding protein C – PbpC	0.741
DIP2339	Putative major secreted protein	Trehalose corynomycolyl transferase A – CmtA	0.346
Cell wall proteins: genetics target of *C. diphtheriae* vaccine			
DIP0139	Hypothetical protein DIP0139	Conserved hypothetical protein	0.669
DIP0235	Putative fimbrial subunit	Putative fimbrial protein	0.739
DIP0237	Putative surface-anchored protein	Surface-anchored protein (fimbrial subunit) – SpaE	0.517
DIP0238	Putative surface-anchored fimbrial subunit	Surface-anchored protein (fimbrial subunit) – SpaF	0.809
DIP0515	Putative transport system secreted protein	ABC-type dipeptide/oligopeptide/nickel transport system – OppA	0.549
DIP0956	Putative peptide transport system secreted protein	ABC transporter solute-binding protein – OppA1	0.502
DIP1740	ABC transporter solute-binding protein	ABC transporter solute-binding protein	0.405
DIP2010	Putative surface-anchored membrane protein	Possible surface-anchored membrane protein	0.565
DIP2013	Putative surface-anchored fimbrial subunit	Putative surface-anchored fimbrial subunit	0.559
DIP2066	Putative surface-anchored fimbrial associated protein	Putative surface-anchored fimbrial associated protein	0.843
DIP2093	Sdr family related adhesin	Putative Sdr-family related adhesin	0.818
DIP2162	ABC transporter solute-binding protein	Periplasmic binding protein-like II	0.221
DIP2226	Surface-anchored fimbrial subunit	Surface-anchored fimbrial subunit – SpaH	0.741
DIP2227	Surface-anchored fimbrial subunit	Surface-anchored fimbrial subunit – SpaG	0.816

Note: RefSeq is the National Center for Biotechnology Information Reference Sequence database.
Abbreviation: ID, identification.

An important characteristic noted was the presence of classical virulence factors such as adhesins (DIP2093), fimbrial subunits (DIP0235, DIP0237, DIP0238, DIP2013, DIP2066, DIP2226, and DIP2227), and adenosine triphosphate-binding cassette transporter proteins connected to the transport of solutes (DIP0515, DIP0956, and DIP1740). These proteins, which are located in the cell wall fraction, were analyzed in silico (Table 4). The knowledge of these elements that promote interaction with the host is vast.[40] As they are some of the first elements to make contact with the host cell, and show no counterpart in the host, they generally are the targets that are used in the development of vaccines. Dealing with the classical elements, which are already well described in the literature and in databases, the reannotation did not result in significant changes in these proteins.

However, the extracellular portion underwent a significant change in the functional annotation. There is a large number of proteins connected to the membrane, such as polysaccharide deacetylase (DIP0225), NlpC/60 protein families (DIP1621 and DIP1622), penicillin-binding proteins (DIP0298 and DIP2294), and unique proteins of actinobacteria, for example, subtilisin-like serine protease (mycosin, DIP0554) with high proteolytic capacity and structural proteins, including trehalose corynomycolyl transferase (DIP2193, DIP2194, and DIP2339).[41] Most of these proteins, connected to the extracellular part of the membrane, show enormous capacity to promote recognition and immune response in the host. It is because of this characteristic that they are good candidates for the development of more effective therapies.

Additionally, one protein on the list shown in Table 4 deserves highlighting. In dealing with a human pathogen such as *C. diphtheriae*, having the ability to colonize the mucosa, the presence of the neuraminidase (sialidase) gene (DIP0543) confers an extra ability to use solutes present only in animal host cells, such as sialic acid. Some pathogens have the capacity to use this sugar as a source of carbon and thereby possess an extra mechanism for surviving within the cell, in a hostile environment.[42]

In addition, the use of this compound can interfere with the defense system of the host by diminishing the viscosity of the mucus and diminishing the activity of inflammatory cells. A rapid and sensitive assay for neuraminidase using peanut lectin hemagglutination was used to study the prevalence of neuraminidase activity among sucrose-fermenting and nonsucrose-fermenting toxigenic *C. diphtheriae* strains. Neuraminidase activity was found in all isolates regardless of biotype, hemagglutinating activity, and site of isolation of bacteria. Besides expressing neuraminidase activity that hydrolyzes sialic acid from glycoconjugates, *C. diphtheriae* was also capable of transferring sialic acid residues from a sialyl-lactose donor. A single molecule probably expresses both neuraminidase and trans-sialidase activity. The trans-sialidase activity was documented by observations of the interactions of bacterial cells with wheat germ agglutinin and peanut lectins. *C. diphtheriae* expressed a trans-sialidase activity located on the cell surface that produced asialoglycoconjugates from a sialyl donor substrate and at the same time generated bacterial sialyl derivatives of beta-galactosidase acceptors.[43] Therefore, the action of this protein can be a strong indication of the ability of *C. diphtheriae* to colonize the mucosa of human airways, escaping from the human immune system and causing disease.

Aside from proposing various genes that may be broadly used as targets in therapy studies, the findings presented here show a bit of the virulence and the pathogenicity of this reemerging and diversifying pathogen, now in a more detailed way.

Description of other operons in *C. diphtheriae* PAIs

Additionally, other genes linked to virulence could be described. Several operons inside the PAIs of *C. diphtheriae* had been assigned a gene name on the genome information, such as *cyd*, *dha*, and *pdx* operons. The *cydABCD* operon codes for an oxygen-scavenging enzyme (cytochrome d) which is reported to be elevated in situations where oxygen is a limiting factor for bacterial growth. Besides, cytochrome d has several different roles in bacteria, including scavenging oxygen that could inactivate oxygen-sensitive nitrogenases, contributing to energy conservation under microaerobiosis and protecting bacteria from oxidative stress.[44,45]

The *dha* operon (PiCd 24) pertains to a family of enzymes that utilize phosphate donors, such as adenosine triphosphate or phosphoproteins, to phosphorylate a toxic compound formed during glycerol metabolism (dihydroxyacetone) into a nontoxic compound (dihydroxyacetone phosphate).[46,47]

Finally, the *pdx* operon (PiCd2) is composed of the genes *pdxS* and *pdxT,* which code for pyridoxal 59-phosphate synthase. Pyridoxal 59-phosphate synthase is regulated by another gene of the *pdx* operon, *pdxR*, and plays an important role in the de novo biosynthesis of vitamin B6, which, in turn, is an essential cofactor for several enzymes catalyzing a variety of biochemical reactions.[48]

Conclusion

C. diphtheriae genome annotation underwent alteration in 57% of its contents, after reannotation. The entire approach was manual, following protocols already established in the literature for the functional annotation of genomes. Updating genomes already available in databases, in addition to supporting research groups performing experiments using the genomic data, also helps in the minimization of annotation errors.

The reannotation resulted in the discovery of new genes in the *C. diphtheriae* genome sequence, correction of ORF strands, and improvement of the functional description of the genome, including classical virulence genes. In addition, it assisted in the search for new gene targets for the development of more effective therapies, information hitherto unpublished in the literature. Nevertheless, the improvement in the description of the proteins linked to the different bacterial defense mechanisms present in the genome, besides providing knowledge of how *C. diphtheriae* may respond to invasion by mobile genetic material, provides indications about its plasticity and the modulation of the genome.

Finally, the protocol used in the present genome can be applied to any genome, whether already on file or not, aimed at the improvement, accuracy of the annotation, and the search for virulence and pathogenicity genes of microorganisms.

Authors' contributions

SCS, VD, AA, ARS, ACP, AACM, CJF, EB, LCG, ME, SSA, VACA, AZN, ARC, LTC, RTJR were involved in all of: predictions of genes, transfer RNA, ribosomal RNA, and conserved domains of proteins; in similarity searches of *C. diphtheriae* genomes against several databases; and, in the functional reannotation. SCS performed the PAIs analysis. ARS located the subcellular proteins in the genome. VD was responsible for searching new genetic targets for the development of vaccines. VA and AM coordinated and participated in the conception, design, and supervision of the whole project. VA, AM, ALMG, RHJ, SCS, VD, ET, AT, AS, and MPS were involved in writing the manuscript.

Acknowledgments

Funding was provided by CNPq (Conselho Nacional de Desenvolvimento Científico) and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Médigue C, Moszer I. Annotation, comparison and databases for hundreds of bacterial genomes. *Res Microbiol*. 2007;158(10):724–736.
2. Salzberg SL. Genome re-annotation: a wiki solution? *Genome Biol*. 2007;8(1):102.
3. Petty NK. Genome annotation: man versus machine. *Nature*. 2010; 8(11):762.
4. Boneca IG, de Reuse H, Epinat J, Pupin M, Labigne A, Moszer I. A revised annotation and comparative analysis of Helicobacter pylori genomes. *Nucleic Acids Res*. 2003;31(6):1704–1714.
5. Luo C, Hu GQ, Zhu H. Genome reannotation of Escherichia coli CFT073 with news insights into virulence. *BMC Genomics*. 2009;10:522.
6. Gundogdu O, Bentley SD, Holden MT, Parkhill J, Dorrell N, Wren BW. Re-annotation and re-analysis of the Campylobacter jejuni NCTC11168 genome sequence. *BMC Genomics*. 2007;8:162.
7. Camus JC, Pryor MJ, Médigue C, Cole ST. Re-annotation of genome sequence of Mycobacterium tuberculosis H37Rv. *Microbiology*. 2002; 148(Pt 10):2967–2973.
8. Dandekar T, Huynen M, Regula JT, et al. Re-annotating the Mycoplasma pneumonia genome sequence: adding value, function and reading frames. *Nucleic Acids Res*. 2000;28(17):3278–3288.
9. Cerdeño-Tárraga AM, Efstratiou A, Dover LG, et al. The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. *Nucleic Acids Res*. 2003;31(22):6516–6523.
10. Rappuoli R, Podda A, Giovannoni F, Nencioni F, Peragallo L, Francolini P. Absence of protective immunity against diphtheria in a large proportion of young adults. *Vaccine*. 1993;11(5):576–577.
11. Pimenta FP, Hirata R Jr, Rosa AC, Milagres LG, Mattos-Guaraldi AL. A multiplex PCR assay for simultaneous detection of Corynebacterium diphtheriae and differentiation between non-toxigenic and toxigenic isolates. *J Med Microbiol*. 2008;57(Pt 11):1438–1439.
12. Gomes DL, Martins CA, Faria LM, et al. Corynebacterium diphtheriae as an emerging pathogen in nephrostomy catheter-related infection: evaluation of traits associated with bacterial virulence. *J Med Microbiol*. 2009;58(Pt 11):1419–1427.
13. Lagesen K, Hallin P, Rødland EA, Staerfeldt H, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res*. 2007;35(9):3100–3108.
14. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*. 1997;25(5):955–964.
15. Benson G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*. 1999;27(2):573–580.
16. Zdobnov EM, Apweiler R. InterProScan – an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*. 2001;17(9):847–848.
17. Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation. *Bioinformatics*. 2000;16(10):944–945.
18. He Y, Xiang Z, Mobley HL. Vaxign: the first web-based vaccine design program for reverse vaccinology and an application for vaccine development. *J Biomed Biotechnol*. 2010;2010:297505.
19. Waack S, Keller O, Asper R, et al. Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinformatics*. 2006;7:142.
20. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 2011;39:W29–W37.
21. Carver TJ, Rutherford KM, Berriman M, Rajandream M, Barrell BG, Parkhill J. ACT: the Artemis Comparison Tool. *Bioinformatics*. 2005;21(16):3422–3423.
22. Zhou CE, Smith J, Lam M, Zemla A, Dyer MD, Slezak T. MvirDB – a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications. *Nucleic Acids Res*. 2007;35:D391–D394.
23. Karp PD, Paley S, Romero P. The Pathway Tools software. *Bioinformatics*. 2002;18 Suppl 1:S225–S232.
24. Caspi R, Foerster H, Fulcher CA, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res*. 2008;36: D623–D631.
25. Caspi R, Karp PD. Using the MetaCyc pathway database and the BioCyc database collection. *Curr Protoc Bioinformatics*. 2007;Chapter 1 (Unit 1):17.
26. Bubeck Wardenburg J, Williams WA, Missiakas D. Host defenses against Staphylococcus aureus infection require recognition of bacterial lipoproteins. *Proc Natl Acad Sci U S A*. 2006;103(37): 13831–13836.
27. Grissa I, Vergnaud G, Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*. 2007;8:172.
28. Marraffini LA, Sontheimer EJ. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nat Rev Genet*. 2010;11(3): 181–190.
29. Zegans ME, Wagner JC, Cady KC, Murphy DM, Hammond JH, O'Toole GA. Interaction between bacteriophage DMS3 and host CRISPR region inhibits group behaviors of Pseudomonas aeruginosa. *J Bacteriol*. 2009;191(1):210–219.
30. Jansen R, Embden JD, Gaastra W, Schouls LM. Identification of genes that are associated with DNA repeats in prokaryotes. *Mol Microbiol*. 2002;43(6):1565–1575.
31. Mojica FJ, Díez-Villaseñor C, García-Martínez J, Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. *J Mol Evol*. 2005;60(2):174–182.
32. Pourcel C, Salvignol G, Vergnaud G. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. *Microbiology*. 2005;151(Pt 3):653–663.
33. Palmer KL, Gilmore MS. Multidrug-resistant enterococci lack CRISPR-cas. *MBio*. 2010;1(4):e00227–e00310.

34. Mokrousov I. Corynebacterium diphtheriae: genome diversity, population structure and genotyping perspectives. *Infect Genet Evol*. 2009;91(1):–15.
35. Mes TH, Doeleman M. Positive selection on transposase genes of insertion sequences in the Crocosphaera watsonii genome. *J Bacteriol*. 2006;188(20):7176–7185.
36. Ooka T, Ogura Y, Asadulghani MD, et al. Inference of the impact of insertion sequence (IS) elements on bacterial genome diversification through analysis of small-size structural polymorphisms in Escherichia coli O157 genomes. *Genome Res*. 2009;19(10):1809–1816.
37. Dybvig K, Sitaraman R, French CT. A family of phase-variable restriction enzymes with differing specificities generated by high-frequency gene rearrangements. *Proc Natl Acad Sci U S A*. 1998;95(23): 13923–13928.
38. Obarska-Kosinska A, Taylor JE, Callow P, Orlowski J, Bujnicki JM, Kneale GG. HsdR subunit of the type I restriction-modification enzyme EcoR124I: biophysical characterisation and structural modelling. *J Mol Biol*. 2008;376(2):438–452.
39. Rappuoli R. Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine*. 2001;19(17–19):2688–2691.
40. Webb SA, Kahler C. Bench-to-bedside review: bacterial virulence and subversion of host defences. *Crit Care*. 2008;12(6):234–242.
41. Ramulu HG, Adindla S, Guruprasad L. Analysis and modeling of mycolyl-transferases in the CMN group. *Bioinformation*. 2006;1(5): 161–169.
42. Lichtensteiger CA, Vimr ER. Neuraminidase (sialidase) activity of Haemophilus parasuis. *FEMS Microbiol Lett*. 1997;152(2):269–274.
43. Mattos-Guaraldi AL, Formiga LC, Andrade AF. Trans-sialidase activity for sialic acid incorporation on Corynebacterium diphtheriae. *FEMS Microbiol Lett*. 1998;168(2):167–172.
44. Iuchi S, Chepuri V, Fu HA, Gennis RB, Lin EC. Requirement for terminal cytochromes in generation of the aerobic signal for the arc regulatory system in Escherichia coli: study utilizing deletions and lac fusions of cyo and cyd. *J Bacteriol*. 1990;172(10):6020–6025.
45. Winstedt L, Yoshida K, Fujita Y, von Wachenfeldt C. Cytochrome bd biosynthesis in Bacillus subtilis: characterization of the cydABCD operon. *J Bacteriol*. 1998;180(24):6571–6580.
46. Bächler C, Schneider P, Bähler P, Lustig A, Erni B. Escherichia coli dihydroxyacetone kinase controls gene expression by binding to transcription factor DhaR. *EMBO J*. 2005;24(2):283–293.
47. Bizzini A, Zhao C, Budin-Verneuil A, et al. Glycerol is metabolized in a complex and strain-dependent manner in Enterococcus faecalis. *J Bacteriol*. 2010;192(3):779–785.
48. Jochmann N, Götker S, Tauch A. Positive transcriptional control of the pyridoxal phosphate biosynthesis genes pdxST by the MocR-type regulator PdxR of Corynebacterium glutamicum ATCC 13032. *Microbiology*. 2011;157(Pt 1):77–88.

Additional File 1 Similarity analyses between new coding sequences and pseudogenes against the nonredundant database of proteins from National Center for Biotechnology Information

Abbreviations: ABC, adenosine triphosphate-binding cassette; ACP, acyl carrier protein; ATP, adenosine triphosphate; CoA, Coenzyme A; CRISPR, clustered regularly interspaced palindromic repeats; DNA, deoxyribonucleic acid; GMP, guanosine monophosphate; GTP, guanosine triphosphate; IS, insertion sequence; MFS, major facilitator superfamily; NAD, nicotinamide adenine dinucleotide, NADH, reduced form of NAD; NAD(P)H, NAD phosphate; PAI, pathogenicity island; RNA, ribonucleic acid; rRNA, ribosomal RNA; SNARE, soluble N-ethylmaleimide-sensitive factor attachment protein receptor; tRNA, transfer RNA.

Additional File 2 Coding sequences of pathogenicity islands predicted by the software PIPS (http://www.genoma.ufpa.br/lgcm/pips) in the reannotated Corynebacterium diphtheriae NCTC13129 genome

Abbreviation: ABC, adenosine triphosphate-binding cassette.

Open Access Bioinformatics

Dovepress

Publish your work in this journal

Open Access Bioinformatics is an international, peer-reviewed, open access journal publishing original research, reports, reviews and commentaries on all areas of bioinformatics. The manuscript management system is completely online and includes a very quick and fair peer-review system. Visit http://www.dovepress.com/testimonials.php to read real quotes from published authors.

Submit your manuscript here: http://www.dovepress.com/open-access-bioinformatics-journal

3.1.6 Projeto pangenoma de *C. pseudotuberculosis*

Até o ano de 2007 existiam quatro espécies pertencentes ao gênero *Corynebacterium* que possuíam seus genomas completamente sequenciados e depositados no NCBI. Em 2009, essa quantidade aumentou para 15 genomas. Esse aumento em um período relativamente curto demonstra o interesse em caracterizar diferentes espécies desse gênero.

Em 2006, a RGMG apoiou o sequenciamento do genoma da bactéria *C. pseudotuberculosis*, linhagem 1002, para caracterizar geneticamente o agente etiológico da LC. O objetivo era identificar proteínas relacionadas à patologia e que pudessem contribuir para o desenvolvimento de uma eventual vacina, de terapias e de kits de diagnóstico. A linhagem C231, sequenciada por pesquisadores australianos, foi montada em conjunto com a linhagem 1002 como consequência de uma parceria com a RGMG e RPGP. Ambos os genomas foram finalizados em maio de 2009, ocasião em que ocorreu o depósito no banco de dados do NCBI sob os números de acesso CP001809 (linhagem 1002) e CP001829 (linhagem C231). Em 2011, esses genomas, bem como uma análise comparativa com os genomas disponíveis no NCBI do gênero *Corynebacterium*, foram publicados em uma revista internacional indexada (Seção 3.1.1).

Em maio de 2009, foi concluído, pelo nosso grupo, o primeiro sequenciamento de próxima geração na América Latina utilizando o sistema SOLiD® (*Supported Oligonucleotide Ligation and Detection*) da empresa *Applied Biosystems*, uma das plataformas de sequenciamento de nova geração. Foram decodificados naquele momento, os genomas de duas novas linhagens de *C. pseudotuberculosis*, isoladas de cavalo (linhagem 258) e camelo (linhagem 162) e em parceria com o LGCM estes organismos foram montados em prazo recorde de 21 dias.

A Tabela 1 mostra as linhagens de *C. pseudotuberculosis* que foram montadas e anotadas manualmente pelo nosso grupo de pesquisa. Apenas a linhagem 1002 teve dados oriundos do sequenciamento didesoxi enquanto as demais utilizaram tecnologias de sequenciamento de próxima geração (SOLiD, 454, Illumina e Ion Torrent).

Linhagem	Tecnologia	Hospedeiro	Biovar	País de isolamento	Número de acesso (*GenBank*)	Parceiros do LGCM
1002	Sanger, 454	Cabra	*ovis*	Brasil	CP001809	RGMG
C231	454	Ovelha	*ovis*	Austrália	CP001829	RGMG/LPDNA-RPGP
FRC41	454	Humano	*ovis*	França	CP002097	LPDNA-RPGP/Bielefeld
I19	SOLiD v2	Boi	*ovis*	Israel	CP002251	LPDNA-RPGP
PAT10	SOLiD v2	Ovelha	*ovis*	Argentina	CP002924	LPDNA-RPGP
CIP 52.97	SOLiD v2	Cavalo	*equi*	Kenia	CP003061	LPDNA-RPGP
42/02-A	Illumina	Ovelha	*ovis*	Austrália	CP003062	LPDNA-RPGP
316	Ion Torrent	Cavalo	*equi*	EUA	CP003077	LPDNA-RPGP
1/06-A	Illumina	Cavalo	*equi*	EUA	CP003082	LPDNA-RPGP
3/99-5	Illumina	Ovelha	*ovis*	Escócia	CP003152	LPDNA-RPGP
P54B96	Ion Torrent, SOLiD v3	Antílope	*ovis*	África do Sul	CP003385	LPDNA-RPGP
267	SOLiD v3	Lhama	*ovis*	EUA	CP003407	LPDNA-RPGP
31	Ion Torrent, SOLiD v3	Búfalo	*equi*	Egito	CP003421	LPDNA-RPGP
258	SOLiD v2	Cavalo	*equi*	Bélgica	CP003540	LPDNA-RPGP
162	SOLiD v2	Camelo	*equi*	Reino Unido	CP003652	LPDNA-RPGP

Tabela 1: Linhagens de *C. pseudotuberculosis* montadas e anotadas pelo nosso grupo de pesquisa e alvos desse trabalho de doutorado.

Esta tese está focada na análise das linhagens 1002, C231, I19, PAT e FRC41 por razões temporais, sendo duas delas (1002 e C231) fonte de um trabalho extensivo sobre o proteoma exportado *in vitro* (Pacheco e cols., 2011; Silva e cols., 2012).

A infecção por *C. pseudotuberculosis* em humanos é considerada uma doença ocupacional, uma zoonose. Uma menina de 12 anos foi infectada na França quando visitava a fazenda de criação de ovinos de seus avôs. Essa bactéria ficou seis meses em contato com o organismo desta criança (Join-Lambert e cols., 2006). Como existem muitos trabalhos de interação de patógenos com hospedeiros humanos, pode-se utilizar as informações obtidas no sequenciamento do genoma da linhagem FRC41 para minerar candidatos a vacinas e drogas (Trost e cols., 2010).

A Figura 7 mostra os genomas cujo processos internos de depósito do NCBI estão finalizados.

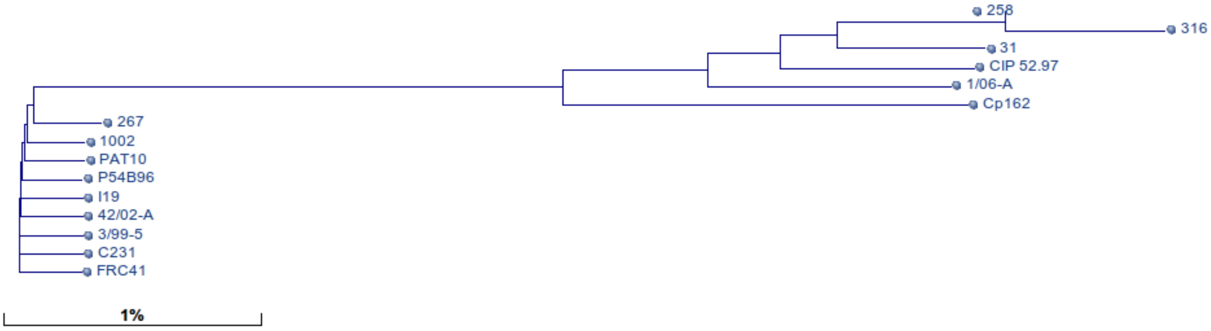


Figura 7: Genomas da *C. pseudotuberculosis* disponíveis através do sítio do NCBI.

Os genomas FRC41, 3/99-5 e 316 são sequências de referência (*RefSeq*) para os genomas da *C. pseudotuberculosis*. A FRC41 é *RefSeq* por ter sido o primeiro genoma dessa bactéria depositada no NCBI; a 316 é *RefSeq* também por ter um genoma com maior diferença em relação à primeira *RefSe*q, mas também deve se levar em conta que pertence ao biovar equi, diferentemente da outra *RefSeq* que é do biovar ovis. Na Figura 7, os genomas da 258, 316, 31, 162, 1/06-A e CIP 52.97 estão filogeneticamente mais distantes dos demais. A explicação pode ser o fato de serem de linhagens do biovar equi, infectando cavalos e búfalos, em comparação com as demais isoladas do biovar ovis. Os genomas do biovar equi poderão ser úteis para estabelecer as bases moleculares de doenças causadas pela *C. pseudotuberculosis* em diferentes hospedeiros.

3.1.7 CpDB: Um banco de dados relacional e ferramentas para armazenamento, recuperação e anotação automática de genomas bacterianos.

Criou-se um esquema relacional de banco de dados para armazenar, compartilhar e facilitar a consulta dos genomas de *C. pseudotuberculosis* denominado CpDB. Porém, apenas um banco de dados relacional não é suficiente para esta demanda, visto que genomas filogeneticamente próximos ao de *C. pseudotuberculosis* também precisam ser comparados. Do mesmo modo, diferentes versões de genomas criados, por exemplo, como consequência de uma anotação melhorada também precisam ser comparados. Para esse propósito, além do banco de dados relacional, também criou-se um programa em linguagem C que interpreta dados em formato EMBL e *GenBank* convertendo-os para o formato de entrada de nosso esquema relacional de banco de dados. Esse programa pertence a uma categoria denominada como *parser* na disciplina de Compiladores, geralmente ofertada em cursos de Ciências da Computação. Um *parser* é um programa conversor entre formatos específicos; nesse caso o *parser* foi denominado '*parseEMBLtoCpDB*'. Esse binômio (CpDB e seu *parser* para entrada de dados) continua sendo uma ferramenta essencial para todos os genomas que o nosso grupo de pesquisa deposita no *GenBank*. Até a data de defesa dessa tese, o CpDB e seu *parser* foram utilizados no auxilio da montagem, anotação e depósito em bancos de dados públicos de 15 genomas de *C. pseudotuberculosis*, um genoma de *Campylobacter fetus* subespécie *venerealis*, dois genomas de *Streptococcus*, um genoma de *Lactococcus lactis* e um genoma de uma *Archae*.

A seção "*Automated functional annotation*", faz parte de um capítulo de livro "*Whole Genome Annotation: in silico Analysis*" (Seção 6.2.6), publicado pelo nosso grupo de pesquisa. Esse capítulo de livro em sua maioria é constituído de material teórico, porém possui uma seção prática que é apresentada em destaque como um dos resultados dessa tese. A referida seção fala sobre a anotação automática de genomas e sobre a transferência de anotação de um genoma manualmente anotado para um genoma o qual se possui apenas uma predição gênica, ambas possibilidades oferecidas pelo CpDB e seu parser de formatação de dados para alimentar o banco de dados. Essa transferência de anotação está sendo extensivamente utilizada no pangenoma em construção da *C. pseudotuberculosis*. Ao final desse texto é apresentado o CpDB como opção para transferência de anotação automática de genomas e é oferecido um endereço (no formato de repositório *subversion*) para que um interessado baixe um tutorial com aproximadamente 30 passos para realizar uma transferência de anotação entre genomas. Esse tutorial possui versões em inglês e português, bem como todos os dados e programas, disponíveis após uma operação de *checkout* do repositório *subversion*. Na sessão de métodos da tese o CpDB foi explicado com detalhes e foram documentados o código fonte do CpDB e do *parser* para entrada de dados (Seção 6.1.2).

3.1.7.1 Discussão

A anotação funcional pode impor desafios de gerenciamento e recuperação de dados ao ser feita por várias pessoas, trabalhando em locais distintos e utilizando sistemas operacionais diversos. Ao mesmo tempo que a anotação funcional do genoma de uma linhagem é iniciada, a montagem de uma outra linhagem pode ser iniciada. Isso impõem o desafio da incorporação da anotação funcional entre linhagens. Nesse contexto de confrontação com problemas clássicos de administração de dados, pensou-se na utilização em uma solução clássica, um banco de dados relacional. Um modelo relacional de banco de dados denominado CpDB (Figura 10) foi criado no SGBD PostgreSQL (POSTGRES, 1993) contendo as principais entidades do formato EMBL para as quais era necessário prover dados, visto que o objeto final desses dados era um arquivo EMBL que seria depositado no banco de dados de genomas do NCBI. Foi construído um *parser* em linguagem de programação C padrão utilizando uma biblioteca de construção de compiladores compatível com o padrão *Lex/YACC*. Esse compilador extrai dados relevantes de arquivos EMBL de modo a prover dados para popular um banco de dados com o esquema relacional CpDB. Uma vez que dados da anotação manual estão presentes no banco de dados é possível, por exemplo, garantir que o identificador único de um ORF não está se repetindo devido a, por exemplo, um erro de edição durante uma anotação manual. Dessa forma, evita-se criar um programa para esse fim. Com todo o formato EMBL corrigido, convertido e integrado no CpDB, é possível exportar os dados em formato EMBL e gerar uma nova versão de uma anotação. Uma consulta em SQL permite extrair dados relevantes ao formato EMBL que preenchem qualificadores de texto do formato EMBL, criando um novo arquivo EMBL. Para reaproveitar a curadoria manual e o posterior tratamento feito aos dados de uma linhagem curada para uma nova linhagem, um arquivo no formato “m8” do BLAST é gerado para criar um elo entre um banco de dados curado manualmente com um banco de dados que está sendo trabalhado. Por meio desse elo mais de 90% dos dados curados podem ser reaproveitados e uma nova exportação de formato EMBL pode ser gerada para posterior verificação manual.

Dentre as principais vantagens do uso de um banco de dados relacionais para armazenar os dados de um genoma pode ser citada a centralização de dados em um servidor com controle de acessos, eliminando o uso de arquivos diversos armazenados em sistemas de arquivos de sistemas operacionais também diversos. Outro ponto que merece destaque é o fato de que muitos procedimentos computacionais que antes precisariam ter um código de programa escrito em uma linguagem de programação, como por exemplo, PERL não são necessários, pois um SGBD possui programas embutidos que garantem a integridade de dados. O CpDB foi utilizado com sucesso para mapear todas as entidades pertinentes de arquivos EMBL, gerados e manipulados pelo programa ARTEMIS (Rutherford e cols., 2000). Assim foi possível importar os dados de um formato EMBL para dentro do banco de dados e depois fazer o caminho inverso, complementando a anotação com dados oriundos de outros bancos de dados, como por exemplo, o banco do *Gene Ontology*.

Dados armazenados no esquema relacional CpDB podem ser exibidos para um pesquisador em dois formatos gráficos. Um dos formatos é o EMBL que o programa ARTEMIS interpreta e exibe em ambiente *desktop;* o outro formato é o GBROWSE utilizado no ambiente *web*. O banco de dados relacionais CpDB permite exportar o formato esperado pelo GBROWSE com uma consulta em formato SQL que possui apenas uma linha de tamanho. Um projeto genoma que utilize um SGBD relacional terá uma plataforma de depósito, análise e recuperação de dado confiável, exaustivamente testada e melhorada bem como uma modelagem intuitiva e adequada a modelos de relacionamentos entre entidades biológicas.

Outro exemplo de como um SGBD economiza a escrita de programas que visam garantir a integridade de dados é fornecido pela criação de relacionamentos entre as entidades do banco. No esquema CpDB foram criados relacionamentos de dependência entre a entidade GENE e entidades relacionadas como SIGNAL que armazena dados relativos à predição de peptídeo sinal principalmente da via de secreção Sec. Esse relacionamento é configurado de modo que um peptídeo sinal somente possa ser inserido no CpDB caso estivesse relacionado com um identificador que pertença a um gene previamente cadastrado. Assim nenhum peptídeo sinal é cadastrado no banco de dados de modo a que ficasse desconectado de um gene. Caso a identificação única de um gene seja alterada, então automaticamente o SGBD modifica esse identificador junto ao seu peptídeo sinal; caso essa modificação seja uma remoção do gene, então o peptídeo sinal correspondente será removido também, sem que o administrador do banco de dados necessite intervir para que essas tarefas de manutenção sejam executadas.

As situações exemplificadas aqui são o cotidiano de um processo de montagem e anotação de genomas e mostram a importância fundamental de um banco de dados para garantir a integridade dos dados de uma anotação. Mostra também que se dispondo de configurações simples entre entidades relacionadas é possível garantir a integridade e consistência de dados sem a necessidade de escrever código para esse fim. Também não são necessárias pessoas dedicadas para executar e manter códigos de programação e manualmente garantir a integridade dos dados. Um SGBD para armazenar dados biológicos não é uma novidade e sim uma necessidade, uma regra que deve ser seguida, pensada e modelada antes que qualquer trabalho de maior proporção tenha início. Empresas notórias e essências sobrevivem atualmente, num contexto de geração de *gigabytes* de dados diários, por conta da confiança que depositam em seus SGBD's e na medida em que são correspondidos. Para profissionalizar o tratamento de dados biológicos, e seguir para etapas mais complexas, é necessario que os bancos de dados sejam imperativos no tratamento desses dados.

3.2 Genômica Funcional

Uma vez definida a estrutura dos genomas de *C. pseudotuberculosis*, criou-se condição para predições *in silico* a respeito da função do genoma. Essas predições podem servir como referência para confirmar dados de experimentos *in vitro*, bem como fornecer uma expectativa dos possíveis resultados esperados em experimentos.

Nesse contexto, são apresentados resultados *in vitro* corroborados pela genômica estrutural. Dois experimentos do exoproteoma de duas linhagens de *C. pseudotuberculosis,* cujos resultados foram complementares, são apresentados em sequência. Em seguida, utilizando-se dados do exoproteoma foi realizado um trabalho de imunoproteômica. Esses três trabalhos foram auxiliados por predições *in silico* do exoproteoma de cinco linhagens de *C. pseudotuberculosis.* Esse último trabalho foi apresentado na forma de um artigo científico submetido à revista *BMC Genomics*, artigo aceito para publicação em abril de 2012. Como forma de tentar selecionar os melhores candidatos do pan exoproteoma predito *in silico*, é apresentado um artigo científico que se propõe a uma releitura da imunoinformática como forma de selecionar os melhores candidatos para vacinas, diagnóstico e drogas. Esse artigo foi submetido após modificações sugeridas por *referees* da revista *Bioinformatics*.

Nesta seção o doutorando é o responsável direto pelos três artigos científicos das seções 3.2.1, 3.2.5 e 3.2.6. Além disso, por meio das análises *in silico* destes artigos, auxiliou a análise e interpretação dos resultados *in vitro* dos outros três artigos científicos, propiciando-lhe a co-autoria nesses trabalhos (seções 3.2.2, 3.2.3 e 3.2.4).

3.2.1 Análise *in silico* do panexoproteoma de cinco linhagens de *C. pseudotuberculosis*

O advento de novas tecnologias de sequenciamento de genomas permitiu a obtenção de genomas bacterianos completos em menor tempo; utilizando menos recursos financeiros e de mão de obra em comparação com a década passada. O primeiro passo da Vacinologia Reversa (VR) é uma análise computacional de um genoma para identificar os candidatos mais prováveis para o desenvolvimento de vacinas contra certos patógenos que geralmente são proteínas exportadas. Estes avanços permitiram estudos pangenômicos destinados a definir os genomas central, dispensável e exclusivo em uma espécie. Na Vacinologia Reversa Pangenômica (VRPG) , ao invés de estudar os genes previstos como exportados em um único genoma, é possível estudar genes presentes no genoma que são compartilhados entre diferentes espécies.

Por razões temporais, os genomas das linhagens 1002, C231, I19, FRC41 e PAT10 da *C. pseudotuberculosis* foram analisados *in silico* para geração de listas de candidatos ao panexoproteoma. Foi gerado um conjunto de 306, 59 e 11 genes homólogos representando os genomas central, dispensável e exclusivo preditos como exportados, respectivamente. Essa análise proporcionou uma lista com 150 genes preditos como secretados (SEC) e 227 genes preditos como potencialmente expostos na superfície (PSE). Neste pangenoma, 55 genes exportados das linhagens 1002 e C231 foram previamente identificados fazendo parte do exoproteoma central. Nossos resultados sugerem que essa quantidade de genes pode ser maior, com o acréscimo de uma fração de pelo menos 35 proteínas dentre as 47 previstas como parte do exoproteoma variante.

As cinco linhagens de *C. pseudotuberculosis* utilizadas neste trabalho foram redepositadas, atualizando os dados no *GenBank*. Estes genomas foram manualmente curados para correção da metionina inicial das proteínas preditas como exportadas com um total de 1885 genes homogeneizados.

A análise *in silico* de proteínas do exoproteoma gerou uma lista de candidatos vacinais presentes nos cinco genomas completos de *C. pseudotuberculosis* e assim estabeleceu-se uma ordem de prioridade para a escolha de genes que devem ser testados em futuras pesquisas de VRPG. Este artigo científico encontra-se submetido na revista *BMC Genomics*. Os arquivos adicionais deste artigo encontram-se na seção de anexos da tese.

The *Corynebacterium pseudotuberculosis in silico* predicted pan-exoproteome

Anderson R. Santos[1], Adriana Carneiro[2], Alfonso Gala-García[1], Anne Pinto[1], Debmalya Barh[3], Eudes Barbosa[1], Flávia Figueira[1], Fernanda Dorella[1], Flávia Souza[1], Louise Cerdeira[2], Luis Guimarães[1], Meritxell Zurita-Turk[1], Rommel Ramos[2], Sintia Almeida[1], Siomar Soares[1], Ulisses Pereira[1], Vinícius C. Abreu[1], Artur Silva[2], Anderson Miyoshi[1], Vasco Azevedo[1§]

[1]Molecular and Celular Genetics Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[2]DNA Polimorfism Laboratory, Universidade Federal do Pará, Campus do Guamá - Belém, PA, Brazil

[3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India

[§]Corresponding author

Email addresses:

ARS: asantos.icb.ufmg.br@gmail.com

AC: carneiroar@gmail.com

AG: alfonsogala25@gmail.com

AP: acybelle@gmail.com

DB: dr.barh@gmail.com

EB: eudesgvb@gmail.com

FF: flavinhafigueira@gmail.com

FD: fernandadorella@gmail.com

FS: flasouz@yahoo.com.br

LC: lcerdeira@gmail.com

LG: luisguimaraes.bio@gmail.com

MT: meritxellzt@gmail.com

RR: rommelramos@ufpa.br

SA: sintiaalmeida@gmail.com

SS: siomars@gmail.com

UP: upaduapereira@gmail.com

VCA: vini.abreu@gmail.com

AS: asilva@ufpa.br

AM: miyoshi@icb.ufmg.br

VA: vasco@icb.ufmg.br

Abstract

Background

Pan-genomic studies aim, for instance, at defining the core, dispensable and unique genes within a species. A pan-genomics study for vaccine design tries to assess the best candidates for a vaccine against a specific pathogen. In this context, rather than studying genes predicted to be exported in a single genome, with pan-genomics it is possible to study genes present in different strains within the same species, such as virulence factors. The target organism of this pan-genomic work here presented is *Corynebacterium pseudotuberculosis*, the etiologic agent of caseous lymphadenitis (CLA) in goat and sheep, which causes significant economic losses in those herds around the world. Currently, only a few antigens against CLA are known as being the basis of commercial and still ineffective vaccines. In this regard, the here presented work analyses, *in silico,* five *C. pseudotuberculosis* genomes and gathers data to predict common exported proteins in all five genomes. These candidates were also compared to two recent *C. pseudotuberculosis in vitro* exoproteome results.

Results

The complete genome of five *C. pseudotuberculosis* strains (1002, C231, I19, FRC41 and PAT10) were submitted to pan-genomics analysis, yielding 306, 59 and 12 gene sets, respectively, representing the core, dispensable and unique *in silico* predicted exported pan-genomes. These sets bear 150 genes classified as secreted (SEC) and 227 as potentially surface exposed (PSE). Our findings suggest that the main *C. pseudotuberculosis in vitro* exoproteome could be greater, appended by a fraction of the 35 proteins formerly predicted as making part of the variant *in vitro* exoproteome. These genomes were manually curated for correct methionine initiation and redeposited with a total of 1885 homogenized genes.

Conclusions

The *in silico* prediction of exported proteins has allowed to define a list of putative vaccine candidate genes present in all five complete *C. pseudotuberculosis* genomes. Moreover, it has also been possible to define the *in silico* predicted dispensable and unique *C. pseudotuberculosis* exported proteins. These results provide *in silico* evidence to further guide experiments in the areas of vaccines, diagnosis and drugs. The work here presented is the first whole *C. pseudotuberculosis in silico* predicted pan-exoproteome completed till today.

Background

Reverse Vaccinology (RV) [1] analyses the genome sequence of a pathogen, which is an expected coded sequence for all the possible expressed genes in the pathogen's life cycle. All Open Reading Frames (ORF's) derived from the genome sequence can be evaluated using a computer program to determine their ability as vaccine candidates, giving special attention to exported proteins, as these are essential in host-pathogen interactions. Examples of such interactions include: (i) adherence to host cells, (ii) invasion of the cell to which there is compliance, (iii) damage to host tissues, (iv) environmental stresses resistance from the defense machinery of the cell being infected, and (v) mechanisms for subversion of host immune response [2-5].

Regarding exported proteins, these can distinguish between those that are exported to the cell wall, and after cleaved, release the mature portion into the extracellular milieu, which are referred to as secreted proteins (SEC), and those proteins exported to the cell wall which, even after cleaved, do not release the mature portion to the extracellular milieu, due to one or more hydrophobic motifs causing anchoring to the cell wall, and which are referred to as potentially surface exposed proteins (PSE). Different PSE subcategories exist according to the presence of a carboxy (C) or amino (N) terminal portion anchored to the cell wall, lipoproteins (E), end terminal loops

(L), retention signals-like such as LGxTG, LysM, GW, Choline binding and PG binding (R), in combination or not with other PSE subcategories [6].

The term 'Reverse' from RV can be explained by the reverse genetics (RG) technique. Before the dawn of genomic, there were attempts to discover the responsible genes from a phenotype, reversing the research path of Crick's Central Dogma [7] (DNA → RNA → Protein) discovery. Holding the likely gene sequence, several techniques can be used to identify gene sequence modifications responsible for changes in the organism's phenotype. Crick's Central Dogma principle is also used for RV, as this technique searches within a gene sequence for possible proteins that could act as antigens capable of stimulating an immune response in a host organism [8].

The concept of RV was adapted to fit a new reality of widespread availability of genomic data [9]. With this technique, instead of searching for targets in a single strain or subspecies of an organism, it is now possible to simultaneously research in dozen of genomes, exploring potential joint antigens or exclusive ones to multiple genomes [10]. The availability of a large number of genomes to implement RV has lead to the emergence of the pan-genomics reverse vaccinology concept [11], which can also apply to the concepts of core, extended (dispensable) and character (unique) genomes. While the core genome is composed of exported genes (genes that transcribe for exported proteins) that are common to these multiple strains and could represent candidates for a vaccine, the dispensable genome consists of genes that are absent in at least one of the strains of the studied species and the unique genome consists of genes that are specific to only a particular a strain [10]. From the standpoint of vaccines, the core genome represents to be a good candidate to compose a vaccine that is suitable for all studied strains. In this regard, the first step to enable any pan-genomic reverse vaccinology study is to predict the core genome, along this work denominated *in silico* predicted pan-exoproteome (ISPPE). The model organism here analyzed *(C. pseudotuberculosis)* is a Gram-positive (GRAM+) bacterium, intracellular facultative parasite that affects small ruminants causing a chronic infectious pyogranulomatous disease characterized by the formation of abscesses in lymph nodes [12]. This pathogen infects mainly goats and sheep causing caseous lymphadenitis, but can also infect a huge variety of hosts throughout the world such as camels, horses, cattle, buffaloes, llamas, alpacas and, more rarely, humans [13-18], causing different diseases with different degrees of severity in each of them [12; 19].

Results and discussion

In silico exoproteome prediction schema

As shown in our proposed prediction schema (Figure 1), the software SurfG+ (Surface Gram positive), specially configured for GRAM+ bacteria, is responsible for most of the sub-cellular classifications, which vary between cytoplasmic (CYT), membrane (MEM), SEC and PSE (Figure 2). SurfG+ was configured for GRAM+ bacteria. Figure 1 represents the prediction schema using SurfG+ and three additional software, TatP 1.0 [20], SecretomeP 2.0 [21] and NclassG+ [22], which are specialized in non-classical secretion prediction. SurfG+ incorporates SignalP 3.0 predictor, responsible for identification of classical putative secreted proteins or exported proteins by the SEC pathway [23].

The results obtained after running SurfG+, TapP, SecretomeP and NClassG+ have gave rise to two gene data sets labeled as SEC and PSE, which correspond to the *C. pseudotuberculosis* ISPPE. These ISPPE data sets are composed of putative proteins present fivefold (5x), fourfold (4x), threefold (3x), twofold (2x) or onefold (1x), where fivefold means that a gene was predicted in all five strains, four fold meaning that a gene was predicted in four strains, and so on. A gene fold was obtained by reciprocal blast results, as described in the methods section. Since not all predicted genes are named, it was necessary to create a pan genome identifier, here denominated pan *locus*, to nominate each unique gene fold. The pan *locus* is unique within a pan genome and is shared by all homologous genes. For example, when a putative exported protein was found within the five strains, each gene copy received the same pan *locus* to facilitate further data processing and identification. Following, it was necessary to confirm these results by systematical manual curation of each gene using the ACT tool from the Artemis software package [24]. Once completed this manual curation, it was possible to answer several questions regarding the correctness of each blast result and, as a consequence, it was possible to identify, for instance, that a gene formerly classified as 1x was indeed a 5x, as the other four gene copies were created starting beyond the signal peptide motif. After initial methionine correction, and also taking into account homologous genes, a new prediction step indicated all remaining putative proteins to be exported, composing the core ISPPE. However, gene's start positions incorporating a less probable signal peptide motif were also observed. In general, genes formerly predicted as Nx proved to be correct by manual curation as the remaining (5-N)x genes were predicted as cytoplasmic, PSE or pseudogenes. These results are particularly interesting because they compose the dispensable and unique ISPPE data sets. These genome annotation corrections, as a consequence of these analyses, were incorporated into the official annotation of the five *C. pseudotuberculosis* strains deposited at GenBank in August, 2011. This genomes are also available in the additional file 3, as EMBL files.

Classical and non-classical secreted putative proteins

Figure 3 exhibits the *in silico* predicted pan secretome results for *C. pseudotuberculosis*, which comprise 150 genes, out of 377 from the whole ISPPE, representing 750 *locus_tags* in the five studied *C. pseudotuberculosis* strains. However, despite representing 750 *locus_tags*, not all were predicted as secreted. If at least one gene copy, within a specific pan *locus*, was not predicted as secreted, it still received the same pan *locus* but was not classified as part of the predicted core secretome. There are 122 genes composing the predicted core secretome (5x), followed by 25 genes constituting the predicted dispensable secretome (4x, 3x and 2x) and just 3 genes as the predicted unique secretome (1x). These results were obtained applying the prediction schema from Figure 1; however, different contributions were obtained from different predictors, as shown in Figure 4.

SurfG+ predicted 104 genes, corresponding 85, 18 and 1 to the predicted core, dispensable and unique secretome respectively. On the other hand, TatP predicted 25 genes, of which 17, 7 and 1 corresponded to the predicted core, dispensable and unique secretome respectively. Finally, SecretomeP and NClassG+ predicted 21 genes, corresponding 20 and 1 to the predicted core and unique secretome respectively. It can be easily observed that the main predicted portion is originated by SurfG+, as it predicts putative proteins possibly secreted by the SEC pathway. A considerable portion of genes (~31%), only within the predicted core secretome, comes from non-classical secretion predictors that cannot be ignored when the subject is about vaccine candidates.

The dispensable and unique *C. pseudotuberculosis* predicted secretomes contain ~8%, or 58 *locus_tags*, not predicted as secreted. Putative proteins predicted as CYT, PSE and putative frame shifts (pseudogenes) account for 22, 24 and 10 *locus_tags* respectively. In the dispensable and unique *C. pseudotuberculosis in silico* predicted secretomes, the numbers of genes identified as membrane integral or absent in a genome are insignificant. Nevertheless, the manual curation step ensured no annotation errors in these predictions, making it possible to claim the hypothesis that these differences could be due to environment adaptations.

Potentially surface exposed (PSE) putative proteins

The SurfG+ software was calibrated by the cell wall thickness for each *C. pseudotuberculosis* strain. Figure 5 shows 184 genes, out of 377 from the whole ISPPE, comprising the predicted core surfaceome (5x), 34 genes composing the predicted dispensable surfaceome (4x, 3x and 2x) and just 9 genes as predicted unique surfaceome (1x). These 227 genes account for 1135 *locus_tags* in all five strains. In this set, homologous genes within a pan *locus* do not ever share the same sub-cellular prediction. Genes predicted as MEM, CYT, SEC and putative pseudogenes account for 29, 23, 20 and 17 distinct *locus_tags*, respectively. Genes predicted as MEM (~3%) compose the second major group. This could be explained by the fact that membrane proteins already contain hydrophobic extension and could be more susceptible to expose or occult parts of a protein to the extracellular milieu. However, the same reasoning does not suit to explain the third major group of *locus_tags* with surfaceome pan *locus* that correspond to proteins predicted as secreted ones. These 20 *locus_tags* that were predicted as secreted, but also received surfaceome pan *locus*, raise a question; do these fit SEC or PSE labels? There exist no simple paths to estimate their sub-cellular compartment by software, since some *locus_tags* were predicted as PSE receiving surfaceome pan *locus* and other were predicted as SEC and also received secretome pan *locus*. Ten pan *locus* (plcppse193, plcppse194, plcppse205, plcppse218, plcppse226, plcpsec096, plcpsec097, plcpsec098, plcpsec100, plcpsec101) faces this question, as some genes appear in both the predicted secretome and surfaceome.

The PSE subcategories show predominance of genes, as presented in Figure 6. Most of the 1045 genes predicted as PSE are cell wall anchored outward C-terminal (~40%) (≥ 50 AA long), followed by lipoproteins (~24%), outward loops (~11%) (≥ 100 AA long) and outward N-terminal (~17%) (≥ 50 AA long), whereas genes containing retention signals (PSE R) account only for ~8%.

The PSE results of all strains were analyzed considering that a significant cell wall thickness difference between strain I19 and the other ones was observed (~34 nm versus ~24 nm). Despite the significant cell wall thickness difference, a small difference was predicted in the genome, which accounts for a decrease in the number of PSE and an increase in the number of MEM genes in *C. pseudotuberculosis* strain I19.

Revised *in vitro* exoproteome results

The 104 observed genes in both TPP/LC-MS^E [25] and 2-DE-MALDI-TOF/TOF [26] experiments were compared with the ISPPE results here presented. This comparison, explained in the methods section, brought novel insights into the *in vitro* exoproteome and showed the possibility of having additional genes in the main *C. pseudotuberculosis in vitro* exoproteome. In Table 1 are listed all 35 proteins of the variant *in vitro* exoproteome (strains 1002 and C231), that correspond to ~23% of the total amount. These proteins were found to be highly conserved in the five compared *C. pseudotuberculosis* strains and comprise the core ISPPE. Moreover, it was verified that three proteins (ADL20466, ADL20097 e ADL19973), previously classified as belonging to the variant *in vitro* exoproteome of strains 1002 [25], did actually belong to the main *in vitro* exoproteome [26]. These findings give raise to the possibility that more proteins of the variant *in vitro* exoproteome indeed make part of the main *in vitro* exoproteome.

This comparison also served as a rebuttal argument against some specific genes. The Cp1002_0369 gene, classified under the plcpsec100 pan *locus* as a pseudogene, was identified by the *in vitro* exoproteome experiment. Interestingly, this gene copy also suits the plcppse226 pan *locus*. Both pan *locus* make part of previous related genes that already showed difficulties to be classified, by software, into any potential sub-cellular compartment, as some genes within the pan *locus* fit both SEC and PSE labels. The *in silico* predictions enforces that there are at least three secreted proteins, inspite of the other two gene copies being predicted as having PSE and CYT labels. Furthermore, the genes plcppse180, plcppse192, plcpsec077, plcpsec095 and plcpsec099 also had both genes found in the main *in vitro* exoproteome of strains 1002 and C231, but were not classified in the ISPPE. The plcppse180 pan *locus* holds a putative pseudogene (CpPAT10_0459), and is therefore not present in the *in silico* predicted core surfaceome. Other genes were predicted as cytoplasmic. It is possible that these genes were wrongly assembled since there is evidence that at least two homologous genes, from strains 1002 and C231, are exported to the extracellular milieu.

Core *C. pseudotuberculosis* ISPPE candidates homologous to *Mtb*

Within the core *C. pseudotuberculosis* ISPPE, homologous genes to those of the previously studied *Mycobacterium tuberculosis* H37Rv (*Mtb*) were observed. In this work we present some of these homologous genes featuring at least 90% protein alignment and 50% identity within this alignment. These cut-offs were obtained during the search for *C. pseudotuberculosis* homologous genes in the *Mtb* genome.

The core *C. pseudotuberculosis* ISPPE, that accounts for ~81% of the total, is composed of 306 genes or 1,530 distinct *locus_tags*, being ~40% predicted as SEC and ~60% predicted as PSE proteins, of which 20 genes present high similarity to *Mtb*'s genes (Table 2); however, not all of these *Mtb* genes have known functions.

In this regard, here we only discuss some of these *Mtb*'s genes with experimental evidence. The plcppse174 pan *locus* shows 51% protein identity with Rv3915 (YP_178027.1), a gene named *cwlM* that was the first autolysin gene identified and cloned from *Mtb*. This finding offers a new drug target class that could alter the permeability of the mycobacterium cell wall and enhance the effectiveness of treatments for tuberculosis [27]. Applying principles of *in vivo* expression technology (IVET), it was possible to identify upregulated genes from *Mtb* in an *in vitro* simulation of anaerobic persistence condition. The upregulated genes under hypoxic condition (dissolved oxygen <1%) include Rv0050 (*ponA1*), a penicillin binding protein that has 52% protein identity to the plcppse165 pan *locus* and 90% alignment extension [28]. The plcpsec122 pan *locus* shows ~58% protein identity with Rv2752c (NP_217268.1), a unique bi-functional *Mtb* gene that owns both β-lactamase and RNase activities. Both activities are lost upon deletion of the 100 AA long C-terminal 100 tail, which contains an additional loop when compared to the RNase J of *Bacillus subtilis* [29]. As it can be observed, the plcppse080 pan *locus* appears twice in Table 2, as it is homologous to both NADH dehydrogenase gene copies of *Mtb*, *ndh* (NP_216370.1) and *ndhA* (NP_214906.1), with ~57% protein identity. In *Mtb*, energy generation is mainly performed by type II dehydrogenases *ndh* and *ndhA*, being both, as such, essential genes [30].

The plcpsec113 pan *locus* is homologous to the *glmU* gene (NP_215534.1), holding ~59% protein identity and more than 90% alignment extension. This gene is essential in *Mtb*, being required for optimal bacterial growth, and has been selected as a possible drug target for structural and functional investigation [31]. *GlmU* is a bifunctional acetyltransferase/uridyltransferase that catalyses the formation of UDP-GlcNAc from GlcN-1-P. UDP-GlcNAc is the substrate for two important biosynthetic pathways: lipopolysaccharide and peptidoglycan synthesis. Due to its important roles, *glmU* had its conformational structure solved [31]. The plcpsec113 pan *locus* for *C. pseudotuberculosis* is an interesting putative drug candidate since it is predicted to be secreted, part of the core ISPPE and is able to infer its conformational structure by homology modeling using *Mtb glmU*.

Several genes involved in mannoglycoconjugate biosynthesis have shown to be involved in virulence, due to their central role in biosynthesis of major surface-associated glycoconjugates. Within these genes, the *Mtb* gene *manB* (Rv3264c) is defined as a GDP-mannose pyrophosphorylase (GDPMP) and disruption of its activity leads to decrease of surface-associated mannosylated lipoglycans. For GDPMP, this decrease correspond directly to reduced virulence in both BALB/c mice and cultured human macrophages [32]. The *Mtb manB* gene holds 69% protein identity to the plcpsec110 pan *locus* and more than 90% alignment extension, making plcpsec110 a considerable putative drug target.

Mycolic acids and multimethyl-branched fatty acids are found uniquely in the cell envelope and are essential for survival, virulence and antibiotic resistance of *Mtb*. Acyl-CoA carboxylases (ACCases) commit acyl-CoAs to the biosynthesis of these unique fatty acids. Previous studies indicate that AccD5 is important for cell envelope lipid biosynthesis and its disruption leads to pathogen death [33]. The *Mtb* gene *accD5* (NP_217797.1) had its structure determined and also shows ~74% protein identity to the plcppse045 pan *locus* in more than 90% alignment extension, making it also a promising candidate for further vaccine candidate evaluations.

Moreover, it was demonstrated that *Mtb* can use heme as an iron source, suggesting that *Mtb* contains a yet-unknown heme acquisition system [34]. We found that the *C. pseudotuberculosis* plcpsec076 pan *locus* holds ~52% protein identity to the *Mtb* gene *hemE* (NP_217194.1) and more than 90% alignment size, therefore also representing an interesting drug target for *C. pseudotuberculosis*.

Candidates filtering

The here presented results provide a plethora of putative vaccine candidates never seen before for *C. pseudotuberculosis*. However, genes predicted as MEM and CYT account respectively for 18% and 65% of the *in silico* predicted pan genome. Despite the 227 surfaceome and 150 secretome genes here presented, these only represents ~16% of the *C. pseudotuberculosis in silico* predicted pan genome. Most of the genes remain inaccessible for the current *in silico* prediction techniques and it is possible that these neglected genes could also be good candidates against *C. pseudotuberculosis*. These findings raise the need for more elaborated and driven software or prediction schemas capable of uncovering these major genome neglected portions. Using the prediction schema here presented, it was possible to include more than ~2% of non-classic secreted putative proteins that compose putative vaccine candidates. However, this low income amount of vaccine candidates is due to the optional parameter selected in our prediction schema, the non-classic secreted score greater than or equal 0.90. If using the default parameter from the software secretomeP and NClassG+, this income would be increased up to ~6% and the final income of putative vaccine candidates would be ~20%, using a couple of motifs predictors as depicted in Figure 1. The current reverse vaccinology software allows obtaining a number of candidates closer to 20% of the *C. pseudotuberculosis* genome. These considerations raise a question: supposing that novel software for unexplored secretion pathways come into scenario, what is the genome's percentage that could be selected as putative vaccine candidates? Supposing that this percentage reaches 40%, how could the problem of choosing between almost one thousand putative vaccine candidates to be used for the next vaccine production stage for *C. pseudotuberculosis* be solved? This dilemma could be solved by using further software prediction just like those addressing epitopes MHC class I and II allele affinity [35]; however, this could be just a part of the solution. There are chances of solving this dilemma by means of broader vaccine projects, which would take into account particular variables for each target organism in order to minimise research efforts and the number of possible vaccine candidates [36].

In silico versus **non-*in silico***

It is broadly known that *in silico* genome investigations could give evidence about the genome's function and structure. It is also known that such *in silico* investigations could only be proved or denied by non-*in silico* experiments. Therefore, such reasonable thinking is not a single-hand avenue. Non-*in silico* experiments could be improved by means of more comprehensive or specific approaches with the objective of getting a closer answer to the reality for biological questions. The fact is that *in silico* analyses cannot vary when executed over and over again and no matter how many folds are run. We know that exactly 122 genes will be always predicted as having classical exportation motifs; on the other hand, we cannot expect the same behavior from non-*in silico* analysis. Some real proteins could be or not be found in an *in vitro* or *in vivo* exoproteome result, due to an uncountable number of factors [21]. Therefore, we suggest that the core *C. pseudotuberculosis* ISPPE could be composed of a larger number of predicted genes, but such confirmation could only be affirmed with additional non-*in silico* exoproteome experiments.

Conclusions

The *in silico* pan-exoproteome prediction methodology applied to the pathogen *C. pseudotuberculosis* helps to raise new insights into putative vaccine candidates against CLA. Additional investigations of the *in vitro* exoproteome of two strains of *C. pseudotuberculosis*, 1002 and C231, showed evidence that the major part of the variant *in vitro* exoproteome is contained in the core ISPPE. A simultaneous curation of the *in silico* predicted core secretome and surfaceome within the five *C. pseudotuberculosis* strains also contributed to homogenize the genome annotations and it was possible to fix the most probable putative methionine proteins. Moreover, putative miss assembled genes, formerly classified as pseudogenes by *in silico* analyses, were also revised. The efforts to create a *C. pseudotuberculosis* ISSPE catalogue proved to be necessary and computationally viable to ensure a uniform set of putative vaccine candidates free of annotation errors.

Methods

Genomes

The analyzed *C. pseudotuberculosis* genomes were obtained from the NCBI website according to the following accession numbers: EMBL: CP001809 (strain 1002), EMBL: CP001809 (strain C231), EMBL: CP002251 (strain I19), EMBL: CP002924 (strain PAT10) and EMBL: NC_014329 (strain FRC41).

Prediction schema

Predicted genes from all five *C. pseudotuberculosis* strain genomes were exported as amino acid fasta files using the Artemis software. These fasta files were passed as parameters to SurfG+ 1.0 (Figure 1), and lists of genes predicted as CYT, SEC, PSE and MEM were created by this software. Genes formerly predicted as CYT by SurfG+ were then submitted to the TapP 1.0 predictor; when a Tat motif was found, the putative protein was automatically classified as SEC, otherwise, another prediction round would took place using two other non-classic secretion predictors, SecretomeP 2.0 and NclassG+ 1.0. With a positive prediction from both software and a prediction score greater than or equal to 0.90, the genes were automatically classified as SEC. The SEC and PSE data sets were finally submitted to a reciprocal blastp processing and posterior filtering, giving rise to the fivefold categories according to folds occurring in each strain: 5x, 4x, 3x, 2x and 1x. The results were then manually curated using the ACT software and strain 1002, the first to be sequenced and annotated. The strain 1002 was disposed, in ACT software, in the middle of two pairs of the other two genome strains, facilitating to exhibit differences among all of them.

SurfG+ 1.0

Sub-cellular localization prediction of *C. pseudotuberculosis* putative proteins was made by *in silico* analysis using the SurfG+ 1.0 software [6]. SurfG+ is a pipeline for protein sub-cellular prediction that incorporates common software, such as SignalP, LipoP and TMHMM to search for motifs. It also creates novel HMMSEARCH profiles to predict cell wall retention signals. SurfG+ starts searching, in the following order for: retention signals, lipoproteins, SEC pathway export motifs and transmembrane motifs. If none of these motifs are found in a protein sequence, it is then characterized as CYT. A novel possible characterization introduced by SurfG+ is its ability to better distinguish between MEM and PSE, by informing an expected cell wall thickness in amino acids. Using the literature or an electronic microscopy it is possible to estimate a reasonable cell wall thickness value for prokaryotic organisms. By means of this last option, *C. pseudotuberculosis* genes were classified into four different sub-cellular locations: CYT, MEM, PSE, or SEC.

TatP 1.0 Server

Twin-arginine signal peptide motifs were predicted using the on line server hosted by http://www.cbs.dtu.dk/services/TatP/ [20]. Only putative proteins formerly classified as CYT by SurfG+ were submitted to the TatP analyses. There were no intersections between SignalP and TatP predictions.

SecretomeP 2.0 and NClassG+ 1.0

Non-classical secreted putative proteins were predicted using the online server hosted by http://www.cbs.dtu.dk/services/SecretomeP/ [21]. NClassG+ [22], a second non-classical secreted protein predictor, was also used; however, the predictions were directly performed contacting the software authors. This double check prediction ensured greater accuracy. Only those genes formerly classified as CYT by SurfG+ and without the twin-arginine signal peptide motifs were submitted to a non-classical secreted analysis. Despite the significant scores of SecretomeP and Nclass+, ranging between 0.5 and 1.0, only those genes with a score greater than or equal to 0.9 were selected, in order to ensure a minimal false positive in future wet lab experiments, the focus of our research group.

Pan genome

To predict the *C. pseudotuberculosis* pan genome, reciprocal blastp results were used. All the putative proteins predicted as SEC were put apart in a single amino acid fasta file to make a reciprocal blast. A similar file was also created for the proteins predicted as PSE. To avoid homologous mismatches, the blastp results obtained using the PAM70 substitution matrix and the 10^{-6} e-value were manually filtered. In this regard, the first step was to establish the alignment size and identity percentages of cut-offs, being 89.58 and 50.00%, respectively, for SEC putative proteins, whereas for PSE putative proteins, these cut-offs were 88.16 and 48.80%, respectively. Identity percentages closer to 50% are explained by frame shifts not annotated until this work.

All the putative proteins from the five strains (query) with alignment size and identity percentages higher than these cut-offs had no more than one group of blast hits (subject) against the others strains. Moreover, within each of these blast hits groups, there was a blast hit from the query protein against it self as subject. The results were manually curated using the ACT software, from the Artemis package [24], using the strain 1002 as reference strain for the other two strains. This ACT view was composed by strains C231-1002-I19 and FRC41-1002-I19. Each putative protein predicted as SEC and PSE was compared against their other four homologues for correct initial methionine, frame shifts and finally annotating the correct sub-cellular location.

Revised *in vitro* exoproteome results

In lists 1 and 2 of the annex are both gene *locus* present in the *C. pseudotuberculosis* ISPPE, together with the quantity of homologous genes present in the all five genomes. These results were inserted in a relational database, denominated *C. pseudotuberculosis* Data Base (CpDB) [37], in a specific table called ‘exopred’. The list of the *in vitro* exoproteome proteins [25; 26] was also inserted to the CpDB into a table called ‘exo’ that discriminates the identification of each protein regarding GenBank (protein id), as well as in which strains it is found. To make a relationship between the ‘exopred’ and ‘exo’ tables, a third table of the CpDB, called ‘gene’, which contains all the functional annotation of the genomes of *C. pseudotuberculosis*, was created. The CpDB is the repository of the pan genome of *C. pseudotuberculosis*, harbouring the genomes since their initial genomic prediction, deposited in the NCBI, as well as the annotation corrections for future deposits. For this last purpose, the CpDB stores the identification of each protein according to the GenBank. In this way, it is possible to make a link between the three tables in the form of a clause of JOIN of the SQL: "*... WHERE gene.locus_tag = exopred.locus_tag AND gene.protein_id = exo.protein_id AND exopred.pangenome_coverage = '5x' ...*". This clause returns the registries of the CpDB whose *locus_tag* in the gene table is equal to the *locus_tag* of the explored table, being this same gene in the protein_id field in the exo table with prediction of belonging to all five genomes. Other conditions can also be included, such as for example, restraining the results to specific genes of a *C. pseudotuberculosis* strain or simultaneously present in the exoproteome of specific strains.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VA encouraged the research, BMC application, provided references, applied biological knowledge and gave final approval of the version to be published. ARS conducted all the software analyses, manually corrected annotation errors in the five genomes, developed the prediction schema and wrote the paper. EB contributed to the manually curation of all pseudogenes from all bacterial strains. MZT made substantial contributions to the design and interpretation of the manuscript. UP, AG, FD, FS, AC, AP, DB, FF, LC, LG, RR, SA, SS, VCA, AS and AM have given final approval of the version to be published.

Acknowledgements

CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil.

References

1 Rappuoli R: **Reverse vaccinology**. *Curr Opin Microbiol* 2000, **3**:445-450.

2 Sibbald MJJB, van Dij JML: **Secretome Mapping in Gram-Positive Pathogens. In Karl Wooldridge (ed.), Bacterial Secreted Protein: Secretory Mechanisms and Role in Pathogenesis**. *Caister Academic Press* 2009, :193-225.

3 Simeone R, Bottai D, Brosch R: **ESX/type VII secretion systems and their role in host-pathogen interaction**. *Curr Opin Microbiol* 2009, **12**:4-10.

4 Stavrinides J, McCann HC, Guttman DS: **Host-pathogen interplay and the evolution of bacterial effectors**. *Cell Microbiol* 2008, **10**:285-292.

5 Bhavsar AP, Guttman JA, Finlay BB: **Manipulation of host-cell pathways by bacterial pathogens**. *Nature* 2007, **449**:827-834.

6 Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M: **Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria**. *Proteomics* 2009, **9**:61-73.

7 Strasser BJ: **A world in one dimension: Linus Pauling, Francis Crick and the central dogma of molecular biology**. *Hist Philos Life Sci* 2006, **28**:491-512.

8 Rappuoli R: **IS15 Developing vaccines in the era of genomics and toll receptors**. *Immunology* 2005, **116**:1.

9 Rinaudo CD, Telford JL, Rappuoli R, Seib KL: **Vaccinology in the genome era**. *J Clin Invest* 2009, **119**:2515-2525.

10 Lapierre P, Gogarten JP: **Estimating the size of the bacterial pan-genome**. *Trends Genet* 2009, **25**:107-110.

11 Bambini S, Rappuoli R: **The use of genomics in microbial vaccine development**. *Drug Discov Today* 2009, **14**:252-260.

12 Dorella FA, Pacheco LG, Seyffert N, Portela RW, Meyer R, Miyoshi A, Azevedo V: **Antigens of Corynebacterium pseudotuberculosis and prospects for vaccine development**. *Expert Rev Vaccines* 2009, **8**:205-213.

13 Afzal M, Sakir M, Hussain MM: **Corynebacterium pseudotuberculosis infection and lymphadenitis (taloa or mala) in the camel**. *Trop Anim Health Prod* 1996, **28**:158-162.

14 Aleman M, Spier SJ, Wilson WD, Doherr M: **Corynebacterium pseudotuberculosis infection in horses: 538 cases (1982-1993)**. *J Am Vet Med Assoc* 1996, **209**:804-809.

15 Silva A, Schneider MPC, Cerdeira L, Barbosa MS, Ramos RTJ, Carneiro AR, Santos R, Lima M, D'Afonseca V, Almeida SS, Santos AR, Soares SC, Pinto AC, Ali A, Dorella FA, Rocha F, de Abreu VAC, Trost E, Tauch A, Shpigel N, Miyoshi A, Azevedo V: **Complete genome sequence of Corynebacterium pseudotuberculosis I19, a strain isolated from a cow in Israel with bovine mastitis**. *J Bacteriol* 2011, **193**:323-324.

16 Selim SA: **Oedematous skin disease of buffalo in Egypt**. *J Vet Med B Infect Dis Vet Public Health* 2001, **48**:241-258.

17 Peel MM, Palmer GG, Stacpoole AM, Kerr TG: **Human lymphadenitis due to Corynebacterium pseudotuberculosis: report of ten cases from Australia and review**. *Clin Infect Dis* 1997, **24**:185-191.

18 Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, Husemann P, Stoye J, Dorella FA, Rocha FS, Soares SDC, D'Afonseca V, Miyoshi A, Ruiz J, Silva A, Azevedo V, Burkovski A, Guiso N, Join-Lambert OF, Kayal S, Tauch A: **The complete genome sequence of Corynebacterium pseudotuberculosis FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights**

into gene-regulatory networks contributing to virulence. *BMC Genomics* 2010, **11**:728.

19 Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AAMC, Lopes DO, Dorella FA, Pacheco LGC, Costa MP, Turk MZ, Seyffert N, Moraes PMRO, Soares SC, Almeida SS, Castro TLP, Abreu VAC, Trost E, Baumbach J, Tauch A, Schneider MPC, McCulloch J, Cerdeira LT, Ramos RTJ, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, Pedrosa AL, Vieira CU, Guimarães CT, Bartholomeu DC, Oliveira DM, Santos FR, Rabelo EM, Lobo FP, Franco GR, Costa AF, Castro IM, Dias SRC, Ferro JA, Ortega JM, Paiva LV, Goulart LR, Almeida JF, Ferro MIT, Carneiro NP, Falcão PRK, Grynberg P, Teixeira SMR, Brommonschenkel S, Oliveira SC, Meyer R, Moore RJ, Miyoshi A, Oliveira GC, Azevedo V: **Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two Corynebacterium pseudotuberculosis Strains**. *PLoS ONE* 2011, **6**:e18551.

20 Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: **Prediction of twin-arginine signal peptides**. *BMC Bioinformatics* 2005, **6**:167.

21 Bendtsen JD, Kiemer L, Fausbøll A, Brunak S: **Non-classical protein secretion in bacteria**. *BMC Microbiol* 2005, **5**:58.

22 Restrepo-Montoya D, Pino C, Nino LF, Patarroyo ME, Patarroyo MA: **NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins**. *BMC Bioinformatics* 2011, **12**:21.

23 Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat Methods* 2011, **8**:785-786.

24 Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B: **Artemis: sequence visualization and annotation**. *Bioinformatics* 2000, **16**:944-945.

25 Pacheco LGC, Slade SE, Seyffert N, Santos AR, Castro TLP, Silva WM, Santos AV, Santos SG, Farias LM, Carvalho MAR, Pimenta AMC, Meyer R, Silva A, Scrivens JH, Oliveira SC, Miyoshi A, Dowson CG, Azevedo V: **A combined approach for comparative exoproteome analysis of Corynebacterium pseudotuberculosis**. *BMC Microbiol* 2011, **11**:12.

26 Silva WM, Seyffert N, Castro TLP, Santos AV, Pacheco LGC, Santos AR, Ciprandi A, Zurita-Turk M, Dorella FA, Andrade HM, Pimenta AMC, Silva A, Miyoshi A, Azevedo V: **Identification of 11 new exoproteins of Corynebacterium pseudotuberculosis through comparative analysis of the secretome**. *Advances in Integrative Omics and Applied Biotechnology* 2012, **1**:22.

27 Deng LL, Humphries DE, Arbeit RD, Carlton LE, Smole SC, Carroll JD: **Identification of a novel peptidoglycan hydrolase CwlM in Mycobacterium tuberculosis**. *Biochim Biophys Acta* 2005, **1747**:57-66.

28 Saxena A, Srivastava V, Srivastava R, Srivastava BS: **Identification of genes of Mycobacterium tuberculosis upregulated during anaerobic persistence by fluorescence and kanamycin resistance selection**. *Tuberculosis (Edinb)* 2008, **88**:518-525.

29 Sun L, Zhang L, Zhang H, He Z: **Characterization of a Bifunctional β-Lactamase/Ribonuclease and Its Interaction with a Chaperone-Like Protein in the Pathogen Mycobacterium tuberculosis H37Rv**. *Biochemistry (Moscow)* 2011, **76**:350-358.

30 Velmurugan K, Chen B, Miller JL, Azogue S, Gurses S, Hsu T, Glickman M, Jacobs WRJ, Porcelli SA, Briken V: **Mycobacterium tuberculosis nuoG is a virulence gene that inhibits apoptosis of infected host cells**. *PLoS Pathog* 2007, **3**:e110.

31 Zhang Z, Bulloch EMM, Bunker RD, Baker EN, Squire CJ: **Structure and function of GlmU from Mycobacterium tuberculosis**. *Acta Crystallogr D Biol Crystallogr* 2009, **65**:275-283.

32 McCarthy TR, Torrelles JB, MacFarlane AS, Katawczik M, Kutzbach B, Desjardin LE, Clegg S, Goldberg JB, Schlesinger LS: **Overexpression of Mycobacterium tuberculosis manB, a phosphomannomutase that increases phosphatidylinositol mannoside biosynthesis in Mycobacterium smegmatis and mycobacterial association with human macrophages**. *Mol Microbiol* 2005, **58**:774-790.

33 Lin T, Melgar MM, Kurth D, Swamidass SJ, Purdon J, Tseng T, Gago G, Baldi P, Gramajo H, Tsai S: **Structure-based inhibitor design of AccD5, an essential acyl-CoA carboxylase carboxyltransferase domain of Mycobacterium tuberculosis**. *Proc Natl Acad Sci U S A* 2006, **103**:3072-3077.

34 Jones CM, Niederweis M: **Mycobacterium tuberculosis can utilize Heme as an iron source**. *J Bacteriol* 2011, **193**:1767-1770.

35 Davies MN, Flower DR: **Harnessing bioinformatics to discover new vaccines**. *Drug Discov Today* 2007, **12**:389-395.

36 Santos A, Ali A, Barbosa E, Silva A, Miyoshi A, Barh D, Azevedo V: **The reverse vaccinology – A contextual overview**. *IIOABJ* 2011, **2**:8-15.

37 Azevedo V, Santos AR, Soares S, Ali A, Pinto A, Magalhaes A, Barbosa E, Ramos R, Cerdeira L, Carneiro A, Abreu V, Almeida S, Schneider P, Silva A,

Miyoshi A: **Automated functional annotation**. In *Bioinformatics - trends and methodologies*. Volume 1. . Edited by Mahdavi MA. InTechOpen; 2011:722.

Figures

Figure 1 – *C. pseudotuberculosis* pan genomic prediction schema

Software used, identified sub-cellular compartments and flow scheme to create the final pan genomic data sets.

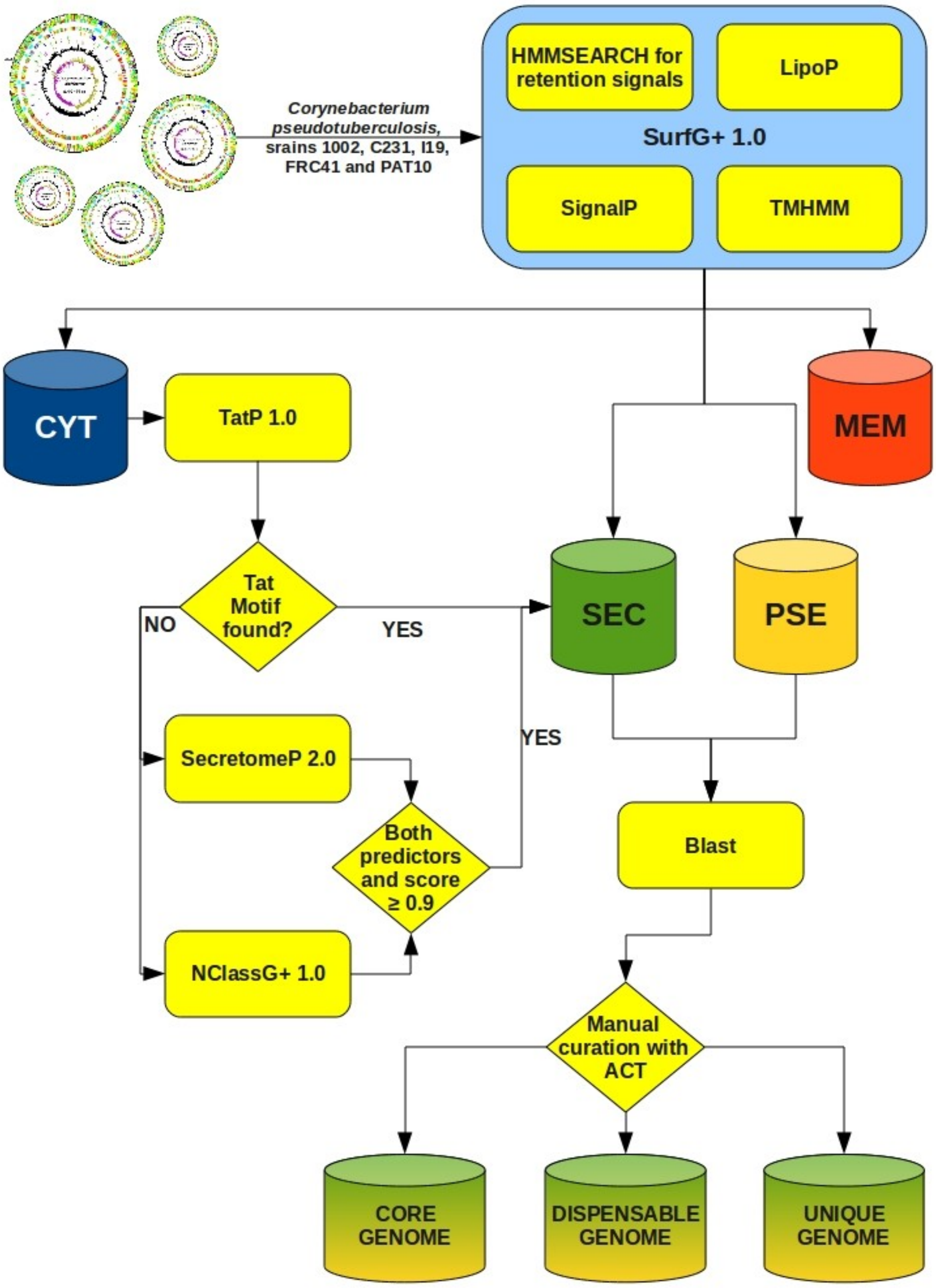


Figure 2 – Predicted gene quantities by sub-cellular compartment from full *C. pseudotuberculosis* genomes

Classification of more than 10,000 distinct genes from the five different *C. pseudotuberculosis* strains in the four sub-cellular categories: cytoplasmic (CYT), membrane (MEM), potentially surface exposed (PSE) and secreted (SEC). Predictions were made using the schema presented in Figure 1.

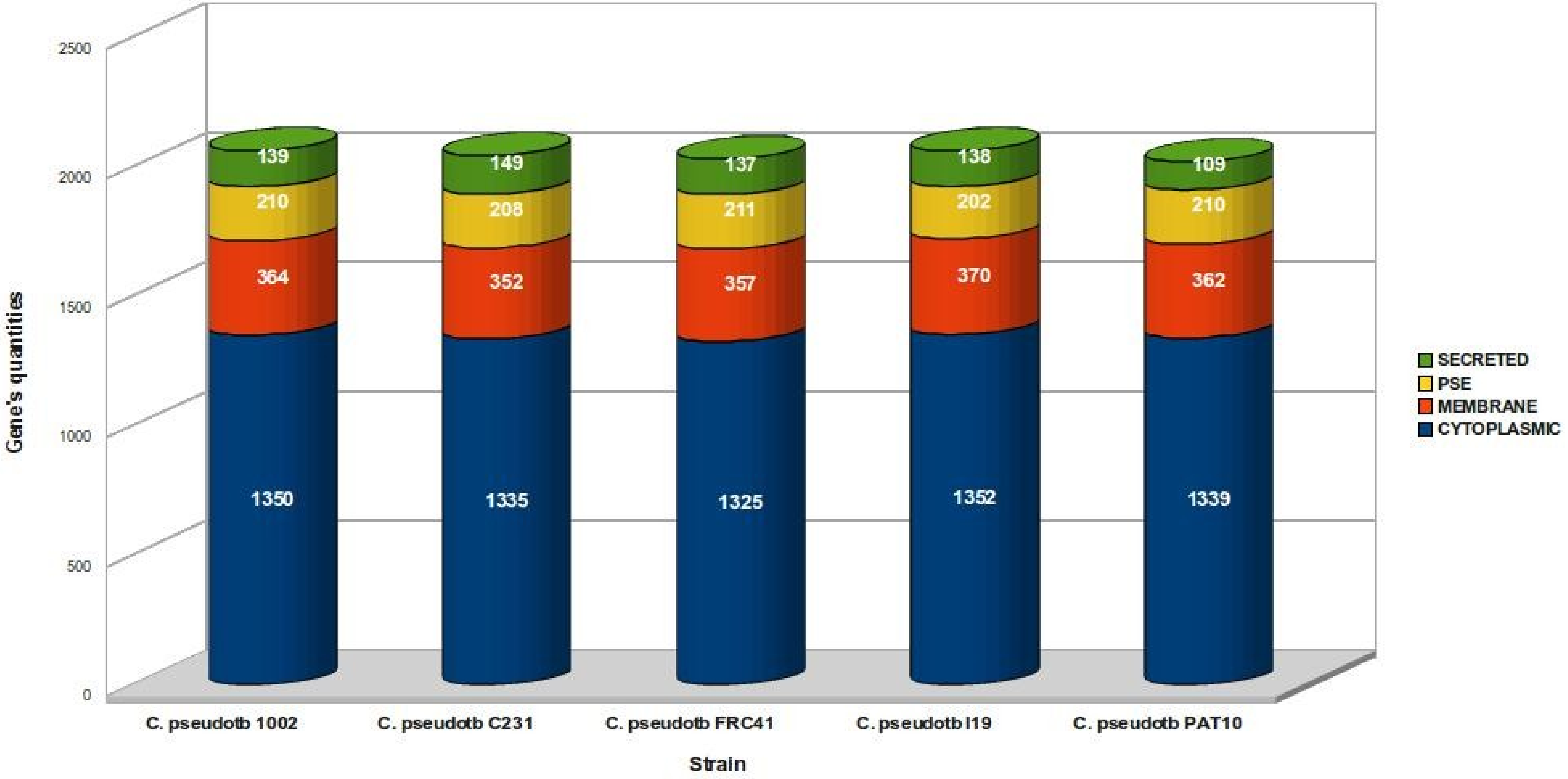


Figure 3 – Predicted *C. pseudotuberculosis* pan secretome

Predictions for 150 genes from strains 1002, C231, I19, FRC41 and PAT10 made by SurfG+ 1.0, TatP 1.0 Server and SecretomeP 2.0 Server.

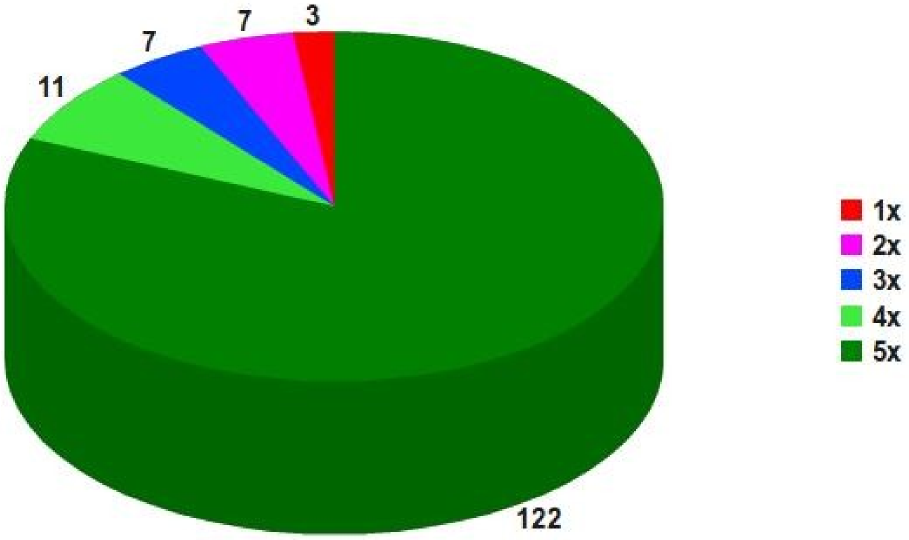


Figure 4 – Predicted *C. pseudotuberculosis* pan secretome by predictor software

Predicted secreted genes coverage in the predicted pan secretome of the five bacterial strains separated by predictor software SurfG+, TatP and SecretomeP.

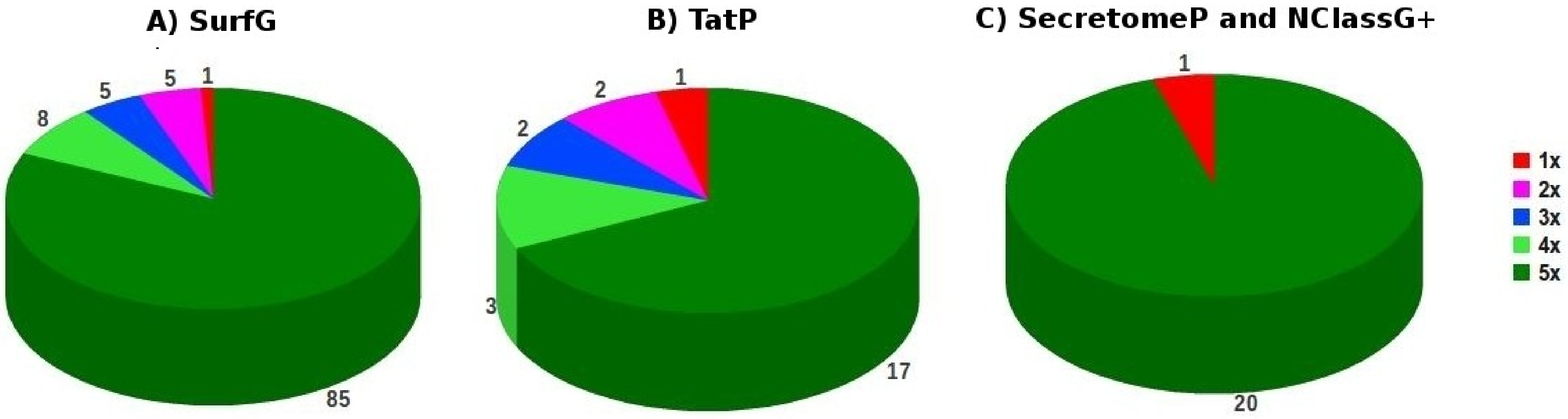


Figure 5 – Predicted *C. pseudotuberculosis* pan surfaceome

Pan surfaceome predictions for 227 genes from strains 1002, C231, I19, FRC41 and PAT10, performed by SurfG+ 1.0.

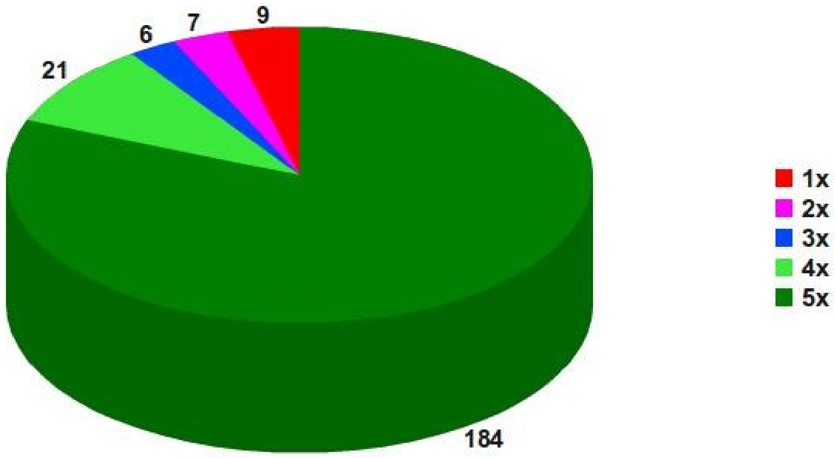
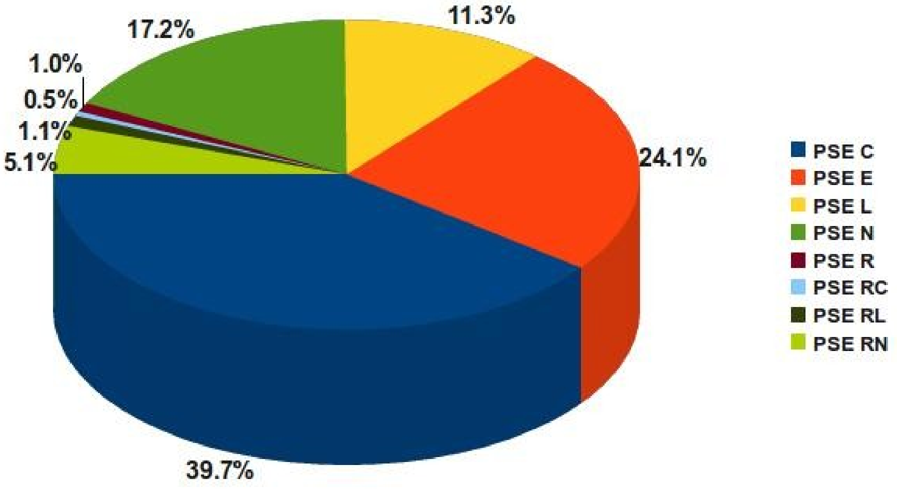


Figure 6 – Predicted *C. pseudotuberculosis* pan surfaceome by PSE subcategories

PSE categories are distributed in outward C-terminal or N-terminal portion greater than or equal 50 AA. Outward N or C terminal greater than 100 AA are classified as L. Lipogenes identified by LipoP are classified as E and retention signals identified by HMMSEARCH profiles are classified as R. These labels can also be conjugated to create other PSE subcategories.



Tables

Table 1 – Core *C. pseudotuberculosis in silico* predicted pan-exoproteome found in the variant *in vitro* exoproteome

The 35 proteins listed in this table were not found in the experimental main *in vitro* exoproteome [25; 26] but were found in the *in silico* predicted pan-exoproteome of all five *C. pseudotuberculosis* strains.

Protein identifier	*locus_tag*	Gene name	Product	Predicted local sub-cellular	GenBank organism identifier
ADL19972	Cp1002_0064		Hypothetical protein	PSE E	CP001809
ADL20140	Cp1002_0237	*slpA*	Surface layer protein A	SEC	CP001809
ADL20222	Cp1002_0320		Hypothetical protein	PSE N	CP001809
ADL20288	Cp1002_0388		L,D-transpeptidase catalytic domain, region YkuD	SEC	CP001809
ADL20391	Cp1002_0497	*malE*	Maltose/maltodextrin transport system substrate-binding protein	PSE E	CP001809
ADL20455	Cp1002_0562	*sprT*	Trypsin	PSE C	CP001809
ADL20477	Cp1002_0584	*cynT*	Carbonic anhydrase	PSE E	CP001809
ADL20508	Cp1002_0615		Hypothetical protein	SEC	CP001809
ADL20574	Cp1002_0681	*rpfB*	Resuscitation-promoting factor RpfB	SEC	CP001809
ADL20656	Cp1002_0766		Hypothetical protein	SEC	CP001809
ADL21028	Cp1002_1144	*yceG*	Amino deoxychorismate lyase	SEC	CP001809
ADL21239	Cp1002_1362		Hypothetical protein	PSE E	CP001809
ADL21302	Cp1002_1425	*ctaC*	Cytochrome c oxidase subunit II	PSE C	CP001809
ADL21537	Cp1002_1669		Hypothetical protein	SEC	CP001809
ADL21667	Cp1002_1802	*lipY*	Secretory lipase	SEC	CP001809
ADL09524	CpC231_0025	*pld*	Phospholipase D	SEC	CP001829
ADL09532	CpC231_0033	*pbpA*	Penicillin-binding protein A	SEC	CP001829
ADL09691	CpC231_0196		Hypothetical protein	SEC	CP001829
ADL09697	CpC231_0203	*pbpB*	Penicillin binding protein transpeptidase	SEC	CP001829

Protein identifier	***locus_tag***	**Gene name**	**Product**	**Predicted local sub-cellular**	**GenBank organism identifier**
ADL09852	CpC231_0360	*oppA1*	Oligopeptide-binding protein oppA	PSE E	CP001829
ADL09871	CpC231_0379		Hypothetical protein	SEC	CP001829
ADL09872	CpC231_0380	*malE*	Maltotriose-binding protein	PSE E	CP001829
ADL09990	CpC231_0503	*lytR*	Transcriptional regulator lytR	PSE C	CP001829
ADL10248	CpC231_0766		Hypothetical protein	SEC	CP001829
ADL10460	CpC231_0982	*ciuA*	Iron ABC transporter substrate-binding protein	PSE E	CP001829
ADL10489	CpC231_1012	*yceI*	Protein yceI	SEC	CP001829
ADL10626	CpC231_1150		Zinc metallopeptidase	PSE C	CP001829
ADL10663	CpC231_1187		Lipoprotein	PSE E	CP001829
ADL10880	CpC231_1409	*pknL*	Serine/threonine protein kinase	PSE N	CP001829
ADL11196	CpC231_1737		Corynomycolyl transferase	SEC	CP001829
ADL11213	CpC231_1756		Hypothetical protein	SEC	CP001829
ADL11326	CpC231_1871		Hypothetical protein	PSE N	CP001829
ADL11338	CpC231_1885		Membrane protein	SEC	CP001829
ADL11339	CpC231_1886		Hypothetical protein	SEC	CP001829
ADL11410	CpC231_1959	*glpQ*	Glycerophosphoryl diester phosphodiesterase	PSE E	CP001829

Table 2 – Core *C. pseudotuberculosis in silico* predicted pan-exoproteome homologous to *Mtb*'s proteins

Related *C. pseudotuberculosis*'s proteins containing at least 50% amino acid identity and 90% alignment size to the *Mtb* H37Rv's proteins.

Corynebacterium pseudotberculosis					***Mycobacterium tuberculosis***				
In silico pan exopro-teome data set	**pan *locus***	**Reference genome *locus_tag***	**ORF size**	**Percentage of amino acid alignment's identity**	**ORF size**	***locus_tag***	**Gene name**	**protein ID**	**Annotated product**
core	plcpsec106	cpfrc_00104	488	69.10	461	Rv3790		NP_218307.1	oxidoreductase
core	plcpsec076	cpfrc_00276	371	51.56	357	Rv2678c	*hemE*	NP_217194.1	uroporphyrinogen decarboxylase
core	plcppse023	cpfrc_00283	535	52.51	529	Rv0528		NP_215042.1	transmembrane protein
core	plcppse045	cpfrc_00491	543	73.72	548	Rv3280	*accD5*	NP_217797.1	propionyl-CoA carboxylase beta chain
core	plcpsec110	cpfrc_00506	362	69.03	359	Rv3264c	*manB*	YP_177951.1	D-alpha-D-mannose-1-phosphate guanylyltransferase MANB (D-alpha-D-heptose-1-phosphate guanylyltransferase)
core	plcpsec111	cpfrc_00508	151	51.45	139	Rv3259		NP_217776.1	hypothetical protein
core	plcpsec113	cpfrc_00705	487	58.67	495	Rv1018c	*glmU*	NP_215534.1	bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase
core	plcpsec115	cpfrc_00945	64	63.33	64	Rv1642	*rpmI*	NP_216158.1	50S ribosomal protein L35
core	plcppse080	cpfrc_01015	452	57.08	470	Rv0392c	*ndhA*	NP_214906.1	membrane NADH dehydrogenase
core	plcppse080	cpfrc_01015	452	58.10	463	Rv1854c	*ndh*	NP_216370.1	NADH dehydrogenase
core	plcpsec041	cpfrc_01074	403	62.96	381	Rv1488		NP_216004.1	hypothetical protein
core	plcpsec119	cpfrc_01121	504	53.71	457	Rv1407	*fmu*	NP_215923.1	Fmu protein (SUN protein)
core	plcppse085	cpfrc_01126	417	55.58	418	Rv1391	*dfp*	NP_215907.1	bifunctional phosphopantothenoylcysteine decarboxylase/phosphopantothenate synthase
core	plcpsec138	cpfrc_01214	79	68.42	82	Rv2708c		NP_217224.1	hypothetical protein
core	plcpsec122	cpfrc_01267	683	57.76	558	Rv2752c		NP_217268.1	hypothetical protein
core	plcpsec124	cpfrc_01393	239	57.83	250	Rv2149c	*yfiH*	NP_216665.1	hypothetical protein
core	plcppse104	cpfrc_01424	412	50.38	429	Rv2195	*qcrA*	NP_216711.1	Rieske iron-sulfur protein QcrA
core	plcpsec128	cpfrc_01757	313	59.42	322	Rv3579c		NP_218096.1	tRNA/rRNA methyltransferase
core	plcppse131	cpfrc_01798	480	62.21	491	Rv2443	*dctA*	NP_216959.1	C4-dicarboxylate-transport transmembrane protein DctA
core	plcppse165	cpfrc_02038	721	52.00	678	Rv0050	*ponA1*	YP_177687.1	bifunctional penicillin-binding protein 1A/1B
core	plcppse174	cpfrc_02102	393	51.41	406	Rv3915	*cwlM*	YP_178027.1	hydrolase

Additional files

Additional file 1 – Predicted *C. pseudotuberculosis* pan secretome
List of the 150 genes for 750 *locus_tags* from the five *C. pseudotuberculosis* strains.

Additional file 2 – Predicted *C. pseudotuberculosis* pan surfaceome
List of the 227 genes for 1135 *locus_tags* from the five *C. pseudotuberculosis* strains.

Additional file 3 – *C. pseudotuberculosis* genomes
The five *C. pseudotuberculosis* genomes here checked, as EMBL files.

3.2.2 Uma abordagem combinada para análise comparativa do proteoma exportado da *C. pseudotuberculosis*

Proteínas exportadas representam componentes chave da interação hospedeiro-patógeno. Assim, buscou-se implementar uma abordagem combinada, que é descrita abaixo, para a caracterização do secretoma da bactéria patogênica *C. pseudotuberculosis*.

Um protocolo otimizado de três fases de particionamento (TPP) para obter proteínas exportadas de *C. pseudotuberculosis* (Paule e cols., 2003) e um método recentemente introduzido de aquisição de dados por espectrometria de massa (LC-MSE) (Geromanos e cols., 2009) foram utilizados para a identificação das proteínas. Além disso, o programa SurfG *plus* (Barinov e cols., 2009) foi utilizado para predição *in silico* de localização subcelular das proteínas identificadas. No total, 93 diferentes proteínas extracelulares de *C. pseudotuberculosis* foram identificadas por essa estratégia; 44 proteínas foram comumente identificadas nas linhagens 1002 e C231, isoladas dos hospedeiros caprino e ovino, respectivamente, compondo assim o exoproteoma central de *C. pseudotuberculosis*. Análises com o programa SurfG *plus* mostraram que mais de 75% (70/93) das proteínas identificadas possuíam motivos de exportação conservados. Além disso, foi encontrada evidência para a provável exportação não clássica da maioria das proteínas remanescentes.

Análises comparativas do exoproteoma destas duas linhagens de *C. pseudotuberculosis*, além da comparação com outros exoproteomas do gênero *Corynebacterium*, são úteis porque permitem uma melhor compreensão de como estas proteínas atuam na virulência destas bactérias.

BMC Microbiology

RESEARCH ARTICLE Open Access

A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*

Luis GC Pacheco[1,2,3], Susan E Slade[4], Núbia Seyffert[2], Anderson R Santos[2], Thiago LP Castro[2], Wanderson M Silva[2], Agenor V Santos[1], Simone G Santos[5], Luiz M Farias[5], Maria AR Carvalho[5], Adriano MC Pimenta[1], Roberto Meyer[3], Artur Silva[6], James H Scrivens[4], Sérgio C Oliveira[1], Anderson Miyoshi[2], Christopher G Dowson[4], Vasco Azevedo[2*]

Abstract

Background: Bacterial exported proteins represent key components of the host-pathogen interplay. Hence, we sought to implement a combined approach for characterizing the entire exoproteome of the pathogenic bacterium *Corynebacterium pseudotuberculosis*, the etiological agent of caseous lymphadenitis (CLA) in sheep and goats.

Results: An optimized protocol of three-phase partitioning (TPP) was used to obtain the *C. pseudotuberculosis* exoproteins, and a newly introduced method of data-independent MS acquisition (LC-MS^E) was employed for protein identification and label-free quantification. Additionally, the recently developed tool SurfG+ was used for *in silico* prediction of sub-cellular localization of the identified proteins. In total, 93 different extracellular proteins of *C. pseudotuberculosis* were identified with high confidence by this strategy; 44 proteins were commonly identified in two different strains, isolated from distinct hosts, then composing a core *C. pseudotuberculosis* exoproteome. Analysis with the SurfG+ tool showed that more than 75% (70/93) of the identified proteins could be predicted as containing signals for active exportation. Moreover, evidence could be found for probable non-classical export of most of the remaining proteins.

Conclusions: Comparative analyses of the exoproteomes of two *C. pseudotuberculosis* strains, in addition to comparison with other experimentally determined corynebacterial exoproteomes, were helpful to gain novel insights into the contribution of the exported proteins in the virulence of this bacterium. The results presented here compose the most comprehensive coverage of the exoproteome of a corynebacterial species so far.

Background

Corynebacterium pseudotuberculosis is a facultative intracellular pathogen that belongs to the so-called CMN (*Corynebacterium-Mycobacterium-Nocardia*) group, a distinct subgroup of the *Actinobacteria* that also includes other highly important bacterial pathogens, such as *Corynebacterium diphtheriae* and *Mycobacterium tuberculosis*. The most distinctive feature of these Gram-positive bacteria is the unique composition of the cell envelope, characterized by the presence of long chain fatty acids, known as mycolic acids, on the surface of the cell [1,2].

The main recognizable disease caused by *C. pseudotuberculosis* is caseous lymphadenitis (CLA) in sheep and goats, though this bacterium can also infect several other hosts, including humans [1,3]. Typical manifestations of CLA in small ruminants include formation of abscesses in superficial and internal lymph nodes, and in visceral organs [3]. Despite the important economic losses caused by this disease to sheep and goat husbandry worldwide, no effective treatment exists, and the efficacy of the currently available vaccines and diagnostic methods is still controversial [4].

The search for *C. pseudotuberculosis* molecular determinants that contribute to CLA pathogenesis lead to the

* Correspondence: vasco@icb.ufmg.br
[2]Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil
Full list of author information is available at the end of the article

BioMed Central

© 2011 Pacheco et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

recognition of two exported proteins as the major virulence-associated factors of this bacterium known to date: a secreted phospholipase D (PLD) [5]; and an ABC-type transporter component of an iron uptake system (FagB) [6]. In fact, one might expect that the majority of the virulence determinants of *C. pseudotuberculosis* would be present in the exoproteome, *i.e.* the entire set of bacterial proteins found in the extracellular milieu [7]. This is because exported proteins participate in essential steps of the host-pathogen interplay, including: (i) adhesion to host cells; (ii) invasion; (iii) damage to host tissues; (iv) resistance to environmental stresses during infection; and (iv) subversion of the host's immune response mechanisms [8-10].

In two previous attempts to characterize the *C. pseudotuberculosis* exoproteome, our group optimized a protocol of salting out of proteins using sulfate and butanol, known as three-phase partitioning (TPP), for isolation of the extracellular proteins of this bacterium [11], and generated a library of *C. pseudotuberculosis* mutant strains possessing transposon insertions in genes coding for probable exported proteins [12]. In the former study, we were able to determine the optimal conditions for obtaining the best recovery of immunoreactive extracellular proteins of *C. pseudotuberculosis* [11]. The second study in turn, enabled us to identify various previously uncharacterized *C. pseudotuberculosis* exported proteins, being that at least two of them are apparently involved in virulence [12]. Now, the very recent conclusion of the *C. pseudotuberculosis* Genome Project by our group, associated to the current availability of high-throughput proteomic technologies, permitted us to perform a much more comprehensive analysis of this bacterium's exoproteome.

In this study, we sought to implement a combined approach for comparative exoproteome analysis of different *C. pseudotuberculosis* strains. The strategy included: (i) the previously optimized TPP protocol for isolation of the extracellular proteins [11]; (ii) a newly introduced method of data-independent LC-MS acquisition (LC-MS^E) for protein identification and quantification [13,14]; and (iii) the recently developed tool SurfG+ for *in silico* prediction of protein sub-cellular localization in Gram-positive bacteria [15]. We believe that the experimental approach used is very suitable for profiling bacterial exoproteomes, as it shown to be easily applicable to different strains with very good reproducibility. This is an advantage over what is commonly observed for proteomic approaches based on two-dimensional (2D) gel electrophoresis, where there is more variability, but is apparently the method of choice for most of the bacterial exoproteome studies published recently [16-20]. Furthermore, the LC-MS^E method provides high subproteome coverage, due to enhanced sensitivity, and allows for label-free analysis of differentially expressed proteins [14]; this latter possibility enables the detection of variations in the exoproteomes of different strains that could be missed by simply profiling the exoproteins, and meets the growing interest in performing physiological proteomic studies of bacteria [21,22].

We were able to identify 93 different *C. pseudotuberculosis* extracellular proteins with high confidence by analyzing the exoproteomes of two strains isolated from different hosts that presented distinct virulence phenotypes under laboratory conditions [23,24]. Most of the identified proteins were predicted *in silico* to have an extracytoplasmic localization. To the best of our knowledge, these results compose the largest inventory of experimentally confirmed exoproteins of a single corynebacterial species to date. Importantly, the comparative exoproteome analyses permitted us to speculate on the probable contributions of different *C. pseudotuberculosis* extracellular proteins to the virulence of this bacterium.

Results and Discussion

Exoproteome analysis of *Corynebacterium pseudotuberculosis*

The extracellular proteins of two *C. pseudotuberculosis* strains, one isolated from a goat (strain 1002) the other from a sheep (strain C231), cultivated in a chemically-defined medium, were extracted/concentrated by the TPP technique. The trypsinized protein samples were then submitted to LC-MS^E analysis.

Seventy soluble extracellular proteins of the 1002 strain could be confidentially identified by this methodology, whereas the number of proteins identified in the exoproteome of the C231 strain was sixty-seven. Altogether, 93 different *C. pseudotuberculosis* exoproteins were identified in this study (Figure 1). These findings agree with the results of previous experiments by our group, in which we have used a 2D-PAGE based strategy for a preliminary appraisal of the *C. pseudotuberculosis* exoproteome (additional file 1). Eighty protein spots, mostly concentrated in the pI range between 3.0 and 6.0, could be reproducibly visible in the 2D gels generated from TPP-extracted extracellular proteins of the 1002 strain (additional file 1). The fact that we have found 70 proteins in the exoproteome of this strain with high confidence when using the LC-MS^E method (Figure 1) indicates that this novel methodology allowed us to identify virtually the complete set of extracellular proteins that are commonly observed in the gel based methodologies (additional file 1). Moreover, the expected existence of protein isoforms among the eighty protein spots observed in the 2D gels, and the identification by LC-MS^E of many proteins out of the pI range 3.0-6.0, suggests that the latter methodology is much more suitable for obtaining a comprehensive coverage of

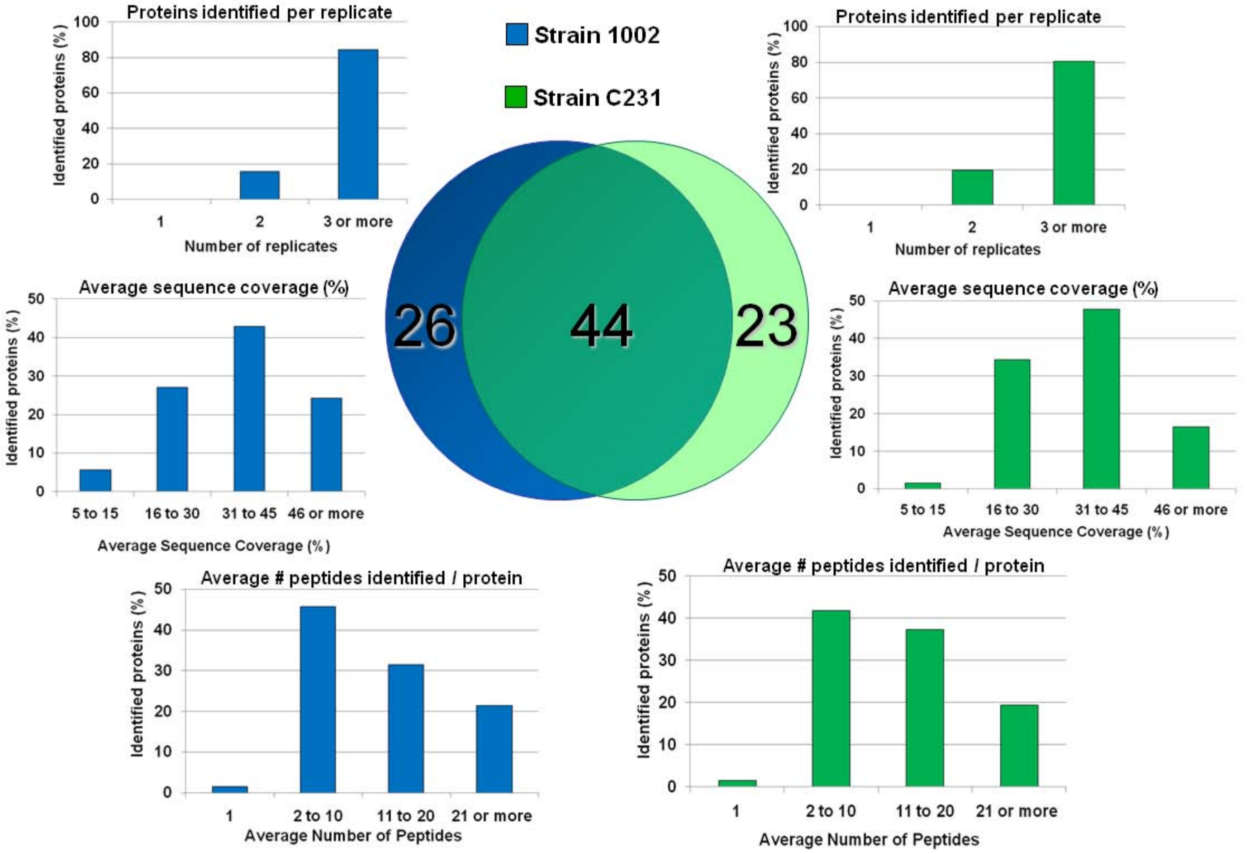


Figure 1 Analysis of the extracellular proteins of two different *C. pseudotuberculosis* strains allowed for identification of the core and variant exoproteomes. TPP-extracted extracellular proteins of the strains 1002 and C231 of *C. pseudotuberculosis* were submitted to LC-MS^E analysis. The Venn-diagram shows the numbers of commonly identified and variant exoproteins between the strains. The number of replicates in which a given protein was observed, the average peptides identified per protein, and the average sequence coverage of the proteins in each exoproteome studied, are shown as frequency distributions for comparison purposes.

the bacterial exoproteome. Noteworthy, is the use of LC-MS^E for exoproteome profiling which required (i) much less time and labor than the gel based proteomic strategy, and (ii) much less protein sample necessary for each experimental replicate, with only 0.5 μg per replicate used in the LC-MS^E compared to 150 μg for the 2D gels [refer to Patel *et al*. [25] for a comprehensive comparison on these proteomic strategies].

The performance of the combined methodology used in the present study (TPP/LC-MS^E) for mapping the *C. pseudotuberculosis* exoproteome was very similar for both strains analyzed, as can be seen by the average numbers of peptides observed per protein in the two proteomes (16.5 and 15.0) and by the average sequence coverage of the proteins identified (37.5% and 35.0%) (Figure 1). Consistent with this, the majority of the proteins detected in each extracellular proteome were shared by the goat and sheep isolates; this permitted us to define a core *C. pseudotuberculosis* exoproteome composed of 44 proteins out of the 93 different extracellular proteins identified. Additional files 2, 3 and 4 list all the proteins identified in the exoproteomes of the two *C. pseudotuberculosis* strains, along with molecular weights, isoelectric points, main orthologs, predicted sub-cellular localizations, number of peptides experimentally observed, and sequence coverage.

Searches of similarity against publicly available protein databases using the Blast-p tool [26] showed that ortholog proteins can be found in the pathogenic *Corynebacterium diphtheriae* for most of the identified *C. pseudotuberculosis* exoproteins (additional files 2, 3 and 4), as would be expected due to the close phylogenetic relationship of these species [27]. Nevertheless, no significant orthologs could be found for six proteins of the *C. pseudotuberculosis* exoproteome, even when using the position-specific iterated BLAST (PSI-BLAST) algorithm [28], namely the proteins [GenBank:ADL09626], [GenBank:ADL21925], [GenBank:ADL11253], [GenBank: ADL20222], [GenBank:ADL09871], and [GenBank: ADL21537] (additional files 2, 3 and 4). With the exception of [GenBank:ADL11253], all these proteins were predicted by different tools as being truly exported proteins. This means they are the only five exoproteins identified in this study which are probably unique for *C. pseudotuberculosis*.

Prediction of sub-cellular localization of the identified proteins

Most of the proteins identified in the exoproteomes of the two *C. pseudotuberculosis* strains were also predicted to have a probable extracytoplasmic localization after *in silico* analysis of the sequences of these proteins with different bioinformatics tools, thereby corroborating our *in vitro* findings (Figure 2, additional file 5). It is important to note here that we are considering the exoproteome as the entire set of proteins released by the bacteria into the extracellular milieu. That means we are looking to: (i) proteins possessing classical signals for active exportation by the different known mechanisms, which are directly secreted into the cell supernatant or that remain exposed in the bacterial cell surface and are eventually released in the growth medium [7]; and (ii) proteins exported by non-classical pathways, without recognizable signal peptides [29]. Besides, one might also expect to observe in the extracellular proteome a small number of proteins primarily known to have cytoplasmic localization; although some of these proteins are believed to be originated from cell lysis or leakage, like in the extreme situation reported by Mastronunzio *et al*. [19], a growing body of evidence suggests that moonlighting proteins (in this case, cytoplasmic proteins that assume diverse functions in the extracellular space) may be commonly found in the bacterial exoproteomes [29-32].

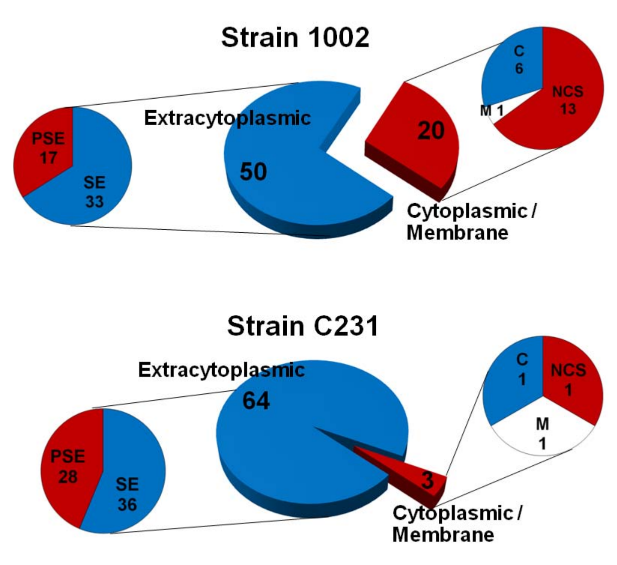


Figure 2 Most of the identified *C. pseudotuberculosis* exoproteins were predicted by the SurfG+ program as having an extracytoplasmic localization. The proteins identified in the exoproteomes of each *C. pseudotuberculosis* strain were analyzed by SurfG+ and attributed a probable final sub-cellular localization. Proteins classified as having a cytoplasmic localization were further analyzed with the SecretomeP tool for prediction of non-classical (leaderless) secretion. Besides, literature evidence for exportation by non-classical pathways was also used to re-classify the cytoplasmic proteins (see text for details). SE = secreted; PSE = potentially surface exposed; C = cytoplasmic; M = membrane; NCS = non-classically secreted.

By using the recently developed tool SurfG+ we were able to classify the identified *C. pseudotuberculosis* proteins into four different categories: (i) secreted, (ii) potentially surface exposed (PSE), (iii) membrane and (iv) cytoplasmic (Figure 2, additional files 2, 3 and 4). Basically, this software brings together the predictions of global protein localizations performed by a series of well-known algorithms, and innovates by allowing for an accurate prediction of PSE proteins [15]. This possibility of classification provides us with valuable information on the proteins identified, as bacterial surface exposed proteins are believed to play important roles in the host-pathogen interactions during infection and many of these proteins have been shown to be highly protective when used in vaccine preparations [33,34].

From a total of 93 different *C. pseudotuberculosis* proteins identified in this study, 75% (70) could be predicted as containing signals for active exportation (secretion or surface exposition) following SurfG+ analysis (Figure 2). Taken together, these proteins represent roughly 50% of all predicted secreted proteins in the recently sequenced genome of *C. pseudotuberculosis*, and around 15% of all predicted PSE proteins of this bacterium (A.R. Santos, pers. comm.).

The concordance of our *in vitro* identification of exoproteins with the *in silico* predictions of protein exportation is higher than what has normally been observed in recent exoproteome analyses of different bacteria [17-19,35,36]. For comparison, Hansmeier *et al*. [17] reported that exportation signals could be predicted in only 42 (50%) out of 85 different proteins identified in the extracellular and cell surface proteomes of *Corynebacterium diphtheriae*. The authors of this study are not the only to speculate on a probably important contribution of cross-contamination of the protein sample during preparation procedures for the observation of high numbers of proteins not predicted as having extracellular location in the bacterial exoproteomes [17,31]. We believe that the proportionally higher identification of proteins possessing exportation signals in the present study could have happened due to a series of different factors, including: (i) our methodology for isolation of the bacterial extracellular proteins might have extracted less "contaminant" cytoplasmic proteins than did other methodologies reported in previous studies; (ii) the combined strategy used by SurfG+ to predict protein sub-cellular localization might have performed better in the identification of exported proteins than happened with other strategies, sometimes based in only one prediction tool; (iii) the fact that we have included in the final exoproteome lists only proteins identified with high confidence, in at least two experimental replicates, reduced significantly the possibilities of false-positive identifications that might account for some of the unexpected proteins; and finally (iv) the lower proportion of proteins primarily regarded as cytoplasmic might be actually a typical characteristic of the *C. pseudotuberculosis* exoproteome.

Non-classically secreted proteins

Intriguingly, a much higher proportion (29.0%) of the exoproteome of the 1002 strain of *C. pseudotuberculosis* was composed by proteins predicted by SurfG+ as not having an extracytoplasmic location, when compared to only 4.5% in the exoproteome of the strain C231 (Figure 2). The possibility of these proteins being non-classically secreted has been evaluated using the SecretomeP algorithm [29]. We have also reviewed the literature for evidence of other bacterial exoproteomes that could support the extracellular localization found for these proteins in our study.

High SecP scores (above 0.5) could be predicted for 5 of the 19 proteins in the exoproteome of the 1002 strain considered by SurfG+ as having a cytoplasmic location (additional files 2 and 3); this could be an indicative that they are actually being secreted by non-classical mechanisms [29]. Nonetheless, 2 of these 5 proteins ([GenBank:ADL09626] and [GenBank:ADL20555]) were also detected in the exoproteome of the C231 strain, in which they were predicted by SurfG+ as possessing an extracytoplasmic location (additional file 2). A comparative analysis of the sequences encoding these proteins in the genomes of the two *C. pseudotuberculosis* strains showed that the disparate results were generated due to the existence of nonsense mutations in the genome sequence of the 1002 strain, which impaired the identification of signal peptides for the two proteins at the time of SurfG+ analysis (data not shown). We believe that it is unlikely that these differences represent true polymorphisms, as the proteins were identified in the extracellular proteome, indicating the real existence of exportation signals. This indeed demonstrates the obvious vulnerability of the prediction tools to the proper annotation of the bacterial genomes. On the other hand, the assignment of high SecP scores to these two proteins, even though they are not believed to be secreted by non-classical mechanisms, would be totally expected, as the SecretomeP is a predictor based on a neural network trained to identify general features of extracellular proteins; this means the prediction tool will attribute SecP scores higher than 0.5 to most of the secreted proteins, regardless the route of export [29].

We have found reports in the literature that strongly support the extracellular localization observed for 8 of the 14 remaining proteins considered as non-secretory by SurfG+ and SecretomeP in the exoproteome of the 1002 strain, and without any detectable signal peptide (additional files 2 and 3, Figure 2). Among these proteins there are the elongation factors Tu and Ts [16,33, 35,37-39]; the glycolytic enzymes triosephosphate isomerase, phosphoglycerate kinase and phosphoglycerate mutase [16-20,37-40]; the chaperonin GroES [16-18, 20,39]; a putative peptidyl prolyl cis trans isomerase [17,18,35,37,41]; and a hydroperoxide reductase enzyme [17,35,39].

Proteins primarily regarded as cytoplasmic have consistently been identified in the exoproteomes of different bacterial species, and moonlighting roles in the extracellular environment have already been demonstrated for some of them [31,32], including evasion of host's immune system [42], adhesion to host cells [43,44], folding of extracytoplasmic proteins [41,45], and interaction between microorganisms [40,46]. Noteworthy, specific evidences for active secretion of such cytoplasmic proteins have been demonstrated for only a few examples to date, and demonstration of an extracellular function is still missing for many of these proteins [30,31].

The variant exoproteome may account for differential virulence of the two *C. pseudotuberculosis* strains

A considerable number (49/93) of the extracellular proteins identified in this work was observed in only one of the two strains studied, then composing a variant experimental *C. pseudotuberculosis* exoproteome (additional files 3 and 4). Highly variant exoproteomes have also been reported recently for other Gram+ bacterial pathogens [20,36,39,47-49], and such a variation may be considered an important factor leading to the observable phenotypic dissimilarities and ultimately to differential virulence of the various strains [50,51]. Hecker *et al.* [36] reported on how the composition of the exoproteome can vary extremely within a single species, *Staphylococcus aureus*, being that only 7 out of 63 identified extracellular proteins were found in all the twenty-five clinical isolates studied.

One of the most intriguing results in the present study was the detection of the phospholipase D (PLD) protein only in the extracellular proteome of the strain C231 (additional file 4). As the regulation of PLD expression was demonstrated to be complex and highly affected by multiple environmental factors [52], we sought to detect this protein in the culture supernatant of the

C. pseudotuberculosis 1002 strain grown in a rich medium (brain-heart infusion broth) instead of only chemically-defined medium (CDM), but these attempts were also unfruitful (data not shown). Besides, we were not able to detect secretion of PLD following total exoproteome analysis of the 1002 strain grown under specific stress generating conditions (Pacheco *et al.*, unpublished). The results strongly indicate that this protein is actually not being secreted by the 1002 strain in culture.

PLD is an exotoxin considered as the major virulence factor of *C. pseudotuberculosis* [5,52]. It possesses sphingomyelinase activity that contributes to endothelial permeability and then to spreading of the bacteria within the host [5]. Mutation of the *pld* gene in *C. pseudotuberculosis* rendered strains no longer capable of causing caseous lymphadenitis (CLA) in sheep and goats; the potential of these strains to be used as live attenuated vaccines was already evaluated [53-55]. Similarly, the strain 1002 of *C. pseudotuberculosis* was already tested as a possible live attenuated vaccine against CLA due to its natural low virulent status, and administration of this bacterium to goats did not cause lesions formation [23,56]. The molecular mechanisms leading to the low virulence of the 1002 strain however remain undetermined so far. We believe that non-secretion of PLD might be one of the main factors responsible for the lowered virulence of the strain. Importantly, we currently cannot affirm that the 1002 strain does not produce this protein while infecting a mammalian host. Besides, this strain still retains the capability of causing localized abscesses and disease in susceptible mice (Pacheco *et al.*, unpublished results).

Other proteins believed to be associated with the virulence of *C. pseudotuberculosis* were also identified exclusively in the exoproteome of the C231 strain, namely FagD and Cp40 (Table 1). The former protein is a component of an iron uptake system, whose coding sequences are clustered immediately downstream of the *pld* gene in the *C. pseudotuberculosis* genome [6]. The latter protein is a secreted serine protease shown to be protective against CLA when used to vaccinate sheep [57].

Strikingly, one variant protein of the *C. pseudotuberculosis* exoproteome, a conserved hypothetical exported protein with a cutinase domain [GenBank:ADL10384], has its coding sequence present in the genome of the C231 strain but absent from the genome of the 1002 strain (additional file 6). The genomic structure of the gene's surroundings is indicative of a region prone to recombination events, such as horizontal gene transfer [58]. In fact, it seems that gene gain and loss are frequent events leading to variations observed in the bacterial exoproteomes [39,59].

Variation of the core exoproteome: differential expression analysis of the common proteins by LC-MS^E

In addition to identifying qualitative variations in the exoproteomes of the two *C. pseudotuberculosis* strains, we were also able to detect relative differences in expression of the proteins common to the two proteomes through label-free protein quantification by the LC-MS^E method. Relative protein quantification by this method can be obtained with basis on the accurate precursor ion mass and electrospray intensity data, acquired during the low energy scan step of the alternating scan mode of MS acquisition [14]. Importantly, this quantitative attribute of the technique opens up new possibilities of utilization, as grows the interest on the so-called physiological proteomics [21].

Thirty-four out of 44 proteins commonly identified in the exoproteomes of the strains 1002 and C231 of *C. pseudotuberculosis* were considered by the PLGS quantification algorithm as having significantly variable expression (score > 250; 95% CI) (Figure 3, additional files 2 and 7). If we further filter these results for the proteins presenting differential expression higher than 2-fold between the strains, we end up with only four proteins up-regulated in the 1002 strain and sixteen in the C231 strain (Figure 3).

Among the group of proteins not presenting considerable variations in expression between the two *C. pseudotuberculosis* strains, proteins probably participating in basic bacterial physiological processes could be easily identified, as would be expected, including cell shape maintenance and cell division (penicillin binding protein, transglycosylases, peptidases, PGRP amidase) [60]; and iron uptake and utilization (HmuT) [61] (Figure 3, additional file 2). In this sense, one might also speculate that the hypothetical proteins identified as non variant in the two strains may have functions associated to the general physiology of *C. pseudotuberculosis*, when grown in minimal medium.

The most up-regulated proteins were observed in the extracellular proteome of the C231 strain, including two cell envelope-associated proteins [62], namely the major secreted (mycoloyltransferase) protein PS1 (10-fold up-regulated), and the S-layer protein A (8-fold up-regulation) (Figure 3). This may be indicative of differences on cell envelope-related activities in the two *C. pseudotuberculosis* strains, such as nutrient acquisition, protein export, adherence and interaction with the host [63]. Dumas *et al.* [49] compared the exoproteomes of *Listeria monocytogenes* strains of different virulence groups, and found that altered expression (up- or down-regulation) of a protein related to the bacterial cell wall could be a marker of specific virulence phenotypes. Additionally, surface associated proteins have been shown to undergo phase and antigenic variation in some bacterial

Table 1 Formerly and newly identified[‡] exported proteins that may be associated with the virulence phenotype of *Corynebacterium pseudotuberculosis* strains

Protein Description[a]	GenBank Accession	Identified in the exoproteome of the strain[b]:		Orhologs found in other Corynebacteria[c]:		References
		1002	C231	Pathogenic	Non-pathogenic	
Phospholipase D (PLD)	ADL09524.1	No	Yes	Yes	No	[54]
Iron siderophore binding protein (FagD)	ADL09528.1	No	Yes	Yes	Yes	[6]
Serine proteinase precursor (CP40)	ADL11339.1	No	Yes	No	No	[57]
Putative iron transport system binding (secreted) protein	ADL10460.1	No	Yes	Yes	No	[12]
Glycerophosphoryl diester phosphodiesterase	ADL11410.1	No	Yes	Yes	No	This work. [72]
Putative surface-anchored membrane protein	ADL20074.1	Yes	Yes	Yes	No	This work.
Putative hydrolase (lysozyme-like)	ADL20788.1	Yes	Yes	Yes	No	This work.
Putative secreted protein	ADL21714.1	Yes	Yes	Yes	No	This work.
Putative sugar-binding secreted protein	ADL09872.1	No	Yes	Yes	No	This work.

[‡] The inclusion criteria followed three main requisites: (i) experimental detection of the proteins in the exoproteomes of the pathogenic *C. diphtheriae* and *C. jeikeium*; (ii) non-detection of the proteins in the exoproteomes of the non-pathogenic *C. glutamicum* and *C. efficiens*; and (iii) *in silico* detection of ortholog proteins in pathogenic, but not in non-pathogenic, corynebacteria through search of similarity against public protein repositories.

[a] This protein list is not meant to be all-inclusive. Rather, it wants to give an overview of the exported proteins identified in this study for which it was possible to speculate on a probable involvement in *C. pseudotuberculosis* virulence after comparative proteomic analyses.

[b] Proteins identified in this study by TPP/LC-MS^E.

[c] Searches of similarity against publicly available protein databases using Blast-p.

pathogens, and ultimately affect the infectivity potential of different strains [50].

Comparative analyses of corynebacterial exoproteomes

Recent studies attempted to characterize the extracellular proteomes of other pathogenic (*C. diphtheriae* and *C. jeikeium*) and non-pathogenic (*C. glutamicum* and *C. efficiens*) corynebacterial species [17,37,64,65]. All these studies used 2D-PAGE to resolve the extracellular proteins of the different corynebacteria, and PMF by MALDI-TOF-MS was the method of choice in most of them for protein identification [17,37,64,65]. Figure 4 shows the numbers of proteins identified in the exoproteomes of all strains studied, in comparison to the numbers obtained in the present study for *C. pseudotuberculosis*. Despite one study with the strain R of *C. glutamicum*, which reports identification of only two secreted proteins [65], all the corynebacterial strains had somehow similar numbers of extracellular proteins identified, ranging from forty-seven in *C. jeikeium* K411 to seventy-four in *C. diphtheriae* C7s(-)$^{tox-}$. Importantly, the fact that we have identified in this study 93 different exoproteins of *C. pseudotuberculosis*, through the analysis of two different strains, means that our dataset represents the most comprehensive exoproteome analysis of a corynebacterial species so far.

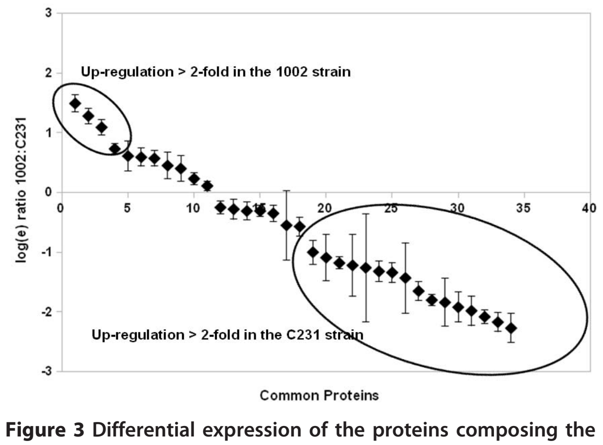


Figure 3 Differential expression of the proteins composing the core *C. pseudotuberculosis* exoproteome, evaluated by label-free relative quantification using LC-MS^E. Results are shown as natural log scale of the relative quantifications (1002:C231) for each protein. Only proteins that were given a variation score higher than 250 by PLGS quantification algorithm are presented. Proteins regulated more than 2-fold in each strain are indicated. Protein identification numbers correspond to additional files 2 and 7: Tables S1 and S4.

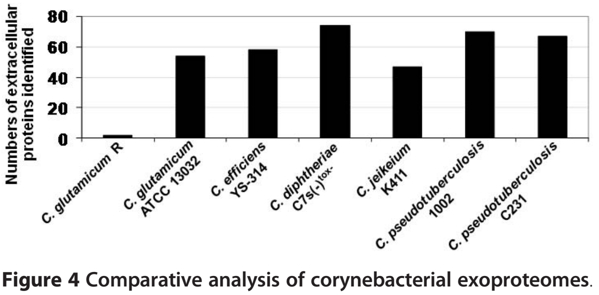


Figure 4 Comparative analysis of corynebacterial exoproteomes. Numbers of extracellular proteins identified in previous corynebacterial exoproteome analyses [17,37,69,70] in comparison to those identified in this study with the two strains of *C. pseudotuberculosis*.

Regardless the different methodologies employed to characterize the exoproteomes of the various corynebacteria, we sought to identify extracellular proteins commonly identified in most of the studies, taking the catalogue of *C. pseudotuberculosis* exoproteins generated in this work as the comparison dataset. Besides corroborating our findings, the objective here was to identify extracellular proteins that could be associated exclusively to pathogenic corynebacterial species.

In total, 34 proteins identified in the exoproteome of the strain 1002 of *C. pseudotuberculosis* were found to be present in the experimentally determined extracellular proteomes of other corynebacteria, whereas the number of common corynebacterial exoproteins in the C231 strain was 32 (Figure 5). Only 6 proteins were consistently identified in all the corynebacterial exoproteomes, including pathogenic and non-pathogenic species: (i) S-layer protein A [62]; (ii) resuscitation-promoting factor RpfB [66]; (iii) cytochrome c oxidase subunit II [67]; (iv) a putative esterase; (v) a NLP/P60 family protein (putative cell wall-associated hydrolase) [68]; and (vi) a trehalose corynomycolyl transferase (Figure 5, additional file 8). Interestingly, three of these six proteins are predicted to be regulated by the same transcription factor [GenBank:ADL09702], a member of the cAMP receptor protein (Crp) family of transcription regulators which are found controlling a diversity of physiological functions in various bacteria [69].

Twelve proteins of the exoproteome of the 1002 strain and fifteen of the C231 strain were also detected experimentally only in the exoproteomes of other pathogenic corynebacteria, namely *C. diphtheriae* and *C. jeikeium* (Figure 5). Altogether, this represents 19 different *C. pseudotuberculosis* proteins (additional file 8). A search of similarity using the sequences of these proteins against publicly available databases, believed to contain the predicted proteomes of all corynebacteria with completely sequenced genomes, showed that 6 of these

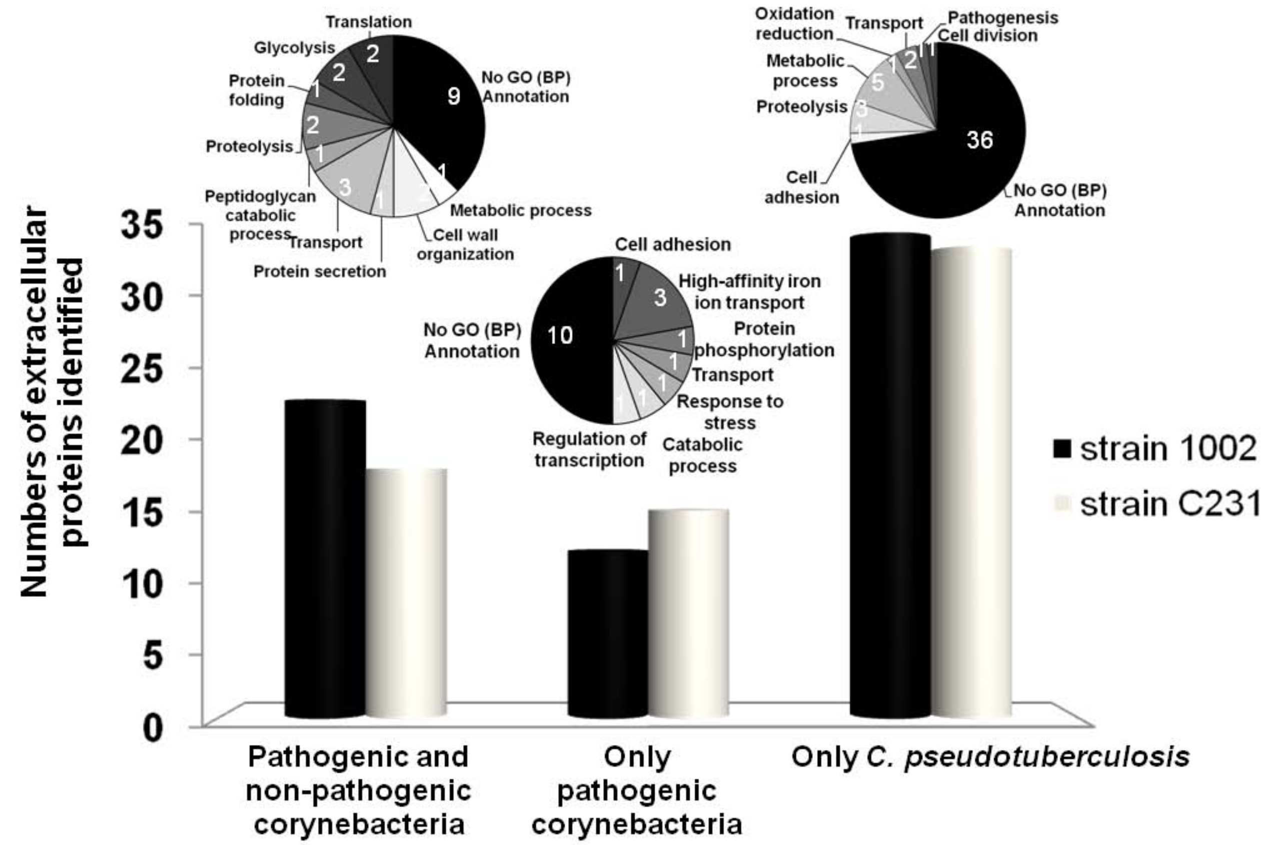


Figure 5 Distribution of orthologous proteins of the *C. pseudotuberculosis* experimental exoproteins throughout other experimentally confirmed corynebacterial exoproteomes. Pathogenic species: *C. diphtheriae* C7s(-)$^{tox-}$ and *C. jeikeium* K411 [17,69]; non-pathogenic species: *C. glutamicum* ATCC13032 and *C. efficiens* YS-314 [37,70]. Pie charts show Gene Ontology (GO) functional annotations for the 93 different *C. pseudotuberculosis* exoproteins identified (24 commonly identified in pathogenic and non-pathogenic corynebacteria; 19 commonly identified only in pathogenic corynebacteria; and 50 only identified in *C. pseudotuberculosis*). Annotations were obtained following analyses with the Blast2GO tool [84], used through the web application available at http://www.blast2go.org/start_blast2go.

19 proteins are apparently absent from non-pathogenic corynebacterial species (Table 1). Moreover, 5 of these proteins are predicted to be part of regulatory networks already shown to be involved in virulence functions, including those regulated by the diphtheria toxin repressor (DtxR)-like protein [70] and the cAMP-binding transcription regulator GlxR [71].

Two proteins presented orthologs highly distributed in various bacterial pathogens: (i) a putative iron transport system binding (secreted) protein [GenBank:ADL10460]; and (ii) a putative glycerophosphoryl diester phosphodiesterase [GenBank:ADL11410]. Interestingly, an ortholog of this latter protein was included recently in a list of seventeen proteins found to be very common in pathogenic bacteria and absent or very uncommon in non-pathogens, representing then probable virulence-associated factors [72]. In fact, reports in the literature can be found that associate orthologs of the two aforementioned proteins with virulence phenotypes [73,74]. Noteworthy, both proteins were detected in this study only in the exoproteome of the C231 strain of *C. pseudotuberculosis*, the more virulent one.

Conclusions

There seems to be a growing interest in profiling the exoproteomes of bacterial pathogens, due to the distinguished roles played by exported proteins on host-pathogen interactions [10]. Classical proteomic profiling strategies, normally involving two-dimensional (2D) gel electrophoresis, have been extensively used for this purpose [16-20]. Nevertheless, the introduction of more high-throughput proteomic technologies brings new perspectives to the study of bacterial exoproteomes, as it makes it easier to analyze multiple phenotypically distinct strains, yielding better subproteome coverage with fewer concerns regarding technical sensitivity and reproducibility [75]. Besides, the currently available methods for label-free quantification of proteins [76] allow us to compare the "dynamic behavior" of the exoproteome across different bacterial strains, and this in turn will help us to better identify alterations of the exoproteome that may contribute to the various virulence phenotypes.

By using a high-throughput proteomic strategy, based on a recently introduced method of LC-MS acquisition (LC-MS^E) [14], we were able to perform a very comprehensive analysis of the exoproteome of an important veterinary pathogen, *Corynebacterium pseudotuberculosis*. Comparative exoproteome analysis of two strains presenting different virulence status allowed us to detect considerable variations of the core *C. pseudotuberculosis* extracellular proteome, and thereby the number of exoproteins identified increased significantly. Most importantly, it was helpful to gain new insights into the probable participation of *C. pseudotuberculosis* exported proteins, other than the well-known PLD and FagB, in the virulence of this bacterium. Several novel targets for future work on *C. pseudotuberculosis* molecular determinants of virulence can be identified from the catalogue of exoproteins generated in this study. Interestingly, around 30% of the proteins identified were predicted by the SurfG+ software [15] as being probably surface exposed in *C. pseudotuberculosis*. Such proteins may represent promising new candidates for composing a CLA vaccine more effective than the ones currently available [4], as has been demonstrated for a series of other bacterial pathogens [33,34]. Therefore, it will be critical to further study the role of this protein set in virulence and vaccine design.

Methods

Bacterial strains and culture conditions

The strains 1002 and C231 of *Corynebacterium pseudotuberculosis* were used in this study. Strain 1002 was isolated from an infected goat in Brazil and has been shown to be naturally low virulent [23,56]; strain C231 was isolated from an infected sheep in Australia, and it showed a more virulent phenotype [24]. Species confirmation was performed by biochemical and molecular methods for both strains, as described [77]. Complete genome sequences of the two strains were generated by Genome Networks in Brazil and Australia (RGMG/RPGP and CSIRO Livestock Industries), and made available for this study (unpublished results).

C. pseudotuberculosis strains were routinely maintained in Brain Heart Infusion broth (BHI: Oxoid, Hampshire, UK) or in BHI 1.5% bacteriological agar plates, at 37°C. For proteomic studies, strains were grown in a chemically defined medium (CDM) previously optimized for *C. pseudotuberculosis* cultivation [78]. The composition of the CDM was as follows: autoclaved 0.067 M phosphate buffer [$Na_2HPO_4 \cdot 7H_2O$ (12.93 g/L), KH_2PO_4 (2.55 g/L), NH_4Cl (1 g/L), $MgSO_4 \cdot 7H_2O$ (0.20 g/L), $CaCl_2$ (0.02 g/L), and 0.05% (v/v) Tween 80]; 4% (v/v) MEM Vitamins Solution 100X (Invitrogen); 1% (v/v) MEM Amino Acids Solution 50X (Invitrogen); 1% (v/v) MEM Non Essential Amino Acids Solution 100X (Invitrogen); and 1.2% (w/v) filter-sterilized glucose.

Three-phase partitioning

Extraction/concentration of the soluble supernatant proteins of *C. pseudotuberculosis* followed the TPP protocol previously optimized by our group [11], with minor modifications. Briefly, overnight cultures (*ca*. 24 hours) of the different *C. pseudotuberculosis* strains were inoculated (1:100) separately into 500 mL of pre-warmed fresh CDM and incubated at 37°C, with agitation at 100 rpm, until reach the mid-exponential growth phase ($OD_{540\ nm}$ = 0.4; LabSystems iEMS Absorbance Plate

Reader). At this point, cultures were centrifuged at room temperature (RT) for 20 min, 4000 rpm, and 400 mL of each supernatant was transferred into new sterile flaks. Following addition of 20 μL Protease Inhibitor Cocktail P8465 (Sigma-Aldrich), supernatants were filtered through 0.22 μm filters; ammonium sulphate was added to the samples at 30% (w/v) and the pH of the mixtures were set to 4.0. Then, *n*-butanol was added to each sample at an equal volume; samples were vigorously vortexed and left to rest for 1 h at RT, until the mixtures separated into three phases. The interfacial precipitate was collected in 1.5 mL microtubes, and re-suspended in 1 mL Tris 20 mM + 10 μL protease inhibitor. Finally, samples were submitted to diafiltration and buffer exchange with NH_4HCO_3 (100 mM), using 5 kDa cut-off spin columns (Millipore).

In-solution tryptic digestion of TPP-extracted proteins

Protein samples were resuspended in 1 mL of 0.1% Rapigest (Waters Corporation, Milford, MA) and concentrated using a 5 kDa cut-off spin column. The solution was heated at 80°C for 15 minutes, reduced with dithiothreitol, alkylated with iodoacetamide and digested with 1:50 (w/w) sequencing grade trypsin for 16 hours. RapiGest was hydrolysed by the addition of 2 μL of 13 M trifluoroacetic acid, filtered using a 0.22 μm spin column and each sample was typically diluted to 1 μg/μL prior to a 1:1 dilution with a 100 fmol/μL glycogen phosphorylase B standard tryptic digest to give a final protein concentration of 500 ng/μL per sample and 50 fmol/μL phosphorylase B.

LC-MS configurations for label-free analysis (LC-MS^E)

Nanoscale LC separations of tryptic peptides for qualitative and quantitative multiplexed LC-MS analysis were performed with a nanoACQUITY system (Waters Corporation) using a Symmetry C_{18} trapping column (180 μm × 20 mm 5 μm) and a BEH C_{18} analytical column (75 μm × 250 mm 1.7 μm). The composition of solvent A was 0.1% formic acid in water, and solvent B (0.1% formic acid in acetonitrile). Each sample (total digested protein 0.5 μg) was applied to the trapping column and flushed with 0.1% solvent B for 2 minutes at a flow rate of 15 μL/min. Sample elution was performed at a flow rate of 250 nL/min by increasing the organic solvent concentration from 3 to 40% B over 90 min. Three technical replicate injections of the TPP-extracted 1002 sample and four technical replicates of the TPP-extracted C231 sample were used for subsequent data analysis in this study. These were from two biological cultures of each *C. pseudotuberculosis* stain.

The precursor ion masses and associated fragment ion spectra of the tryptic peptides were mass measured with a Q-ToF Ultima Global or Synapt HDMS mass spectrometer (Waters Corporation) directly coupled to the chromatographic system. The time-of-flight analyzers of both mass spectrometers were externally calibrated using the MS/MS spectrum from [Glu^1]-Fibrinopeptide B (human - Sigma Aldrich, UK) obtained from the doubly charged peptide ion at *m/z* 785.8426. The monoisotopic mass of the doubly charged species in MS mode was also used for post-acquisition data correction. The latter was delivered at 500 fmol/μL to the mass spectrometer via a NanoLockSpray interface using the auxiliary pump of a nanoACQUITY system at a flow rate of 500 nL/min, sampled every 60 seconds.

Accurate mass data were collected in data independent mode of acquisition by alternating the energy applied to the collision cell/s between a low and elevated energy state (MS^E). The spectral acquisition scan rate was typically 0.9 s with a 0.1 s interscan delay. On the Synapt HDMS instrument in the low energy MS mode, data were collected at constant trap and transfer collision energies (CE) of 3 eV and 1 eV respectively. In elevated energy MS mode, the trap collision energy was ramped from 15 eV to 30 eV with the transfer collision energy at 10 eV. On the Ultima Global instrument a low energy of 6 eV was applied to the collision cell, increasing from 6 eV to 35 eV in elevated MS mode.

Data processing for label-free acquisitions (MS^E)

The LC-MS^E data were processed using ProteinLynx Global Server v2.4 (Waters Corporation, Milford, MA) (see additional file 9). In brief, lockmass-corrected spectra are centroided, deisotoped, and charge-state-reduced to produce a single accurately mass measured monoisotopic mass for each peptide and the associated fragment ion. The initial correlation of a precursor and a potential fragment ion is achieved by means of time alignment. The detection and correlation principles for data independent, alternate scanning LC-MS^E data have been described [14].

Database searches

All data were searched using PLGS v2.4 against a *Corynebacterium pseudotuberculosis* database (NCBI Genome Project ID: 40687 and 40875), released in November 2009, to which the glycogen phosphorylase B and trypsin sequences had been appended. The database was randomised within PLGS generating a new concatenated database consisting of the original sequences plus one additional sequence for each entry with identical composition but randomly scrambled residues. This database contained a total of 4314 entries. A fixed modification of carbamidomethyl-C was specified, and variable modifications included were acetyl N-terminus, deamidation N, deamidation Q and oxidation M. One missed trypsin cleavage site was permitted.

For the MS^E data, the time-based correlation applied in data processing is followed by a further correlation process during the database search that is based on the physicochemical properties of peptides when they undergo collision induced fragmentation. The precursor and fragment ion tolerances were determined automatically. The initial protein identification criteria used by the $Identity^E$ algorithm within PLGS for a single replicate data file, required the detection of at least three fragment ions per peptide, seven fragment ions and a minimum of one peptide per protein.

A process analogous to the Bayesian model described by Nesvizhskii *et al.* [79] was used by PLGS to assign probability values to scores of peptide and protein identifications. Two automated mechanisms determined peptide and protein threshold identification criteria providing a 95% identification confidence interval. A background search is conducted by the search algorithm creating a discriminating decoy identification distribution. The determined peptide cut-off score, typically a log value of 6.25 for the expected 95% identification probability is automatically applied to the results.

Further more stringent filtering was then applied to the database search results from each sample to improve the confidence in the protein observations and quantitative measurements. The results from each of the *individual* replicate analyses from each sample were combined and proteins were removed that were observed in only one of the replicates. Using this additional and rigorous filter the false discovery rate was further reduced to 0.2% for this study, with an average of 16.5 peptides/protein and 37.5% sequence coverage for the TPP-extracted 1002 sample and 15 peptides/protein with 35% sequence coverage for the respective C231 sample. Proteins were observed on average in 2.81 technical replicates in the 1002 sample where 3 replicate analyses were used and 3.52 for the C231 sample in which 4 replicates were included.

Protein quantification using label-free system (MS^E)

Relative quantitative analysis between samples was performed by comparing normalized peak area/intensity of each identified peptide [80]. For relative quantification, automatic normalization was applied to the data set within PLGS using the total peptide complement of each sample. The redundant, proteotypic quantitative measurements generated from the tryptic peptide identifications from each protein were used to determine an average, relative protein fold-change, with a confidence interval and a regulation probability. The confidently identified peptides to protein ratios were automatically weighted based on their identification probability. Binary comparisons were conducted to generate an average normalized intensity ratio for all matched proteins. The entire data set of differentially expressed proteins was further filtered by considering only the identified proteins that replicated in at least two technical replicates with a score > 250 and likelihood of regulation value greater than 0.95 for upregulation and lower than 0.05 for downregulation as determined by the PLGS quantification algorithm.

In silico predictions of protein sub-cellular localization

Prediction of sub-cellular localization was performed initially for the identified proteins by using the SurfG+ program v1.0, run locally in a Linux environment, as described [15] (see additional file 9). For prediction of potentially surface exposed (PSE) proteins, a cut-off value of 73 amino acids was calculated as the minimum distance from the *C. pseudotuberculosis* outermost membrane until the surface of the cell-wall, based on electron microscopy of this bacterium's cell envelope (data not shown).

The programs TatP v1.0 and SecretomeP v2.0 were used through the web applications available at http://www.cbs.dtu.dk/services/, for prediction of twin-arginine pathway-linked signal peptides and non-classical (leaderless) secretion, respectively [29,81].

Comparative analyses of multiple corynebacterial exoproteomes

A list of experimentally observed extracellular proteins of pathogenic (*C. diphtheriae* and *C. jeikeium*) and non-pathogenic (*C. glutamicum* and *C. efficiens*) corynebacteria was identified in previously published studies [17,37,64,65]. The amino acid sequences of these proteins were retrieved from public repositories of protein sequences to create a local database. This database was used in similarity searches with the Blast-p algorithm (E-value < 10^{-4}) [26], taking the group of proteins identified in the *C. pseudotuberculosis* exoproteome as the input sequences. Additionally, transitivity clustering [82] was used to identify proteins (i) commonly detected in the exoproteomes of pathogenic and non-pathogenic corynebacteria, and proteins detected in exoproteomes of (ii) only pathogenic corynebacteria or (iii) only *C. pseudotuberculosis*. A more detailed description on the transitivity clustering analysis can be found in the supplementary material (additional file 9). The amino acid sequences of the identified *C. pseudotuberculosis* exoproteins were also used in similarity searches against public databases, namely NCBI nr and Swissprot.

Transcriptional regulation of the identified exoproteins

The search for transcription factors that regulate expression of the identified corynebacterial exoproteins was performed through the CoryneRegNet database, as described previously [83].

Accession numbers

The sequences of all proteins identified in this work are accessible through GenBank and correspond to the *Corynebacterium pseudotuberculosis* Genome Projects deposited in NCBI (IDs: 40687 and 40875).

Additional material

Additional file 1: Figure S1. Comparison between the experimental (A) and virtual (B) 2-D gels of the exoproteome of the strain 1002 of *C. pseudotuberculosis*. (A) 2D-gel with 150 μg of TPP extracted extracellular proteins of the 1002 strain. Proteins were separated in the first dimension by isoelectric focusing using strips of 3.0-5.6 NL pI range (GE Healthcare). Visualization was by Colloidal Coomassie staining. (B) The virtual 2D-gel was generated with the theoretical pI and MW values of the proteins identified by LC-MS^E.

Additional file 2: Table S1. Proteins composing the core *C. pseudotuberculosis* exoproteome, identified by LC-MS^E.

Additional file 3: Table S2. Variant exoproteome of the strain 1002 of *Corynebacterium pseudotuberculosis*.

Additional file 4: Table S3. Variant exoproteome of the strain C231 of *Corynebacterium pseudotuberculosis*.

Additional file 5: Figure S2. Predictions of LPXTG motif-containing proteins, lipoproteins and Tat-pathway associated signal peptides in the exoproteomes of the strains 1002 and C231 of *C. pseudotuberculosis*.

Additional file 6: Figure S4. A conserved hypothetical exported protein present in the Genome of the strain C231 but absent from the strain 1002 of *C. pseudotuberculosis*. The two sequenced Genomes were aligned using the Artemis Comparison Tool (ACT). The arrows point to tRNA genes.

Additional file 7: Table S4. Relative expression analysis of the extracellular proteins common to the strains 1002 and C231 of *Corynebacterium pseudotuberculosis*.

Additional file 8: Figure S5. Distribution of orthologous proteins of the *C. pseudotuberculosis* experimental exoproteins throughout other experimentally confirmed exoproteomes of pathogenic corynebacteria, as determined through transitivity clustering analysis. The 19 *C. pseudotuberculosis* exoproteins only identified in the exoproteomes of other pathogenic corynebacteria are presented in the table. *Cp = C. pseudotuberculosis*; *Cd = C. diphtheriae*; *Cj = C. jeikeium*.

Additional file 9: Supplementary information on the bioinformatics tools used in this study.

List of abbreviations

CDM: chemically defined medium; CLA: caseous lymphadenitis; LC-MS: liquid chromatography - mass spectrometry; NCS: non-classically secreted; PLD: phospholipase D; PLGS: ProteinLynx Global Server; PMF: peptide mass fingerprinting; PSE: potentially surface exposed; RGMG: Minas Gerais Genome Network; RPGP: Genome and Proteome Network of the State of Pará; TPP: Three-Phase Partitioning.

Acknowledgements

We are thankful to the Minas Gerais Genome Network (RGMG) and to the Genome and Proteome Network of the State of Pará (RPGP). We thank Dr. Robert Moore (CSIRO Livestock Industries) for providing the C231 strain of *C. pseudotuberculosis*.

This work was supported by grants from the Funding Agencies CNPq (grant CNPq/MAPA/SDA) and FAPEMIG, in Brazil; and by The Medical Research Fund and Advantage West Midlands, in the UK.

Author details

[1]Department of Biochemistry and Immunology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. [2]Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. [3]Institute of Health Sciences, Universidade Federal da Bahia, Av. Reitor Miguel Calmon, Salvador, 40.110-902, Brazil. [4]School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom. [5]Department of Microbiology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, Belo Horizonte, 31.270-901, Brazil. [6]Genome and Proteome Network of the State of Pará, Universidade Federal do Pará, R. Augusto Corrêa, Belém, 66.075-110, Brazil.

Authors' contributions

LGCP, SES, LMF, MARC, AMCP, RM, AS, JHS, SCO, AM, CGD, and VA conceived the idea, participated in the design of the study, and critically read the manuscript. LGCP, SES, NS, TLPC, WMS, AGV, and SGS performed microbiological and/or proteomic experiments. LGCP, SES and ARS performed bioinformatical analyses. LGCP and SES wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 15 October 2010 Accepted: 17 January 2011
Published: 17 January 2011

References

1. Dorella FA, Pacheco LGC, Oliveira SC, Miyoshi A, Azevedo V: *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet Res* 2006, **37**:201-218.
2. Ventura M, Canchaya C, Tauch A, Chandra G, Fitzgerald GF, Chater KF, van Sinderen D: Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol Mol Biol Rev* 2007, **71**:495-548.
3. Baird GJ, Fontaine MC: *Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis. *J Comp Pathol* 2007, **137**:179-210.
4. Dorella FA, Pacheco LG, Seyffert N, Portela RW, Meyer R, Miyoshi A, Azevedo V: Antigens of *Corynebacterium pseudotuberculosis* and prospects for vaccine development. *Expert Rev Vaccines* 2009, **8**:205-213.
5. Hodgson AL, Bird P, Nisbet IT: Cloning, nucleotide sequence, and expression in *Escherichia coli* of the phospholipase D gene from *Corynebacterium pseudotuberculosis*. *J Bacteriol* 1990, **172**:1256-1261.
6. Billington SJ, Esmay PA, Songer JG, Jost BH: Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*. *FEMS Microbiol Lett* 2002, **208**:41-45.
7. Desvaux M, Hébraud M, Talon R, Henderson IR: Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol* 2009, **17**:139-145.
8. Bhavsar AP, Guttman JA, Finlay BB: Manipulation of host-cell pathways by bacterial pathogens. *Nature* 2007, **449**:827-834.
9. Stavrinides J, McCann HC, Guttman DS: Host-pathogen interplay and the evolution of bacterial effectors. *Cell Microbiol* 2008, **10**:285-292.
10. Sibbald MJJB, van Dij JML: Secretome Mapping in Gram-Positive Pathogens. In *Bacterial secreted protein: secretory mechanisms and role in pathogenesis* Edited by: Karl Wooldridge 2009, 193-225.
11. Paule BJA, Meyer R, Moura-Costa LF, Bahia RC, Carminati R, Regis LF, Vale VLC, Freire SM, Nascimento I, Schaer R, Azevedo V: Three-phase partitioning as an efficient method for extraction/concentration of immunoreactive excreted-secreted proteins of *Corynebacterium pseudotuberculosis*. *Protein Expr Purif* 2004, **34**:311-316.
12. Dorella FA, Estevam EM, Pacheco LGC, Guimarães CT, Lana UGP, Gomes EA, Barsante MM, Oliveira SC, Meyer R, Miyoshi A, Azevedo V: In vivo insertional mutagenesis in *Corynebacterium pseudotuberculosis*: an efficient means to identify DNA sequences encoding exported proteins. *Appl Environ Microbiol* 2006, **72**:7368-7372.

13. Silva JC, Gorenstein MV, Li G, Vissers JPC, Geromanos SJ: **Absolute quantification of proteins by LCMSE a virtue of parallel MS acquisition.** *Mol Cell Proteomics* 2006, **5**:144-156.
14. Geromanos SJ, Vissers JPC, Silva JC, Dorschel CA, Li G, Gorenstein MV, Bateman RH, Langridge JI: **The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS.** *Proteomics* 2009, **9**:1683-1695.
15. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M: **Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria.** *Proteomics* 2009, **9**:61-73.
16. Trost M, Wehmhöner D, Kärst U, Dieterich G, Wehland J, Jänsch L: **Comparative proteome analysis of secretory proteins from pathogenic and nonpathogenic *Listeria* species.** *Proteomics* 2005, **5**:1544-1557.
17. Hansmeier N, Chao T, Kalinowski J, Pühler A, Tauch A: **Mapping and comprehensive analysis of the extracellular and cell surface proteome of the human pathogen *Corynebacterium diphtheriae*.** *Proteomics* 2006, **6**:2465-2476.
18. Målen H, Berven FS, Fladmark KE, Wiker HG: **Comprehensive analysis of exported proteins from *Mycobacterium tuberculosis* H37Rv.** *Proteomics* 2007, **7**:1702-1718.
19. Mastronunzio JE, Huang Y, Benson DR: **Diminished exoproteome of *Frankia* spp. in culture and symbiosis.** *Appl Environ Microbiol* 2009, **75**:6721-6728.
20. Dumas E, Desvaux M, Chambon C, Hébraud M: **Insight into the core and variant exoproteomes of *Listeria monocytogenes* species by comparative subproteomic analysis.** *Proteomics* 2009, **9**:3136-3155.
21. Hecker M, Reder A, Fuchs S, Pagels M, Engelmann S: **Physiological proteomics and stress/starvation responses in *Bacillus subtilis* and *Staphylococcus aureus*.** *Res Microbiol* 2009, **160**:245-258.
22. Becher D, Hempel K, Sievers S, Zühlke D, Pané-Farré J, Otto A, Fuchs S, Albrecht D, Bernhardt J, Engelmann S, Völker U, van Dijl JM, Hecker M: **A proteomic view of an important human pathogen–towards the quantification of the entire *Staphylococcus aureus* proteome.** *PLoS One* 2009, **4**:e8176.
23. Ribeiro OC, Silva JAH, Oliveira SC, Meyer R, Fernandes GB: **Preliminary results on a living vaccince against caseous lymphadenitis.** *Pesquisa Agropecuaria Brasileira* 1991, **26**:461-465.
24. Simmons CP, Dunstan SJ, Tachedjian M, Krywult J, Hodgson AL, Strugnell RA: **Vaccine potential of attenuated mutants of *Corynebacterium pseudotuberculosis* in sheep.** *Infect Immun* 1998, **66**:474-479.
25. Patel VJ, Thalassinos K, Slade SE, Connolly JB, Crombie A, Murrell JC, Scrivens JH: **A comparison of labeling and label-free mass spectrometry-based proteomics approaches.** *J Proteome Res* 2009, **8**:3752-3759.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
27. Khamis A, Raoult D, La Scola B: **rpoB gene sequencing for identification of *Corynebacterium* species.** *J Clin Microbiol* 2004, **42**:3925-3931.
28. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
29. Bendtsen JD, Kiemer L, Fausbøll A, Brunak S: **Non-classical protein secretion in bacteria.** *BMC Microbiol* 2005, **5**:58.
30. Vanet A, Labigne A: **Evidence for specific secretion rather than autolysis in the release of some *Helicobacter pylori* proteins.** *Infect Immun* 1998, **66**:1023-1027.
31. Bendtsen JD, Wooldridge KG: **Non-Classical Secretion.** In *Bacterial secreted proteins: secretory mechanisms and role in pathogenesis* Edited by: Karl Wooldridge 2009, 225-239.
32. Jeffery CJ: **Moonlighting proteins–an update.** *Mol Biosyst* 2009, **5**:345-350.
33. Rodríguez-Ortega MJ, Norais N, Bensi G, Liberatori S, Capo S, Mora M, Scarselli M, Doro F, Ferrari G, Garaguso I, Maggi T, Neumann A, Covre A, Telford JL, Grandi G: **Characterization and identification of vaccine candidate proteins through analysis of the group A *Streptococcus* surface proteome.** *Nat Biotechnol* 2006, **24**:191-197.
34. Doro F, Liberatori S, Rodríguez-Ortega MJ, Rinaudo CD, Rosini R, Mora M, Scarselli M, Altindis E, D'Aurizio R, Stella M, Margarit I, Maione D, Telford JL, Norais N, Grandi G: **Surfome analysis as a fast track to vaccine discovery: identification of a novel protective antigen for Group B *Streptococcus* hypervirulent strain COH1.** *Mol Cell Proteomics* 2009, **8**:1728-1737.
35. Barbey C, Budin-Verneuil A, Cauchard S, Hartke A, Laugier C, Pichereau V, Petry S: **Proteomic analysis and immunogenicity of secreted proteins from *Rhodococcus equi* ATCC 33701.** *Vet Microbiol* 2009, **135**:334-345.
36. Hecker M, Becher D, Fuchs S, Engelmann S: **A proteomic view of cell physiology and virulence of *Staphylococcus aureus*.** *Int J Med Microbiol* 2010, **300**:76-87.
37. Hansmeier N, Chao T, Pühler A, Tauch A, Kalinowski J: **The cytosolic, cell surface and extracellular proteomes of the biotechnologically important soil bacterium *Corynebacterium efficiens* YS-314 in comparison to those of *Corynebacterium glutamicum* ATCC 13032.** *Proteomics* 2006, **6**:233-250.
38. Schaumburg J, Diekmann O, Hagendorff P, Bergmann S, Rohde M, Hammerschmidt S, Jänsch L, Wehland J, Kärst U: **The cell wall subproteome of *Listeria monocytogenes*.** *Proteomics* 2004, **4**:2991-3006.
39. Sibbald MJJB, Ziebandt AK, Engelmann S, Hecker M, de Jong A, Harmsen HJM, Raangs GC, Stokroos I, Arends JP, Dubois JYF, van Dijl JM: **Mapping the pathways to staphylococcal pathogenesis by comparative secretomics.** *Microbiol Mol Biol Rev* 2006, **70**:755-788.
40. Furuya H, Ikeda R: **Interaction of triosephosphate isomerase from the cell surface of Staphylococcus aureus and alpha-(1->3)-mannooligosaccharides derived from glucuronoxylomannan of *Cryptococcus neoformans*.** *Microbiology* 2009, **155**:2707-2713.
41. Söderberg MA, Cianciotto NP: **A *Legionella pneumophila* peptidyl-prolyl cis-trans isomerase present in culture supernatants is necessary for optimal growth at low temperatures.** *Appl Environ Microbiol* 2008, **74**:1634-1638.
42. Kunert A, Losse J, Gruszin C, Hühn M, Kaendler K, Mikkat S, Volke D, Hoffmann R, Jokiranta TS, Seeberger H, Moellmann U, Hellwage J, Zipfel PF: **Immune evasion of the human pathogen *Pseudomonas aeruginosa*: elongation factor Tuf is a factor H and plasminogen binding protein.** *J Immunol* 2007, **179**:2979-2988.
43. Tsugawa H, Ito H, Ohshima M, Okawa Y: **Cell adherence-promoted activity of *Plesiomonas shigelloides* groEL.** *J Med Microbiol* 2007, **56**:23-29.
44. Feng Y, Pan X, Sun W, Wang C, Zhang H, Li X, Ma Y, Shao Z, Ge J, Zheng F, Gao GF, Tang J: ***Streptococcus suis* enolase functions as a protective antigen displayed on the bacterial cell surface.** *J Infect Dis* 2009, **200**:1583-1592.
45. Pissavin C, Hugouvieux-Cotte-Pattat N: **Characterization of a periplasmic peptidyl-prolyl cis-trans isomerase in *Erwinia chrysanthemi*.** *FEMS Microbiol Lett* 1997, **157**:59-65.
46. Bergonzelli GE, Granato D, Pridmore RD, Marvin-Guy LF, Donnicola D, Corthésy-Theulaz IE: **GroEL of *Lactobacillus johnsonii* La1 (NCC 533) is cell surface associated: potential role in interactions with the host and the gastric pathogen *Helicobacter pylori*.** *Infect Immun* 2006, **74**:425-434.
47. He X, Zhuang Y, Zhang X, Li G: **Comparative proteome analysis of culture supernatant proteins of *Mycobacterium tuberculosis* H37Rv and H37Ra.** *Microbes Infect* 2003, **5**:851-856.
48. Sumby P, Whitney AR, Graviss EA, DeLeo FR, Musser JM: **Genome-wide analysis of group a streptococci reveals a mutation that modulates global phenotype and disease specificity.** *PLoS Pathog* 2006, **2**:e5.
49. Dumas E, Meunier B, Berdagué J, Chambon C, Desvaux M, Hébraud M: **Comparative analysis of extracellular and intracellular proteomes of *Listeria monocytogenes* strains reveals a correlation between protein expression and serovar.** *Appl Environ Microbiol* 2008, **74**:7399-7409.
50. van der Woude MW, Bäumler AJ: **Phase and antigenic variation in bacteria.** *Clin Microbiol Rev* 2004, **17**:581-611, table of contents.
51. Behr MA, Sherman DR: **Mycobacterial virulence and specialized secretion: same story, different ending.** *Nat Med* 2007, **13**:286-287.
52. McKean SC, Davies JK, Moore RJ: **Expression of phospholipase D the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays a role in macrophage death.** *Microbiology* 2007, **153**:2203-2211.
53. Hodgson AL, Krywult J, Corner LA, Rothel JS, Radford AJ: **Rational attenuation of *Corynebacterium pseudotuberculosis*: potential cheesy gland vaccine and live delivery vehicle.** *Infect Immun* 1992, **60**:2900-2905.
54. McNamara PJ, Bradley GA, Songer JG: **Targeted mutagenesis of the phospholipase D gene results in decreased virulence of *Corynebacterium pseudotuberculosis*.** *Mol Microbiol* 1994, **12**:921-930.
55. Moore RJ, Rothel L, Krywult J, Radford AJ, Lund K, Hodgson AL: **Foreign gene expression in *Corynebacterium pseudotuberculosis*: development of a live vaccine vector.** *Vaccine* 1999, **18**:487-497.

56. Meyer R, Carminati R, Bahia R, Vale V, Viegas S, Martinez T, Nascimento I, Schaer R, Silva J, Ribeiro M, Regis L, Paule B, Freire S: Evaluation of the goats humoral immune response induced by the *Corynebacterium pseudotuberculosis* lyophilized live vaccine. *J Med Biol Sci* 2002, **1**:42-48.
57. Walker J, Jackson HJ, Eggleton DG, Meeusen EN, Wilson MJ, Brandon MR: Identification of a novel antigen from *Corynebacterium pseudotuberculosis* that protects sheep against caseous lymphadenitis. *Infect Immun* 1994, **62**:2562-2567.
58. Koonin EV, Makarova KS, Aravind L: Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol* 2001, **55**:709-742.
59. Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EPC: Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr Biol* 2009, **19**:1683-1691.
60. Hett EC, Rubin EJ: Bacterial growth and cell division: a mycobacterial perspective. *Microbiol Mol Biol Rev* 2008, **72**:126-56, table of contents.
61. Allen CE, Schmitt MP: HtaA is an iron-regulated hemin binding protein involved in the utilization of heme iron in *Corynebacterium diphtheriae*. *J Bacteriol* 2009, **191**:2638-2648.
62. Puech V, Chami M, Lemassu A, Lanéelle MA, Schiffler B, Gounon P, Bayan N, Benz R, Daffé M: Structure of the cell envelope of corynebacteria: importance of the non-covalently bound lipids in the formation of the cell wall permeability barrier and fracture plane. *Microbiology* 2001, **147**:1365-1382.
63. Jordan S, Hutchings MI, Mascher T: Cell envelope stress response in Gram-positive bacteria. *FEMS Microbiol Rev* 2008, **32**:107-146.
64. Hansmeier N, Chao T, Daschkey S, Müsken M, Kalinowski J, Pühler A, Tauch A: A comprehensive proteome map of the lipid-requiring nosocomial pathogen *Corynebacterium jeikeium* K411. *Proteomics* 2007, **7**:1076-1096.
65. Suzuki N, Watanabe K, Okibe N, Tsuchida Y, Inui M, Yukawa H: Identification of new secreted proteins and secretion of heterologous amylase by *C. glutamicum*. *Appl Microbiol Biotechnol* 2009, **82**:491-500.
66. Hartmann M, Barsch A, Niehaus K, Pühler A, Tauch A, Kalinowski J: The glycosylated cell surface protein Rpf2, containing a resuscitation-promoting factor motif, is involved in intercellular communication of *Corynebacterium glutamicum*. *Arch Microbiol* 2004, **182**:299-312.
67. Sakamoto J, Shibata T, Mine T, Miyahara R, Torigoe T, Noguchi S, Matsushita K, Sone N: Cytochrome c oxidase contains an extra charged amino acid cluster in a new type of respiratory chain in the amino-acid-producing Gram-positive bacterium *Corynebacterium glutamicum*. *Microbiology* 2001, **147**:2865-2871.
68. Tsuge Y, Ogino H, Teramoto H, Inui M, Yukawa H: Deletion of cgR_1596 and cgR_2070, encoding NlpC/P60 proteins, causes a defect in cell separation in *Corynebacterium glutamicum* R. *J Bacteriol* 2008, **190**:8204-8214.
69. Körner H, Sofia HJ, Zumft WG: Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol Rev* 2003, **27**:559-592.
70. Oram D, Avdalovic A, Holmes R: Analysis of genes that encode DtxR-like transcriptional regulators in pathogenic and saprophytic corynebacterial species. *Infect Immun* 2004, **72**:1885-1895.
71. Kohl T, Baumbach J, Jungwirth B, Puhler A, Tauch A: The GlxR regulon of the amino acid producer *Corynebacterium glutamicum*: in silico and in vitro detection of DNA binding sites of a global transcription regulator. *J Biotechnol* 2008, **135**:340-350.
72. Stubben CJ, Duffield ML, Cooper IA, Ford DC, Gans JD, Karlyshev AV, Lingard B, Oyston PCF, de Rochefort A, Song J, Wren BW, Titball RW, Wolinsky M: Steps toward broad-spectrum therapeutics: discovering virulence-associated genes present in diverse human pathogens. *BMC Genomics* 2009, **10**:501.
73. Janson H, Melhus A, Hermansson A, Forsgren A: Protein D the glycerophosphodiester phosphodiesterase from *Haemophilus influenzae* with affinity for human immunoglobulin D influences virulence in a rat otitis model. *Infect Immun* 1994, **62**:4848-4854.
74. Braun V: Iron uptake mechanisms and their regulation in pathogenic bacteria. *Int J Med Microbiol* 2001, **291**:67-79.
75. Roe MR, Griffin TJ: Gel-free mass spectrometry-based high throughput proteomics: tools for studying biological response of proteins and proteomes. *Proteomics* 2006, **6**:4678-4687.
76. Panchaud A, Affolter M, Moreillon P, Kussmann M: Experimental and computational approaches to quantitative proteomics: status quo and outlook. *J Proteomics* 2008, **71**:19-33.
77. Pacheco LGC, Pena RR, Castro TLP, Dorella FA, Bahia RC, Carminati R, Frota MNL, Oliveira SC, Meyer R, Alves FSF, Miyoshi A, Azevedo V: Multiplex PCR assay for identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples. *J Med Microbiol* 2007, **56**:480-486.
78. Moura-Costa LF, Paule BJA, Azevedo V, Freire SM, Nascimento I, Schaer R, Regis LF, Vale VLC, Matos DP, Bahia RC, Carminati R, Meyer R: Chemically defined synthetic medium for *Corynebacterium pseudotuberculosis* culture. *Rev. Bras. Saúde e Produção Animal* 2002, **3**:1-9.
79. Nesvizhskii AI, Keller A, Kolker E, Aebersold R: A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem* 2003, **75**:4646-4658.
80. Silva JC, Denny R, Dorschel CA, Gorenstein M, Kass IJ, Li G, McKenna T, Nold MJ, Richardson K, Young P, Geromanos S: Quantitative proteomic analysis by accurate mass retention time pairs. *Anal Chem* 2005, **77**:2187-2200.
81. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: Prediction of twin-arginine signal peptides. *BMC Bioinformatics* 2005, **6**:167.
82. Wittkop T, Emig D, Lange S, Rahmann S, Albrecht M, Morris JH, Böcker S, Stoye J, Baumbach J: Partitioning biological data with transitivity clustering. *Nat Methods* 2010, **7**:419-420.
83. Baumbach J, Wittkop T, Kleindt CK, Tauch A: Integrated analysis and reconstruction of microbial transcriptional gene regulatory networks using CoryneRegNet. *Nat Protoc* 2009, **4**:992-1005.
84. Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, Dopazo J, Conesa A: High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res* 2008, **36**:3420-3435.

doi:10.1186/1471-2180-11-12
Cite this article as: Pacheco *et al.*: A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*. *BMC Microbiology* 2011 **11**:12.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

3.2.3 Identificação de 11 novas proteínas exportadas da *C. pseudotuberculosis* utilizando análise comparativa do secretoma

Este estudo, que encontra-se aceito para publicação, envolve a comparação entre os proteomas exportados das linhagens 1002 e C231 da *C. pseudotuberculosis*. A razão da escolha dessas linhagens deve-se ao fato de que a doença linfadenite caseosa causa grandes perdas econômicas mundiais por infectar rebanhos de caprinos e ovinos, principais hospedeiros dessas linhagens. No estudo anterior (Seção 3.2.2), com base em um sistema livre de gel (TPP-LC/MSE), foram identificados 70 proteínas exportados para a linhagem 1002 e 67 para a linhagem C231, totalizando 93 diferentes proteínas extracelulares identificadas para a espécie. Seguindo uma estratégia complementar, no presente trabalho, utilizou-se a eletroforese 2D para encontrar proteínas extracelulares em ambas as linhagens. Essas proteínas foram posteriormente degradadas com tripsina, analisadas por MALDI-TOF/TOF e, finalmente, identificadas com o programa *Mascot*. Um total de 45 proteínas extracelulares de *C. pseudotuberculosis* foram identificadas através deste método. A análise comparativa entre as linhagens 1002 e C231 mostrou 13 e 3 proteínas, respectivamente, pertencentes ao proteoma exportado variante, enquanto que um subconjunto de 11 destas proteínas não haviam sido previamente detectadas no exoproteoma central. As proteínas recentemente identificadas podem desempenhar um papel importante na fisiologia e virulência de *C. pseudotuberculosis*, além de confirmar as predições *in silico*.

Identification of 11 new exoproteins of *Corynebacterium pseudotuberculosis* through comparative analysis of the secretome

Wanderson M. Silva[1], Núbia Seyffert[1], Thiago L. P. Castro[1], Agenor V. Santos[2], Luis G. C. Pacheco[3], Anderson R. Santos[1], Alessandra Ciprandi[2], Meritxell Zurita-Turk[1], Fernanda A. Dorella[1], Hélida M. Andrade[4], Adriano M. C. Pimenta[5], Artur Silva[2], Anderson Miyoshi[1] and Vasco Azevedo[1]

[1]Departamento de Biologia Geral. Instituto de Ciências Biológicas. Universidade Federal de Minas Gerais. Av. Antonio Carlos, 6627 - Pampulha, CP 486, CEP 31.270-901. Belo Horizonte- MG, Brazil.

[2]Instituto de Ciências Biológicas, Universidade Federal do Pará, Rua Augusto Corrêa, 01 - Guamá, Belém, PA, Brazil.

[3]Instituto de Ciências da Saúde. Universidade Federal da Bahia. Av. Reitor Miguel Calmon, s/n - Vale do Canela, CEP 40110100. Salvador-BA, Brazil.

[4]Departamento de Parasitologia. Instituto de Ciências Biológicas. Universidade Federal de Minas Gerais. Av. Antonio Carlos, 6627 - Pampulha, CP 486, CEP 31.270-901. Belo Horizonte- MG, Brazil.

[5]Departamento de Bioquímica e Imunologia. Instituto de Ciências Biológicas. Universidade Federal de Minas Gerais. Av. Antonio Carlos, 6627 - Pampulha, CP 486, CEP 31.270-901. Belo Horizonte- MG, Brazil.

Corresponding authors: Vasco Azevedo[1]; **mail:** vasco@icb.ufmg.br and Wanderson M. Silva[1]; **mail:** silvamarques@yahoo.com.br

Emails:

WMS:	silvamarques@yahoo.com.br
NS:	nbseyffert@gmail.com
TLPC:	castrotlp@gmail.comA
AVS:	valadaresantos@gmail.com
LGCP:	lgcpacheco@gmail.com
ARS:	anderson2010@gmail.com
AC:	alecip@gmail.com
MZ:	meritxellzt@gmail.com
FAD:	fernandadorella@gmail.com
HMA:	helidandrade@gmail.com
AMCP:	apimenta@icb.ufmg.br
AS:	asilva@ufpa.br
AM:	miyoshi@icb.ufmg.br
AV:	vasco@icb.ufmg.br

Abstract

This study involves the comparison between the exoproteomes of two different strains of *Corynebacterium pseudotuberculosis*, the etiologic agent of caseous lymphadenitis in small ruminants. In a previous study, based on a gel-free system (TPP-LC/MS^E), we have identified 70 exoproteins for the strain 1002 and 67 for the strain C231, totalizing so far 93 different identified extracellular proteins for the species. Following a supplementary strategy, in the present work we have used 2D electrophoresis to resolve both strains extracellular proteins, which were thereafter digested with trypsin, analyzed by MALDI-TOF/TOF and finally identified with the MASCOT software. A total set of 45 extracellular proteins of *C. pseudotuberculosis* has been identified through this methodology. The comparative analysis among the strains 1002 and C231 showed 13 and 3 respectively belonging unique proteins, whereas a subset of 11 of these proteins had not been previously detected. The newly identified proteins can play an important role in physiology and virulence of *C. pseudotuberculosis.*

Background

Corynebacterium pseudotuberculosis is a Gram positive facultative intracellular pathogen, responsible for infectious diseases in small ruminants, horses, bovine and, occasionally, in humans [1]. In goats and sheep, *C. pseudotuberculosis* is the etiologic agent of caseous lymphadenitis (CLA), a chronic disease characterized by the formation of abscesses in on the skin (surface) and in internal organs [2]. CLA is distributed worldwide and responsible for great economic losses in countries where there is intense activity in the goat and sheep industry [2]. Currently, there are no efficient vaccines, neither an effective sub-clinical diagnosis method available for detection of this disease in its early stage.

The main virulence factor of *C. pseudotuberculosis* that contributes to the pathogenic mechanism of CLA is phospholipase D (PLD), a potent exotoxin with sphingomyelinase activity that favours the spread of this pathogen in the host [3]. In addition to PLD, other virulence factors include iron transporters belonging to the ABC proteins family (important to disease progression) [4], and a serine protease [5].

To increase our knowledge on the molecular basis of genes related to virulence factors and to better understand *C. pseudotuberculosis* physiology, our research group has carried out sequencing, assembling and annotation of the genome of two *C. pseudotuberculosis* strains: 1002, isolated from a goat in Brazil, and C231, isolated from a sheep in Australia [6].

Several studies demonstrate that extracellular proteins are important virulence factors, such as those associated with cell adhesion, cell invasion, survival and proliferation in the host cell, and escape from the immune system [7]. Some studies have been performed to characterize the most immunogenic fractions of *C. pseudotuberculosis* extracellular proteins, to be used as targets for the development of vaccines and diagnostics to combat CLA [8-10]. Recently, our research group has conducted a detailed study of the extracellular proteins of *C. pseudotuberculosis*, comparing the exoproteomes of strains 1002 and C231. This study combined the technique of three-phase partitioning (TPP) [11] and a gel-free separation method using liquid chromatography coupled with mass spectrometry (LC-MS), called TPP-LC/MS^E. Such strategy enabled us to characterize 93 extracellular proteins of *C. pseudotuberculosis*, including new factors probably associated with virulence [12].

In this paper, a complementary method of extracellular bacterial proteome analysis using two-dimensional electrophoresis (2-DE) associated with mass spectrometry MALDI-TOF/TOF type, was used to generate exoproteome maps of the 1002 and C231 strains. The gel-dependent approach

(2-DE-MALDI-TOF/TOF) is widely used in proteomics to characterize protein maps, besides allowing the identification of post-translational modifications [13,14]. Based on this methodology, it was possible to identify 11 extracellular proteins of *C. pseudotuberculosis* that had not been detected in the previous study by Pacheco *et al.* [12]. Combining the results obtained by TPP-LC/MS^{E} [12] and 2-DE-MALDI-TOF/TOF methodologies, a total of 104 extracellular proteins were successfully characterized for *C. pseudotuberculosis*.

Results

The proteome reference maps of two *C. pseudotuberculosis* strains and protein identifications by MALDI-TOF/TOF

After obtaining the extracellular proteins of *C. pseudotuberculosis* strains by TPP technique, the proteins were resolved using 2-DE. For each strain, 300 µg of proteins were applied on strips of 18 cm with a pH range between 3-10 N.L (Figures 1 and 2). Three biological replicates of 2-DE gels were scanned using an Image Scanner (GE Healthcare) and the Image Master 2D Platinum 7 (GE Healthcare) software was used to analyze the generated images. Fine adjustments for the gels images were done based on the spots that were reproducibly detected by the software and further manual analysis was carried out to eliminate possible artifacts. Based on these parameters, 85 spots were detected in the extracellular proteome of strain 1002 (isolated from goat), while 80 spots were detected for strain C231 (isolated from sheep). It was observed that electrophoretic patterns between the strains were similar, with the majority of spots concentrated in the acidic range of the gels (p.I 3-5).

After obtaining the results from images of the gels, spots were excised from gels, submitted to tryptic digestion and analyzed by MALDI-TOF MS/MS. Researches in databases using the MASCOT software allowed the identification of 55 protein spots (65%) in strain 1002 (Figure 1) and 45 (56%) in strain C231 (Figure 2). Some spots could not be identified possibly due to low protein amounts and occurrence of post-translational modifications. By this approach, it was possible to identify 45 *C. pseudotuberculosis* proteins, with 29 of them common to both strains (Additional file 1). Among the identified proteins as exclusive, we distinguished 13 unique of strain 1002 (Additional file 2) and 3 unique of strain C231 (Additional file 3). Some protein spots were identified as the same protein, revealing the presence of isoforms possibly due to pos-translational modifications.

Comparative analysis between the strains exoproteomes

Proteins exclusively found in *C. pseudotuberculosis* 1002 gels (Additional file 2) are related to several biological processes, such as response to heat and oxidative stresses, energy production and molecules transport mechanisms. Bioinformatics analysis revealed that all 13 proteins present only in 1002 have correspondent Open Reading Frames (*ORF)* in this strain genome, but 12 of them also hold correspondent *ORFs* in the C231 genome. The *ORF* that is not in the C231 genome but is present in 1002 and in another 2 *C. pseudotuberculosis* strains (FRC41 and I19) encodes for a putative DsbG protein (ADL21555), which belongs to the superfamily of thioredoxin and acts as a chaperone in the formation of disulfide bonds in proteins secreted by *E. coli* [15].

In *C. pseudotuberculosis* C231, the proteins exclusively (Additional file 3) are probably related to virulence and unknown functions (hypothetical protein). DNA sequences encoding for all these proteins were observed in the genomes of both 1002 and C231 strains.

Same proteins with unknown functions were also identified for both strains. According to the databases, one specific hypothetical protein (ADL09626) showed no similarity to any other protein of the *Corynebacterium* genus or other bacterial genera, suggesting that it is unique to the *C. pseudotuberculosis* species. Therefore, important to perform specific studies to try to assess whether it is possible that the ADL09626 protein plays an important role in the pathogenesis and virulence of this microorganism.

Predicting the sub-cellular localization of identified exoproteins

In silico predictions of sub-cellular localization of *C. pseudotuberculosis* extracellular proteins were performed using the software SurfG+ [16]. The predictions point out the presence of 9 cytoplasmic proteins, 39 secreted proteins and 4 proteins possibly associated to the cell wall. Likewise, other studies have detected cytoplasmic or cell surface associated proteins in extracellular extracts [12,17,18].

Analysis of *C. pseudotuberculosis* extracellular proteins using different approaches

Pacheco *et al.* [12] characterized the exoproteome of *C. pseudotuberculosis* strains 1002 (70 proteins) and C231 (67 proteins) by TPP/LC-MS^E, resulting so far in a total of 93 extracellular proteins identified for the specie. The present work, based on the combination of 2-DE and MALDI-TOF/TOF techniques, allowed the identification of 42 proteins in strain 1002 and 32 in strain C231, resulting in a total of 45 different extracellular proteins identified for *C. pseudotuberculosis*. Despite of the lower number of identifications given herein, it was possible to demonstrate the presence of 11 proteins that had not been previously detected by Pacheco *et al.* [12]

(Table 1). Combining the results obtained by both studies, 81 and 73 extracellular proteins were characterized for strains 1002 and C231 respectively, totaling 104 exported proteins for *C. pseudotuberculosis* (Additional file 4). Figure 3 shows the distribution of proteins identified by the different approaches.

Discussion

Comparative protein studies of different bacterial strains have become a powerful approach to characterize the proteome of an organism, contributing for the understanding of the functional state of its genome and for the identification of new virulence factors [7,14]. In this study, a comparative analysis between extracellular proteins of C231 and 1002 *C. pseudotuberculosis* strains was carried out. Exoproteome differences observed between strains may be associated to the fact that they have been isolated from different hosts in different geographic locations; whereas C231 strain was isolated from a sheep in Australia, 1002 strain was isolated from a goat in Brazil. Among the differences observed, what really draws attention is that, in 1002 gels, it was not possible to detect spots referent to PLD and CP40 proteins. PLD is the most important virulence determinant identified in *C. pseudotuberculosis* [19] and CP40 is a secreted serine protease that, when used as an antigen for vaccine, has caused substantial protection against CLA [5]. Similarly, these same proteins were not identified by Pacheco *et al.* [12]. The lack of PLD production in 1002 can be explained as a result of attenuation of this bacterial strain or due to the fact that this bacterium only expresses these proteins during its infection process. A study based on the analysis of *pld* gene regulation showed that its expression is related to multiple factors, which include dependence on cell density and thermal regulation. Furthermore, it was demonstrated that *pld* expression is significantly increased during infection of macrophages *in vitro* [19].

In this study, cytoplasmic proteins were identified among the extracellular extracts, possibly because both strains may feature an alternative secretion pathway, i.e., such proteins may be exported by non-classical secretion. The detected proteins elongation factor Tu, GroEL, Enolase, Glyceraldehyde-3-phosphate dehydrogenase and superoxide dismutase (SodA) were found to be, according to other studies, dependent on a non-classical secretion pathway via SecA [20-22]. After sequencing the genomes of strains 1002 and C231, it could be observed that both carry two *SecA* genes (*SecA1* and *SecA2*), possibly involved in the *C. pseudotuberculosis* secretion systems [6]. *SecA2* has been described in *Bacillus subtilis* [21] and pathogenic bacteria such as *Listeria monocytogenes* [22] and *Mycobacterium tuberculosis* [23].

Braunstein *et al.* [23], generated mutant strains for the *SecA2* gene in *M. tuberculosis* and showed that some secreted proteins related to virulence factors were not expressed in two-dimensional gels, while attenuating the virulence of *M. tuberculosis* in mice. Among these proteins, superoxide dismutase, a protein related to oxidative stress response, which has a great importance for intracellular pathogens, was missing. This protein has been identified in several proteomic studies as a cytoplasmic protein, due to the lack of a signal peptide sequence. This study demonstrated that SodA is a secreted protein, being dependent on an alternative secretion pathway. Lenz *et al.* [22], generated a mutant strain for the *SecA* gene in *Listeria monocytogenes,* and detected proteins with signal peptide sequences and protein sequences without signal, demonstrating that this route can export proteins both with and without signal peptide sequences. This *Sec2* secretion pathway is important for virulence factor export; however the general mechanism of how proteins are exported by this route and other non-classical secretory mechanisms is still unknown.

A classic proteomics technique, based on 2-DE, led us to identify 11 proteins not detected using the TPP-LC/MS (gel-independent) approach in a previous study [12]. The combination of both techniques evidenced the complementarity of them, once the previously number of identified extracellular proteins has increased from 93 to 104. The differences observed in both studies may be associated to the separation methods, which involve distinct buffers for sample solubilization, and physic-chemical properties of each protein. The use of more than one methodology in large-scale proteomics is widely applied; each complements the technical limitations of the other, resulting in increased percentage of identified proteins [24-26].

Another question that also deserves attention for proteomics involving prokaryotes is the growth phase in which the studies are held. This is a possible explanation for the non-identification by Pacheco *et al.* [12] of 11 proteins newly described in the present work, since that whereas the extractions of proteins in this study were held at the late exponential growth phase, Pacheco *et al.* [12] held their extractions in the early exponential phase. Among the 11 proteins newly identified in this study (Table 1), 3 proteins have unknown functions (hypothetical proteins) and 8 have been related to various physiological functions and virulence factors.

GroEL (HSP60) and DnaK (HSP70) are two of the major and best-studied chaperone proteins; they act in folding and translocation of proteins and are essential for cell growth at physiologically relevant temperatures [27]. In order to understand the role of DnaK and GroEL in the physiology of *Streptococcus mutants*, Lemos *et al.* [28] conducted a study where the knockout strains SM12 (Δ*dnaK*) and SM13 (Δ*groEL*) showed slower growth and changes in their proteomes when compared to the wild-type strain. At higher temperature (44ºC) or lower pH (5.0), SM12 was impaired in its capacity to grow and was more susceptible to H_2O_2, while SM13 showed equivalent growth profiles compared to the wild-strain under these conditions. Moreover, SM12 and SM13 showed affected biofilm-forming capacities [28]. These results demonstrate that chaperones DnaK and GroEL play a key role in bacterial physiology in such a way that they can influence the process of bacterial infection.

Elongation factor P (EFP) is a highly conserved protein in prokaryotes that stimulates the peptidyltransferase activity of 70S ribosomes and enhances the synthesis of certain dipeptides initiated by *N*-formylmethionine [29]. Aoki *et al.* [30] demonstrated that the *efp* gene is essential for *E. coli* growth and protein synthesis. Navarre *et al.* [31], observed that the interaction between *poxA*, *yjek* and *efp* are of great importance in virulence and stress response of *Salmonella enterica*.

Enolase is an enzyme of the glycolytic pathway which catalyzes the reversible conversion of 2-phosphoglycerate (2-PGE) into phosphoenolpyruvate (PEP). Studies suggest that enolases from prokaryotic cells possess fibronectin-binding activity and influence in bacterial colonization, invasion and persistence in host tissues [32]. This protein has also been characterized as an immunodominant antigen in *Bacillus anthracis* [33]. In addition, a recombinant enolase of *B. anthracis* is capable of binding to laminin present on mammalian extracellular matrix, besides interacting with plasminogen. These results suggest that enolase may contribute to raising the invasive potential of *B. anthracis* [34].

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH), a highly conserved enzyme in both prokaryotes and eukaryotes, participates in the glycolytic pathway and is associated to the mechanisms of oxidation and phosphorylation of glyceraldehyde 3-phosphate to 1,3-bisphosphoglycerate [35]. This protein is present in cellular surfaces and secretomes of various pathogens, and is also associated to *Streptococcus agalactiae* virulence mechanism due to its ability to modulate the host immune system during infection [36].

ABC-type transporters in prokaryotes are classified in three functional categories: (i) importers in nutrient uptake; (ii) exporters that are associated to the secretion of various molecules, and (iii) involved in translation of mRNA and in DNA repair [37]. ABC transporters can also play important roles in the viability, virulence and bacterial pathogenicity [38], like the iron ABC uptake system. Bacteria that features this system secrete molecules called siderophores, which in turn have high affinity for iron and act as chelators of this metal in host cells. Studies have demonstrated that bacteria carrying defective genes for siderophores have reduced ability to cause disease [38].

Carbonic anhydrase catalyzes the reversible hydration of CO_2 to bicarbonate, helping maintain pH homeostasis and performing various physiological functions, such as respiration, ions transportation and bacterial growth [39,40]. Studies in *Neisseria spp. H.* pylori, *B. suis* and *S. pneumoniae* have demonstrated that some enzymes can inhibit the action of the carbonic anhydrase, thereby inhibiting bacterial growth *in vivo* [41].

Manganese superoxide dismutase (MnSOD) is a isoform to the superoxide dismutase family and is an important component of the antioxidant defense mechanism that act to eliminate reactive oxygen species (ROS) [42,43]. In *Enterococcus faecalis* the MnSOD contributes to survival of the bacterium in macrophages, could even important during the pathogenesis process [44]. However, studies showed that this protein is involved in response to acid stress [45,46].

Thus, the results obtained in this study demonstrated that the use of complementary proteomic techniques is an efficient strategy for characterization of bacterial secretomes.

Conclusion

This study allowed the characterization of new extracellular proteins of *C. pseudotuberculosis* that have not been reported in other studies. The achieved results increase the extracellular protein catalogue of *C. pseudotuberculosis* and validate former *in silico* predictions, hereby extending our knowledge regarding the functional status of the genome of this bacterium under exponential growth condition. Therefore, this work clearly demonstrates that comparative analysis combining different proteomic approaches, i.e. LC/MS^E and 2-DE-MALDI-TOF/TOF, is an attractive strategy to characterize a bacterial proteome. These new identified proteins are associated with physiology and virulence, probably playing an important role in the pathogenicity mechanism of CLA. In addition, these proteins may represent new potential targets for use in prophylaxis against this disease.

Methods

Bacterial strains and culture conditions

The wild-type *C. pseudotuberculosis* strain 1002 was provided by Dr. Roberto Meyer of the Institute of Health Sciences, University of Bahia, Brazil; and wild-type *C. pseudotuberculosis* strain C231 was provided by Dr. Robert Moore of the CISRO Livestock Industries, Australian Animal Health Laboratory, Australia. The bacterial strains were routinely maintained in Brain Heart Infusion (BHI) broth or bacteriological agar plates at 37°C. For extraction of extracellular proteins, the strains of *C. pseudotuberculosis* were cultured in chemically defined medium (CDM) [($Na_2HPO_4{\cdot}7H_2O$ (12.93 g/L), KH_2PO_4 (2.55 g/L), NH_4Cl (1 g/L), $MgSO_4.7H_2O$ (0.20 g/L), $CaCl_2$ (0.02 g/L), and 0.05% (v/v) Tween 80]; 4% (v/v) MEM Vitamins Solution 100X (Invitrogen); 1% (v/v) MEM Amino Acids Solution 50X (Invitrogen); 1% (v/v) MEM Non Essential Amino Acids

Solution 100X (Invitrogen); and 1.2% (w/v) filter-sterilized glucose) at 37°C [47], until the set point of exponential growth (DO_{600nm} = 1.3) is reached.

Three-phase partitioning

The TPP protocol, optimized by our group, was used to extract extracellular proteins [11]. After cultivation of three biological replicates for each strain, cultures were centrifuged for 20 min at 2700 xg. Supernatants were filtered using 0.22 µm filters; ammonium sulphate was added to samples at 30% (w/v) and the pH of mixtures was set to 4.0. N-butanol was then added to each sample at an equal volume; samples were vigorously vortexed and left to rest for 1 h at RT; centrifuged for 10 min at 1350 xg at 4 ºC. The interfacial precipitate was collected in 1.5 mL microtubes, and re-suspended in 1 mL Tris 20 mM + 10 µL protease inhibitor. The protein concentration was determined using a standard curve [48].

Two-dimensional electrophoresis (2-DE)

Approximately 300 µg of proteins were dissolved in 450 µL of rehydration buffer (Urea 7M, thiourea 2M, 3-[(3-Cholamidopropyl)dimethylammonio]-1-propanesulfonate 2%, Tris 40 mM, bromophenol blue 0,002%, DTT 75 mM, IPG Buffer 1%). The samples were applied to 18 cm pH 3-10 NL strips (GE Healthcare). Isoelectric focusing (IEF) was performed using the apparatus IPGphor 2 (GE Healthcare) under the following voltages: 100V 1hr, 500V 2hr, 1000V 2hr, 10000V 3hr, 10000V 60000Vhr, 500V 4 hr. The strips were kept at equilibrium for 15 minutes in 10 mL of equilibration buffer I (Tris-HCl 50 mM pH 8.8, Urea 6M, Glycerol 30%, SDS 2%, bromophenol blue 0,002%, 100mg dithiothreitol) and in 10 mL of equilibration buffer II (Tris-HCl 50 mM pH 8.8, urea 6M, Glycerol 30%, SDS 2%, bromophenol blue 0,002%, iodoacetamide 250mg). The isolated proteins were separated in 12 % acrylamide/bis-acrylamide gels with an Ettan DaltSix II system (GE Healthcare). To visualize the separated proteins, gels were stained with Coomassie blue G-250 staining solution.

In-gel tryptic digestion of proteins

Protein spots were excised from the gels using an *Ettan Spot Picker* (GE Healthcare) and fragments containing the excised spots were washed with ultra sterile water for 5 minutes and dehydrated with acetonitrile (ACN) for 20 minutes. Following, the fragments were dried in speed vac. Protein digestion was carried out by adding 10 µL of a stock solution of trypsin (Promega, Sequencing Grade Modified Trypsin) (33 ng / mL - ca. 1.5 µM) to each tube for 60 minutes at 4°C. After removal of trypsin excess, samples were incubated at 58°C for 30 minutes. Digestion was interrupted by adding 1 mL of 5% formic acid (v/v). The extraction of peptides was performed using 30 mL of formic acid solution 5% - 50% ACN and the sample subjected to ultrasound. The peptides were concentrated to a volume of 10 mL using speed vac and desalinated and concentrated using ZIP-TIP C18 tips (Eppendorf) [49]. The samples were stored at -20°C for subsequent analysis by mass spectrometry.

Mass spectra and database search

Analyses by MS and MS/MS modes were performed using a MALDI-TOF/TOF mass spectrometer Autoflex III^{TM} (Brucker Daltonics, Billerica USA). The equipment was controlled in a positive/reflector using the FlexControl^{TM} software. Calibration was made using samples of standard peptides (Angiotensin II, Angiotensin I, Substance P, Bombesin, ACTH clip 1-17, ACTH clip 18-39, Somatostatin 28, Bradykinin Fragment 1-7, Renin Substrate tetradecapeptide porcine) (Brucker Daltonics, Billerica, USA). The peptides were added to the alpha-cyano-4-hydroxycinnamic acid matrix, applied on an AnchorChipTM 600 plate (Brucker Daltonics, Billerica, USA) and analyzed by Autoflex III. The following search parameters were: peptide mass fingerprint; enzyme; trypsin; fixed modification, carbamido methylation (Cys); variable modifications, oxidation (Met); mass values, monoisotopic; maximum missed cleavages, 1; and peptide mass tolerance of 0.05% Da (50 ppm). The results obtained by MS/MS were used for identifying proteins with the program MASCOT$^{®}$ (http://www.matrixscience.com) and compared with the NCBI databases.

In silico predictions of protein sub-cellular localization

To predict the sub-cellular localization of proteins we used the following programs: SurfG+ v1.0 [16], SecretomeP v2.0 [20] and TatP v1.0 [50].

Abbreviations

CLA: caseous lymphadenitis; LC-MS: liquid chromatography - mass spectrometry; PLD: phospholipase D; PSE: potentially surface exposed; TPP: Three-Phase Partitioning; 2-DE: two-dimensional electrophoresis; MALDI-TOF matrix-assisted laser desorption/ionization -time-of-flight; MS: mass spectrometer; HSP: heat shock protein; RT: room temperature.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WMS, NS, TLPC, LGCP, FAD, MZT, performed microbiological experiments. WMS, NS, AVS, LGCP, AC, HMA, performed proteomic experiments. WMS, ARS performed bioinformatical analyses. WMS, AMCP, AS, AM and VA participated in the design of the study, and critical analysis of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We are thankful to the Center for Study of Structure and Function of Biomolecules of the Institute of Biological Sciences of the Federal University of Minas Gerais and to the Genomic and Proteomic network of the State of Pará. The current work was supported by CNPq, FAPEMIG and FAPESPA.

References

1. Dorella FA, Pacheco LG, Oliveira SC, Miyoshi A, Azevedo V: ***Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence.** *Vet. Res* 2006, **37:**201-218.

2. Williamson LH: **Caseous lymphadenitis in small ruminants.** *Vet Clin North Am Food Anim Prac* 2001, **17:**359-371.

3. Songer JG: **Bacterial phospholipases and their role in virulence.** *Trends Microbiol* 1997, **5:**156-160.

4. Billington SJ, Esmay PA, Songer JG, Jost BH: **Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*.** *J Bacteriol* 2002, **180:**3233-3236.

5. Wilson MJ, Brandon MR, Walker J: **Molecular and Biochemical Characterization of a Protective 40- Kilodalton Antigen from *Corynebacterium pseudotuberculosis*.** *Infection and immunity* 1995, **63:**206-211.

6. Ruiz JC, D'Afonseca V, Silva A, Ali A, Pinto AC, Santos AR, Rocha AA, Lopes DO, Dorella FA, Pacheco LG, Costa MP, Turk MZ, Seyffert N, Moraes PM, Soares SC, Almeida SS, Castro TL, Abreu VA, Trost E, Baumbach J, Tauch A, Schneider MP, McCulloch J, Cerdeira LT, Ramos RT, Zerlotini A, Dominitini A, Resende DM, Coser EM, Oliveira LM, Pedrosa AL, Vieira CU, Guimarães CT, Bartholomeu DC, Oliveira DM, Santos FR, Rabelo ÉM, Lobo FP, Franco GR, Costa AF, Castro IM, Dias SR, Ferro JA, Ortega JM, Paiva LV, Goulart LR, Almeida JF, Ferro MI, Carneiro NP, Falcão PR, Grynberg P, Teixeira SM, Brommonschenkel

S, Oliveira SC, Meyer R, Moore RJ, Miyoshi A, Oliveira GC, Azevedo V: **Evidence for Reductive Genome Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains.** *Plos One* 2011, **6:**e18551.

7. Wu Z, Zhang W, Lu C: **Comparative proteome analysis of secreted proteins of *Streptococcus suis* serotype 9 isolates from diseased and healthy pigs.** *Microbial Pathogenesis* 2008, **45:**159-166.

8. Meyer R, Regis L, Vale V, Paule B, Carminati R, Bahia R, Moura-Costa L, Schaer R, Nascimento I, Freire S: **In vitro IFN-gamma production by goat blood cells after stimulation with somatic and secreted *Corynebacterium pseudotuberculosis* antigens.** *Vet Immunol immunolpathol* 2005, **15:**249-54.

9. Moura-Costa LF, Bahia RC, Carminati R, Vale VL, Paule BJ, Portela RW, Freire SM, Nascimento I, Schaer R, Barreto LM, Meyer R: **Evaluation of the humoral and cellular immune response to different antigens of *Corynebacterium pseudotuberculosis* in Caninde´ goats and their potential protection against caseous lymphadenitis.** *Vet Immun Immunopath* 2008, **126:**131-141.

10. Rebouças MF, Portela RW, Lima DD, Loureiro D, Bastos BL, Moura-Costa LF, Vale VL, Miyoshi A, Azevedo V, Meyer R: ***Corynebacterium pseudotuberculosis* secreted antigen-induced specific gamma-interferon production by peripheral blood leukocytes: potential diagnostic marker for caseous lymphadenitis in sheep and goats.** *J Vet Diagn Invest* 2011, **23:**213-220.

11. Paule BJ, Meyer R, Moura-Costa LF, Bahia RC, Carminati R, Regis LF, Vale VL, Freire SM, Nascimento I, Schaer R, Azevedo V: **Three-phase partitioning as an efficient method for extraction/concentration of immunoreactive excreted-secreted proteins *of Corynebacterium pseudotuberculosis.*** *Protein Expr Purif* 2004, **34:**311-166.

12. Pacheco LG, Slade SE, Seyffert N, Santos AR, Castro TL, Silva WM, Santos AV, Santos SG, Farias LM, Carvalho MA, Pimenta AM, Meyer R, Silva A, Scrivens JH, Oliveira SC, Miyoshi A, Dowson CG, Azevedo V: **A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*.** *BMC Microbiol* 2011, **11:**12.

13. Halligan BD, Ruotti V, Jin W, Laffoon S, Twigger SN, Dratz EA: **ProMoST (Protein Modification Screening Tool): a web-based tool for mapping protein modifications on two-dimensional gels.** *Nucleic Acids Research* 2004, **32:**638-644.

14. Pocsfalvi G, Cacace G, Cuccurullo M, Serluca G, Sorrentino A, Schlosser G, Blaiotta G, Malorni A: **Proteomic analysis of exoproteins expressed by enterotoxigenic *Staphylococcus aureus* strains.** *Proteomics* 2008, **8:**2462-2476.

15. Bessette PH, Cotto JJ, Gilbert HF, Georgiou G: ***In Vivo* and *in Vitro* Function of the *Escherichia coli* Periplasmic Cysteine Oxidoreductase DsbG.** *J biological chemistry* 1999, **274:**7784-7792.

16. Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Maquin E, van de Guchte M: **Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria.** *Proteomics* 2009, **9:**61-73.

17. Hansmeier N, Chao TC, Daschkey S, Müsken M, Kalinowski J, Pühler A, Tauch A: **A comprehensive proteome map of the lipid-requiring nosocomial pathogen *Corynebacterium jeikeium* K411.** *Proteomics* 2007, **7:**1076-1096.

18. Hansmeier N, Chao TC, Kalinowski J, Puhler A, Tauch A: **Mapping and comprehensive analysis of the extracellular and cell surface proteome of the human pathogen *Corynebacterium diphtheriae*.** *Proteomics* 2006, **6:**2465-2476.

19. McKean SC, Davies JK, Moore RJ: **Expression of phospholipase D, the major virulence factor of *Corynebacterium pseudotuberculosis*, is regulated by multiple environmental factors and plays role in macrophage death.** *Microbiology* 2007, **153:**2203-2211.

20. Bendtsen JD, Kiemer L, Fausboll A, Brunak S: **Non-classical protein secretion in bacteria.** *BMC Microbiol* 2005, **5:**58.

21. Hirose I, Sano K, Shioda I, Kumano M, Nakamura K, Yamane K: **Proteome analysis of *Bacillus subtilis* extracellular proteins: a two-dimensional protein electrophoretic study.** *Mycrbobiology* 2000, **146:**65-75.

22. Lenz LL, Mohammadi S, Geissler A, Portnoy DA: **SecA2-dependent secretion of autolytic enzymes promotes *Listeria monocytogenes* pathogenesis.** *Proc Natl Acad Sci USA* 2003, **14:**12432-12437.

23. Braunstein M, Espinosa BJ, Chan J, Belisle JT, Jacobs WR Jr: **SecA2 functions in the secretion of superoxide dismutase A and in the virulence of *Mycobacterium tuberculosis*.** *Mol Microbiol* 2003, **48:**453-464.

24. Wolff S, Otto A, Albrecht D, Zeng JS, Büttner K, Glückmann M, Hecker M, Becher D: **Gel-free and gel-based proteomics in Bacillus subtilis: a comparative study.** *Mol Cell Proteomics* 2006, **5:**1183-1192.

25. Dumpala PR, Lawrence ML, Karsi A: **Proteome analysis of *Edwardsiella ictaluri*.** *Proteomics* 2009, **9:**1353-1363.

26. Van Cutsem E, Simonart G, Degand H, Faber AM, Morsomme P, Boutry M: **Gel-based and gel-free proteomic analysis of Nicotiana tabacum trichomes identifies proteins involved in secondary metabolism and in the (a)biotic stress response.** *Proteomics* 2011, **3:**440-454.

27. Craig EA, Cambill BD, Nelson RJ: **Heat shock proteins: molecular chaperones of protein biogenesis.** *Microbiol Rev* 1993, **57:**402-414.

28. Lemos JA, Luzardo Y, Burne RA: **Physiologic Effects of Forced Down-Regulation of *dnaK* and *groEL* Expression in *Streptococcus mutans*.** *J Bacteriology* 2007, **189:**1582-1588.

29. Aoki H, Dekany K, Adams SL, Ganoza MC: **Molecular characterization of the prokaryotic efp gene product involved in a peptidyltransferase reaction.** *Biochimie* 1997, **79:**7-11.

30. Aoki H, Adams SL, Turner MA, Ganoza MC: **The gene encoding the elongation factor P protein is essential for viability and is required for protein synthesis.** *J Biol Chem* 1997, **19:**32254-32259.

31. Navarre WW, Zou SB, Roy H, Xie JL, Savchenko A, Singer A, Edvokimova E, Prost LR, Kumar R, Ibba M, Fang FC: **PoxA, YjeK and Elongation Factor P Coordinately Modulate Virulence and Drug Resistance in *Salmonella entérica*.** *Mol Cell* 2010, **30:**209-221.

32. Pancholi V, Fischetti VA: **Alpha-enolase, a novel strong plasmin (ogen) binding protein on the surface of pathogenic streptococci.** *J Biol Chem* 1998, **273:**14503-14515.

33. Chitlaru T, Gat O, Grosfeld H, Inbar I, Gozlan Y, Shafferman A: **Identification of in vivo-expressed immunogenic proteins by serological proteome analysis of the *Bacillus anthracis* secretome.** *Infect. Immun* 2007, **6:**2841-2852.

34. Agarwal S, Kulshreshtha P, Bambah MD, Bhatnagar R: **α-Enolase binds to human plasminogen on the surface of *Bacillus anthracis*.** *Biochim Biophys Acta* 2008, **1784**:986-994.

35. Nelson K, Whittam TS, Selander RK: **Nucleotide polymorphism and evolution in the glyceraldehyde-3- phosphate dehydrogenase gene (gapA) in natural populations of *Salmonella* and *Escherichia coli*.** *Proc Nati Acad Sci USA* 1991, **88:**6667-6671.

36. Madureira P, Baptista M, Vieira M, Magalhães V, Camelo A, Oliveira L, Ribeiro A, Tavares D, Trieu-Cuot P, Vilanova M, Ferreira P: ***Streptococcus agalactiae* GAPDH is a virulence-associated immunomodulatory protein.** *J. Immunol* 2007, **178:**1379-1387.

37. Davidson AL, Dassa E, Orelle C, Chen J: **Structure, Function, and Evolution of Bacterial ATP-Binding Cassette Systems.** *Microbiol Mol Biol Rev* 2008, **72:**317-364.

38. Henderson DP, Payne SM: ***Vibrio cholerae* Iron Transport Systems: Roles of Heme and Siderophore Iron Transport in Virulence and Identification of a Gene Associated with Multiple Iron Transport Systems.** *Infection and immunity* 1994, **62:**5120-5125.

39. Hewett-Emmett D, Tashian RE: **Functional diversity, conservation, and convergence in the evolution of the alpha-,beta-, and gamma-carbonic anhydrase gene families.** *Mol Phylogenet Evol* 1996, **5:**50-77.

40. Burghout P, Cron LE, Gradstedt H, Quintero B, Simonetti E, Bijlsma JJ, Bootsma HJ, Hermans PW: **Carbonic anhydrase is essential for Streptococcus pneumoniae growth in environmental ambient air.** *J Bacteriol* 2010, **192:**4054-4062.

41. Supuran CT: **Bacterial carbonic anhydrases as drug targets: toward novel antibiotics.** *Front Pharmacol* 2011, **2:**34.

42. Poyart C, Berche P, Trieu-Cuot P: **Characterization of superoxide dismutase genes from Gram-positive bacteria by polymerase chain reaction using degenerate primers.** *FEMS Microbiol Lett* 1995, **131:**41-45.

43. Sanders JW, Leenhouts KJ, Haandrikman AJ, Venema G, Kok J: **Stress Response in *Lactococcus lactis*: Cloning, Expression Analysis and Mutation of the Lactococcal Superoxide Dismutase Gene.** *J bacterial* 1995, **177:**5254-5260.

44. Peppoloni S, Posteraro B, Colombari B, Manca L, Hartke A, Giard JC, Sanguinetti M, Fadda G, Blasi E: **Role of the (Mn)superoxide dismutase of Enterococcus faecalis in the in vitro interaction with microglia.** *Microbiology* 2011, **157:**1816-1822.

45. Clements MO, Watson SP, Foster SJ: **Characterization of the major superoxide dismutase of *Staphylococcus aureus* and its role in starvation survival, stress resistance, and pathogenicity.** *J Bacteriol* 1999, **181:**3898-3903.

46. Bruno-Bárcena JM, Azcárate-Peril MA, Hassan HM: **Role of antioxidant enzymes in bacterial resistance to organic acids.** *Appl Environ Microbiol* 2010, **76:**2747-2753.

47. Moura-Costa LF, Paule BJA, Freire SM, Nascimento I, Schaer R, Regis LF, Vale VLC, Matos DP, Bahia RC, Carminati R, Meyer R: **Meio sintético quimicamente definido para o cultivo de *Corynebacterium pseudotuberculosis*.** *Rev Bras Saúde Prod An* 2002, **3:**1-9.

48. Simpson RJ: *Proteins and proteomics: A laboratory manual.* New York: Cold Spring Harbor Laboratory Press, 2003.

49. Havlis J, Thomas H, Sebela M, Shevchenko A: **Fast-response proteomics by accelerated in-gel digestion of proteins.** *Analyti Chemist* 2003, **75:**1300-1306.

50. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S: **Prediction of twin-arginine signal peptides.** *BMC Bioinformatics* 2005, **6:**167.

Figures

Figure 1: 2-DE of extracellular proteins of strain 1002 stained with colloidal Coomassie. Electrophoretic profile using 3-10 NL strips. Spots with numbers were identified by MS. Spots with blue circle indicate protein spot taken as exclusive of strain 1002.

Figure 2: 2-DE of extracellular proteins of strain C231 stained with colloidal Coomassie. Electrophoretic profile using 3-10 NL strips. Spots with numbers were identified by MS. Spots with blue circle indicate protein spot taken as exclusive of strain C231.

Figure 3: Distribution of the identification of extracellular proteins of *C. pseudotuberculosis* by different approaches. In blue, identified proteins detected by both techniques (2DE/MALDI-TOF/TOF + TPP-LC-MS^E); in red, proteins identified by 2DE/MALDI-TOF/TOF technique and in green, proteins identified by Pacheco et al. (2011) (TPP-LC-MS^E).

Tables

Table 1: List of the 11 proteins not identified by TPP/LCMS^E.

Protein	ID (NCBI)^P of strains		Gel identification	
	1002	C231	1002	C231
Chaperone GroEL	ADL21673	ADL11255	Yes	Yes
Chaperone protein DnaK	ADL21757	ADL11343	Yes	Yes
Hypothetical protein	ADL21702	ADL11289	Yes	Yes
Hypothetical protein	ADL21703	ADL11290	Yes	Yes
Hypothetical protein	ADL21704	ADL11291	Yes	Yes
Elongation factor P	ADL21021	ADL10611	Yes	Yes
Enolase	ADL20605	ADL10195	Yes	No
Glyceraldehyde-3-phosphate dehydrogenase	ADL20991	ADL10581	Yes	No
ABC-type transporter	ADL20391	ADL09988	Yes	No
Carbonic anhydrase	ADL20477	ADL20477	Yes	No
Manganese superoxide dismutase	ADL21849	ADL11437	Yes	No

(P) Access number of proteins (NCBI genome project at 40687 and 40875).

Additional files

Additional file 1: Table S1. List of the extracellular proteins common to 1002 and C231 *C. pseudotuberculosis* strains.

Additional file 2: Table S2. List of the extracellular unique proteins 1002 *C. pseudotuberculosis* strain.

Additional file 3: Table S3. List of the extracellular unique proteins to C231 *C. pseudotuberculosis* strain.

Additional file 4: Table S4. List of 104 extracellular proteins of *C. pseudotuberculosis* detected by both approaches (2-DE-MALDI-TOF/TOF and TPP-LC/MS^E).

3.2.4 Análise sorológica preliminar do secretoma da *C. pseudotuberculosis*

Neste artigo científico, as proteínas identificadas como secretadas, referidas na seção 3.2.2, foram resolvidas em gel de 2-D. Na sequência, foi feito *western blotting* com soro de animais positivos para a linfadenite caseosa. Seis proteínas foram as mais imunorreativas. Estes estudos nos permitem sugerir que estas proteínas são alvos preferenciais para serem utilizadas como alvos para o desenvolvimento de novas vacinas e diagnóstico. Essas proteínas secretadas são um critério adicional que pode ser utilizado para a escolha de candidatos vacinais.

A Tabela 2 relaciona estas seis proteínas mais imunorreativas com os resultados apresentados nos artigos científicos da seções 3.2.1 (*Análise in silico do panexoproteoma de cinco linhagens de C. pseudotuberculosis*), 3.2.2 (*Uma abordagem combinada para análise comparativa do proteoma exportado da C. pseudotuberculosis*) e 3.2.2 (*Identificação de 11 novas proteínas exportadas da C. pseudotuberculosis utilizando análise comparativa do secretoma*).

locus_tag	**Local subcelular predito *in silico***	**Identificador do *GenBank***	***Pan locus***	**Idenficador do *NCBI***	**Cobertura**	**Presente no exoproteoma central**
Cp1002_0237	SECRETED	ADL20140	plcpsec012	gi|302329946	5x	Sim
Cp1002_0681	SECRETED	ADL20574	plcpsec029	gi|302330380	5x	Sim
Cp1002_0766	SECRETED	ADL20656	plcpsec033	gi|302330462	5x	Não
Cp1002_1416	SECRETED	ADL21293	plcpsec048	gi|302331099	5x	Sim
Cp1002_1887	PSE C	ADL21747	plcppse143	gi|302331553	5x	Sim
CpC231_0642	PSE C	ADL10125	plcppse060	gi|302205783	5x	Não

Tabela 2: Seis proteínas mais imunorreativas e presentes no core panexoproteoma predito *in silico* da *C. pseudotuberculosis*.

Na Tabela 2, as células da coluna denominada "*Presente no exoproteoma central*" marcadas com "*Sim*" significam que a proteína foi encontrada nos dois experimentos do exoproteoma de *C. pseudotuberculosis;* enquanto as células marcadas com "*Não*" subentendem que as proteínas foram encontradas no exoproteoma variante de uma das linhagens. A proteína identificada pelo *locus_tag* CpC231_0642, referente à linhagem C231, possui homologia com a proteína com o *locus_tag* Cp1002_0643, da linhagem 1002, porém o identificador do *NCBI* presente no artigo a seguir identifica unicamente a proteína da linhagem C231.

JOURNAL OF INTEGRATED OMICS
A METHODOLOGICAL JOURNAL
HTTP://WWW.JIOMICS.COM

PRELIMINARY COMMUNICATION

Preliminary serological secretome analysis of *Corynebacterium pseudotuberculosis*

Núbia Seyffert[1,2], Luis G.C. Pacheco[1,3,4], Wanderson M. Silva[1], Thiago L.P. Castro[1], Agenor V. Santos[3,5], Anderson Santos[1], John A. McCulloch[1,5], Maira R. Rodrigues[1], Simone G. Santos[2], Luiz M. Farias[2], Maria A.R. Carvalho[2], Adriano M.C. Pimenta[3], Artur Silva[5], Roberto Meyer[4], Anderson Miyoshi[1], Vasco Azevedo[1*].

[1]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Belo Horizonte 31.270-901, Brazil; [2]Departamento de Microbiologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Belo Horizonte 31.270-901, Brazil; [3]Departamento de Bioquímica, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Belo Horizonte 31.270-901, Brazil; [4]Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador 40.110-902, Brazil; [5]Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém 66.075-900, Brazil.

Received: 02 November 2010 **Accepted:** 19 June 2011 **Available Online**: 20 June 2011

Abstract

Caseous lymphadenitis (CLA) is a chronic disease affecting small ruminants that is caused by *Corynebacterium pseudotuberculosis* and is responsible for significant economic losses. Various *C. pseudotuberculosis* secreted proteins are known to react with sera from infected goats. Mapping of the secretome would help us understand the pathogenesis of CLA. We identified six immunoreactive secreted proteins of *C. pseudotuberculosis* by 2D-Western blotting, using sera from goats with CLA, and characterized them by mass spectrometry. This preliminary information will give support to future studies aimed at the development of efficient vaccines and diagnostic kits.

Keywords: *C. pseudotuberculosis*, Caseous lymphadenitis, Secretome.

1. Introduction

Caseous lymphadenitis (CLA) is a chronic disease affecting small ruminants; it is caused by infection with *Corynebacterium pseudotuberculosis* and is responsible for significant worldwide economic losses due to decreases in both the productivity and the reproductive performance of infected animals [1]. The lack of efficient immunoprophylaxis against CLA results in ineffective management of this disease in animals, facilitating its dissemination [2]. Efficient vaccines against CLA and diagnostic kits for this disease are still not available, in part due to a lack of sufficient information concerning newly-characterized *C. pseudotuberculosis* virulence determinants [3].

Only a few genes and their products have been identified as factors that contribute to the virulence of *C. pseudotuberculosis*, including phospholipase D (PLD) [4,5], the *fagABC* operon involved in iron acquisition by the cell [6] and the protease CP40 [7]. Chaplin et al. [8] developed a DNA vaccine encoding PLD to immunize sheep, but they achieved only partial protection against challenge with *C. pseudotuberculosis.* Similar results were obtained when sheep were immunized with a formalin-inactivated subunit vaccine [9]. CP40 protease has been reported as a possible candidate for the development of vaccines, based on Western blot analysis with serum samples from sheep experimentally infected with *C. pseudotuberculosis* [10].

To date, the search for immunogenic proteins has been carried out in a non-exhaustive manner, using various extraction and separation techniques [11,12]. A comprehensive analysis of the entire set of proteins expressed by *C. pseudotuberculosis* strains is needed in order to identify the best

***Corresponding author:** Dr. Vasco Azevedo. Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Belo Horizonte 31.270-901, Brazil. E-mail Address: vasco@icb.ufmg.br.

IN PRESS | JIOMICS 2011

candidate proteins for immunoprophylactic or diagnostic applications. Bacterial secreted proteins have a various biological functions, ranging from toxicity to more subtle alterations of the host cell for the benefit of the invader; they are an important part of the pathogenic process [13].

There has not been much conclusive research concerning pathogenesis or even the immune response against *C. pseudotuberculosis* infection, compared to research on other important veterinary pathogens, such as *Mycobacterium tuberculosis* [14]. Mapping of the *C. pseudotuberculosis* secretome, followed by characterization of expressed proteins and assessment of their immunogenic potential would be ideal for shedding light on the pathogenesis of specific strains and the host immune response that they provoke, paving the way for the development of more efficient vaccines and diagnostic kits. We examined antigenic proteins of the secretome of *C. pseudotuberculosis* strain 1002 cultivated in chemically defined medium (CDM) using serological proteome analysis (SERPA) [15].

2. Material and Methods

2.1 *Bacterial strain and growth conditions*

Corynebacterium pseudotuberculosis strain 1002, originally isolated from an infected goat in Brazil [16], was routinely maintained in Brain Heart Infusion broth (BHI) and characterized by biochemical and molecular methods, as previously described [17]. For SERPA, bacteria were grown at 37ºC under agitation (100 rpm), for 24 h in 1 L of chemically defined medium (CDM), until reaching the exponential growth phase (OD600nm = 1.3). The CDM contained 0.067M of phosphate buffer, 0.05% (v/v) Tween 80 (Sigma), 4% (v/v) 100X minimal Essential Medium (MEM) Vitamin Solution (Invitrogen), 1% (v/v) of 50X MEM Amino Acids Solution (Invitrogen), 1% (v/v) 100X MEM Non-Essential Amino Acids Solution (Invitrogen) and 1.2% (w/v) filter-sterilized glucose, as previously described [18].

2.2 *Extraction of secreted proteins*

Corynebacterium pseudotuberculosis exoproteins were obtained according to a previously described three-phase partitioning (TPP) protocol [19]. Briefly, bacterial cells were separated from the supernatant by centrifugation at 4,000 rpm for 20 min at 4ºC. The supernatant was filtered through a 0.22 μm membrane (filter) and 30% (w/v) ammonium sulphate was added. The pH was adjusted to 4.0 and *n*-butanol was added at a ratio of 1:1, and the sample was vortexed. After 1h of incubation at room temperature, the precipitate at the interface was collected and re-suspended in 1 mL of 20 mM Tris-HCl buffer pH 7.4 with 10 μL of protease inhibitor (GE).

2.3 *2D-PAGE-Western blot*

Two-dimensional electrophoretic separation was carried out, as previously described [20], with minor modifications. Secreted proteins (150 μg) were dissolved in 2-DE sample buffer (8 M urea, 2 M thiourea, 4% CHAPS, 1% (v/v) carrier ampholyte pH 3.0-5.6, 80 mM dithiothreitol (DTT), 40 mM Tris-base and bromophenol blue. The mixture was used for overnight rehydration of 11 cm immobilized pH gradient (IPG) strips (Immobiline DryStripTM Gels, pH 3.0-5.6 NL [GE Healthcare]). Isoelectric focusing (IEF) was carried out at room temperature for 24.5 h (maximum voltage of 3,500 V and maximum current of 50 μA). After equilibration for 15 min in a 50 mM Tris HCl (pH 8.8) buffer solution containing 6 M urea, 2% (w/v) SDS, 30% (v/v) glycerol and 0.001% (v/v) bromophenol blue and 10 mg/mL DTT, the strips were equilibrated for 15 min in the same solution, except that the DTT was replaced by 25 mg/mL iodoacetamide. The proteins were resolved in 2D electrophoresis in 12% polyacrilamyde gels under denaturing conditions, using a Protean IIxi system (Biorad). Protein spots were visualized by staining with Coomassie blue G-250 (GE Healthcare). For each protein sample, three 2D gels were stained to visualize proteins and six 2D gels were electroblotted onto polyvinylidene difluoride membranes (*Owl* system) for 1 h, with an electric current of 0.4 A. The membranes were blocked overnight at 4º C in 5% non-fat milk in phosphate buffered saline pH 7.5 with 0.05% Tween 20 (PBS-T). The membranes were then incubated at room temperature for 1 h in PBS-T with sera (at a proportion of 1:100 v/v PBS-T:serum) obtained from animals either infected or uninfected with *C. pseudotuberculosis*. The membranes were then washed with PBS-T three times for 5 min and incubated for 1h with an anti-goat IgG peroxidase antibody produced in rabbits (Sigma), diluted 1:1000 in PBS-T solution. Antibody-tagged protein spots were detected with DAB peroxidase substrate solution.

2.4 *Identification of immunoreactive proteins*

Membranes were digitally scanned and immunoreactive proteins matched to 2D gel images of the samples were identified using the Melanie software (GeneBio). All spots reactive in 2D-Western blots were selected from an analogous 2D stained gel and manually excised. The excised gel fragments were incubated overnight with 25mM bicarbonate/50% acetonitrile (ACN) solution until completely destained. After drying, gel fragments were placed in 50 mM ammonium bicarbonate solution with 20ng/μL sequencing-grade modified trypsin (Promega Biosciences, CA, USA). Digestion was run at 37º C overnight. The peptides were extracted using 5% formic acid/50% acetonitrile solution, concentrated in a SpeedVac (Savant, USA) to a volume of about 10 μL, desalted using ZipTip® C18 plates (C18 resin, P10; Millipore Corporation, Bedford, MA, USA) and eluted with 0.1% trifluoroacetic acid solution containing 50% ACN. The sample extract was mixed at a 1:1 ratio with matrix (10 mg/mL recrystallized α-

IN PRESS | JIOMICS 2011

cyano-4-hydroxycinnamic acid) to a final volume of 1 μL and then spotted onto an MTP AnchorChip™ 600/384 (Bruker Daltonics) for matrix-assisted laser desorption/ ionization time-of-flight tandem mass spectrometry (MALDI-TOF-MS/MS) (LIFT technology, Autoflex III™; BrukerDaltonics, Billerica, USA) analysis. Ionization was performed in MS/MS (PSD-LIFT technology) by irradiation of a nitrogen laser (337 nm) operating at 50 Hz. Data were acquired at a maximum accelerating potential of 25 kV in the positive and reflector modes. Trypsin and keratin contamination peaks were excluded from the mass spectra and MS/MS results were used to search the *C. pseudotuberculosis* 1002 (Gen Bank: CP001809.1) protein database using MASCOT software (http://www.matrixscience.com/).

3. Results and Discussion

Bacterial growth within a host resulting in infection is a consequence of colonization, adherence, invasion, evasion of the immune response and toxigenesis caused by the bacterial cell. This feat can be accomplished by a bacterial strain through temporal expression of a panoply of virulence genes (the virulon), in response to appropriate environmental stimuli. Characterization of when, which and what amounts of virulence factors are expressed in response to certain stimuli is necessary for understanding the pathogenesis of bacterial species [13]. The dynamics of the immune response to infection can only be fully understood if we characterize the bacterial proteins responsible for eliciting the immune response. The advent of genomics has made this approach feasible, since information concerning an immunogenic protein can be traced back to the genome, and thence to the regulon that is involved. Several species of Actinobacteria have been the subject of proteomic analysis, including *Mycobacterium avium*, *Mycobacterium tuberculosis*, *Rhodococcus equi* and *Corynebacterium diphtheriae*, yielding insight into the relationships between the host and the bacterial parasite [21,22,23,24,25]. The secretome of *C. pseudotuberculosis* was first studied by Braithwaite et al. [12], who extracted proteins from a culture supernatant with ammonium sulphate; they found seven proteins with molecular weights between 14 and 64 kDa, five of which reacted with sera obtained from goats infected with *C. pseudotuberculosis*. A later study described the reaction of a pool of sera obtained from goats suffering from CLA against 11 *C. pseudotuberculosis* secreted proteins with molecular weights ranging from 24 to 125 kDa [19]; these proteins induced an increase in the serum concentration of IFN-γ in goats infected with this bacterium [26]. We made a follow up of that study. The excreted-secreted antigens of *C. pseudotuberculosis* were obtained by culturing the 1002 strain in CDM [18], with subsequent extraction of secreted proteins by TPP [19], and a 2D-PAGE-Western blot. Twenty-three immunoreactive spots were detected using sera obtained from animals with CLA; due to time and budget constraints, only six of these proteins were identified by MALDI-TOF-MS/MS (Figure 1 and Table 1). Immunoproteomic methods, such as SERPA [15], have been used to identify biomarkers and target antigens for developing diagnostic kits based on antibody/antigen detection, as well as to develop vaccines and treatments for various infectious diseases. Due to the unfeasibility of targeting many proteins simultaneously, our objective was to

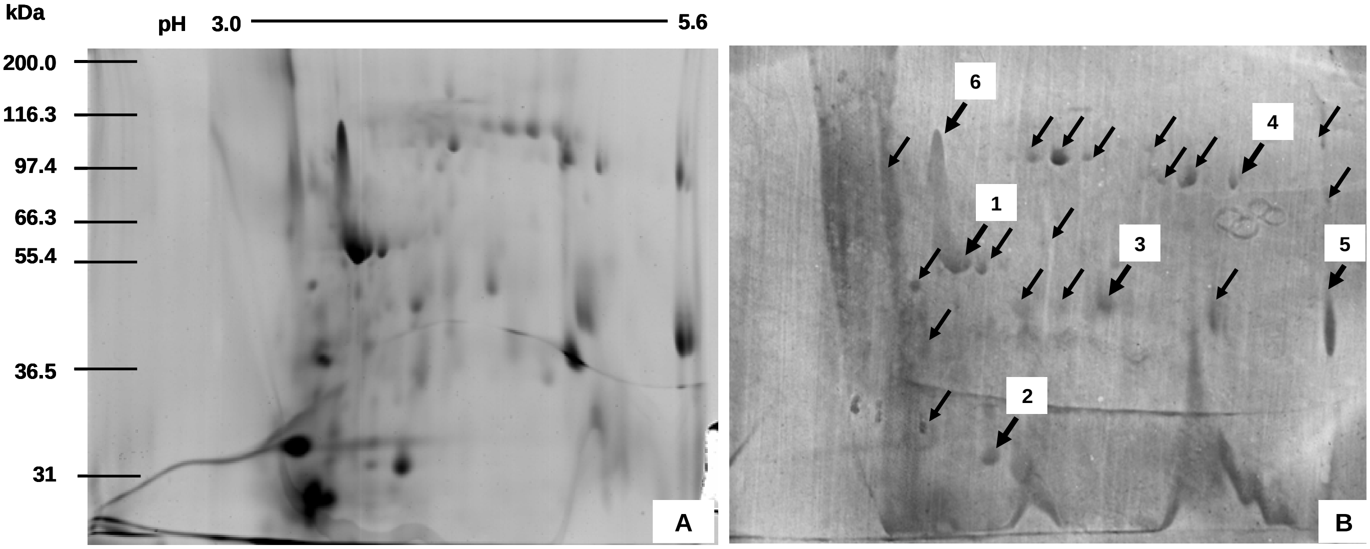
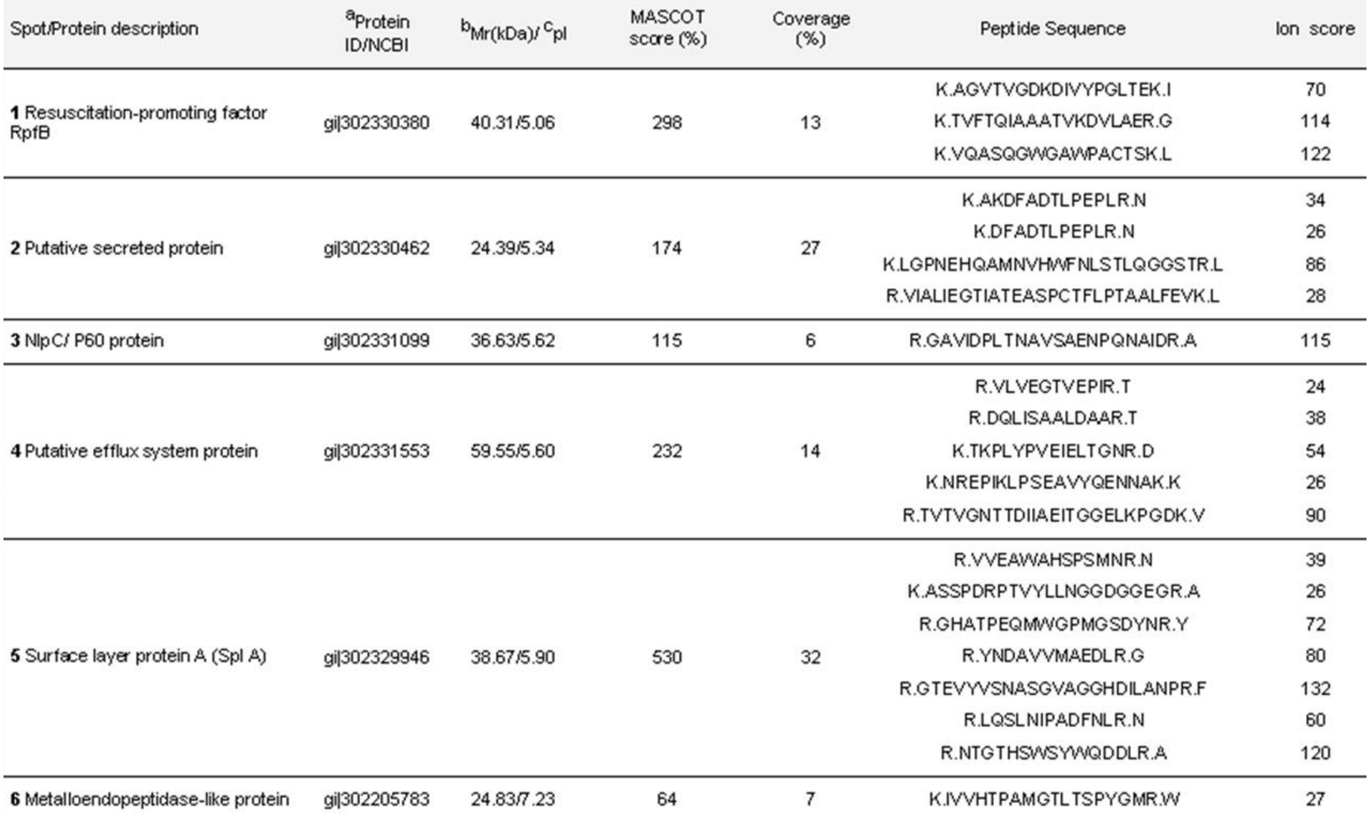


Figure 1. Serological proteome analysis of secreted proteins of *Corynebacterium pseudotuberculosis* using serum from infected goats. A) 2D-PAGE with 150μg of sample for analysis of *C. pseudotuberculosis* secreted proteins. Spots were detected in the gels stained with Coomassie G-250. B) 2D-PAGE-Western blot analysis of *C. pseudotuberculosis* secreted proteins with 150μg sample. The black arrows indicate 23 immunoreactive spots detected by anti-goat IgG peroxidase antibody produced in rabbits. Numbers correspond to the proteins identified in Table 1.

IN PRESS | JIOMICS 2011

Table 1. Antigenic proteins of *Corynebacterium pseudotuberculosis* identified by MALDI-TOF MS/MS. [a] Accession numbers in Entrez Protein (NCBI Genome CP001809.1). [b] Theoretical molecular weights (Mr). [c] Theoretical isoelectric points (pI)



Spot/Protein description	[a]Protein ID/NCBI	[b]Mr(kDa)/ [c]pI	MASCOT score (%)	Coverage (%)	Peptide Sequence	Ion score
1 Resuscitation-promoting factor RpfB	gi|302330380	40.31/5.06	298	13	K.AGVTVGDKDIVYPGLTEK.I	70
					K.TVFTQIAAATVKDVLAER.G	114
					K.VQASQGWGAWPACTSK.L	122
2 Putative secreted protein	gi|302330462	24.39/5.34	174	27	K.AKDFADTLPEPLR.N	34
					K.DFADTLPEPLR.N	26
					K.LGPNEHQAMNVHWFNLSTLQGGSTR.L	86
					R.VIALIEGTIATEASPCTFLPTAALFEVK.L	28
3 NlpC/ P60 protein	gi|302331099	36.63/5.62	115	6	R.GAVIDPLTNAVSAENPQNAIDR.A	115
4 Putative efflux system protein	gi|302331553	59.55/5.60	232	14	R.VLVEGTVEPIR.T	24
					R.DQLISAALDAAR.T	38
					K.TKPLYPVEIELTGNR.D	54
					K.NREPIKLPSEAVYQENNAK.K	26
					R.TVTVGNTTDIIAEITGGELKPGDK.V	90
5 Surface layer protein A (Spl A)	gi|302329946	38.67/5.90	530	32	R.VVEAWAHSPSMNR.N	39
					K.ASSPDRPTVYLLNGGDGGEGR.A	26
					R.GHATPEQMWGPMGSDYNR.Y	72
					R.YNDAVVMAEDLR.G	80
					R.GTEVYVSNASGVAGGHDILANPR.F	132
					R.LQSLNIPADFNLR.N	60
					R.NTGTHSWSYWQDDLR.A	120
6 Metalloendopeptidase-like protein	gi|302205783	24.83/7.23	64	7	K.IVVHTPAMGTLTSPYGMR.W	27

identify polypeptide chains that display immunoreactivity, thus narrowing the number of targets for further experiments. Six of the spots that displayed immunoreactivity by SERPA (Figure 1) and were identified by MALDI-TOF-MS/MS (Table 1) were consistent with five proteins identified by data-independent MS acquisition (LC-MSE) and six proteins predicted *in silico* by SurfG *plus* studies performed by our team to characterize the total *C. pseudotuberculosis* exoproteome, independent of the 2D-PAGE Western blot analysis [27]. Four immunoreactive secreted proteins that we identified, namely, resuscitation-promoting factor B (RpfB), Nlp/P60 protein, putative efflux system protein and surface layer protein A (SlpA), have previously been reported from other bacterial species. Rpf homologues are widespread throughout the Actinobacteria [28] and have the ability to stimulate the growth of dormant mycobacteria, especially *Mycobacterium tuberculosis* [29]. NlpC/P60 belongs to the peptidase family and plays a role in turnover of the bacterial cell wall [30]. Previously characterized NlpC/P60 proteins include *Listeria monocytogenes* secreted autolysin P60 [31], *Bacillus subtilis* autolysins [32] and *Escherichia coli* membrane-associated lipoprotein [33]. The bacterial efflux system is composed of proteins that act as a continuous channel for the extrusion of substrates within the cell envelope into the external environment. This mechanism may be involved in various transport functions, including efflux of toxins, metabolites and drugs [34,35]. Surface layer protein A (Spl A) has been described from several bacterial species, in which it has various roles, including nutrient uptake, colonization, antiphagocytosis and exclusion of noxious substances [36]. Putative secreted protein (spot 2) was found to be similar to protein sequences of other Actinobacteria, based on protein BLAST/NCBI (National Center for Biotechnology Information), but it has not been described. Another secreted protein, metalloendopeptidase-like protein belongs to the family of metalloproteases [37], which are common in pathogenic bacteria, including *Pseudomonas aeruginosa* (degradation of host connective tissues) [38], *Clostridium spp.* (neurotoxin activity) [39], *Bacillus anthracis* (lethal toxin) [40] and *Listeria monocytogenes* (enzyme maturation) [41]. Secreted proteins of similar molecular weight have been found in other studies [12,19], through extraction of proteins by other methods and resolution by unidimensional electrophoresis [12,14], but they were not identified as being the CP40 protease [7]. Our preliminary results indicate that SERPA coupled with mass spectrometry analysis is a useful strategy for the identification of these six antigens. Studies are underway to develop a protocol for the detection of the other spots that remained unidentified. These findings may help identify proteins that can induce protective immunity or elicit immune responses with diagnostic value for CLA.

IN PRESS | JIOMICS 2011

Acknowledgements

This work was supported by CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brasil), CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil) and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais, Brasil).

References

1. M.C.Fontaine, G.J.Baird, Small Rumin. Res. 76 (2008) 42–48.DOI:10.1016/j.smallrumres.2007.12.025
2. N. Seyffert, A.S. Guimarães, L.G.C. Pacheco, R.W. Portela, B.L. Bastos, F.A. Dorella, M.B. Heinemann, A.P. Lage, A.M.G. Gouveia, R. Meyer, A. Miyoshi, V. Azevedo, Res. Vet. Sci. 88 (2010) 50-55. DOI:10.1016/j.rvsc.2009.07.002
3. F.A.Dorella, L.G. Pacheco, N. Seyffert, R.W. Portela, R. Meyer, A. Miyoshi, V. Azevedo, Expert. Rev. Vaccines 8 (2009) 205-213. DOI:10.1586/14760584.8.2.205
4. B.A. Lipsky, A.C. Goldberger, L.S. Tompkins, J.J. Plorde, Clin. Infect. Dis. 4 (1982) 1220–1235. DOI:10.1093/clinids/4.6.1220
5. A.L. Hodgson, K. Carter, M. Tachedjian, J. Krywult, L.A. Corner, M. McColl, A. Cameron, Vaccine 17 (1999) 802–808. DOI:10.1016/S0264-410X(98)00264-3
6. S.J. Billington, P.A. Esmay, J.G. Songer, B.H. Jost, FEMS Microbiol. Lett. 208 (2002) 41-45. DOI: 10.1111/j.1574-6968.2002.tb11058.x
7. M.J. Wilson, M.R. Brandon, J. Walker, Infect. Immun. 63 (1995) 206-211.
8. P. J. Chaplin, R. De Rose, J. S. Boyle, P. McWaters, J. Kelly, J. M. Tennent, A. M. Lew, J. P. Scheerlinck, Infect. Immun. 67 (1999) 6434-6438.
9. K. Stanford, K.A. Brogden, L.A. McClelland, G.C. Kozub, F. Audibert, Can. J. Vet. Res. 62(1998) 38–43.
10. J. Walker, H.J. Jackson, D.G. Eggleton, E.N.T. Meeusen, M.J. Wilson, M.R. Brandon, Infect. Immun. 62 (1994) 2562–2567.
11. C.A. Muckle, P.I. Menzies, Y. Li, Y.T. Hwang, M. van Wesenbeeck, Vet. Microbiol. 30 (1992) 47-58.
12. C.E. Braithwaite, E.E. Smith, J.G. Songer, A.H. Reine, Vet. Microbiol. 38 (1993) 59-70.
13. K. Wooldridge, Bacterial secreted proteins secretory mechanisms and role in pathogenics, Norfolk, UK, 2009.
14. I. Smith, Clin. Microbiol, Rev. 16 (2003) 463–496. DOI:10.1128/CMR.16.3.463-496.2003.
15. N. Falisse-Poirrier, V. Ruelle, B. Elmoualij, D. Zorzi, O. Pierard, E. Heinen, E. De Pauw, W. Zorzi, J. Microbiol. Methods. 67 (2006) 593-596. DOI:10.1016/j.mimet.2006.05.002.
16. R. Meyer, R. Carminati, R. Bahia, V. Vale, S. Viegas, T. Martinez, I. Nascimento, R. Schaer, J. Silva, M. Ribeiro, L. Régis, B. Paule, S. Freire, J. Med. Biol. Sci. 1(2002), 42-48.
17. L.G.C. Pacheco, R.R. Pena, T.L.P. Castro, F.A. Dorella, R.C. Bahia, R. Carminati, M.N.L. Frota, S.C. Oliveira, R. Meyer, F.S.F. Alves, A. Miyoshi, V. Azevedo, J. Med. Microbiol, 56 (2007) 480-486. DOI: 10.1099/jmm.0.46997-0.
18. L.F. Moura-Costa, B.J.A. Paule, V. Azevedo, S. M. Freire, I. Nascimento, R. Schaer, L.F. Regis, V.L.C. Vale, D.P. Matos, R.C. Bahia, R. Carminati, R. Meyer, Rev. Bras. Saúde e Produção Animal, 3(2002) 1-9.
19. B.J.A. Paule, R. Meyer, L.F. Moura-Costa, R.C. Bahia, R. Carminati, L.F. Regis, V.L.C. Vale, S.M. Freire, I. Nascimento, R. Schaer, V. Azevedo, Protein. Expr. Purif. 34 (2004) 311-316. DOI: 10.1016/j.pep.2003.12.003.
20. R. J. Simpson. Cold Spring Harbor:CSHL Press, 2003
21. V. Hughes, J.P. Bannantine, S. Denham, S. Smith, A. Garcia-Sanchez, J. Sales, M.L. Paustian, K. MClean, K. Stevenson, Clin. Vaccine Immunol. 15(2008) 1824-1833. DOI: 10.1128/CVI.00099-08.
22. P.R. Jungblut, U.E. Schaible, H.J. Mollenkopf, U. Zimny-Arndt, B. Raupach, J. Mattow, P. Halada, S. Lamer, K. Hagens, S.H. Kaufmann, Mol. Microbiol. 33(1999) 1103-17. DOI: 10.1046/j.1365-2958.1999.01549.x.
23. S. Sinha, K. Kosalai, S. Arora, A. Namane, P. Sharma, A.N. Gaikwad, P. Brodin, S.T. Cole, Microbiol. 151 (2005) 2411-2419. DOI:10.1099/mic.0.27799-0.
24. C. Barbey, A. Budin-Verneuil, S. Cauchard, A. Hartke, C. Laugier, V. Pichereau, S. Petry, Vet. Microbiol. 135 (2009) 334-45. DOI:10.1016/j.vetmic.2008.09.086.
25. N. Hansmeier, T. Chao, J. Kalinowski, A. Pühler, A. Tauch, Proteomics 6 (2006) 2465–2476. DOI: 10.1002/pmic.200500360.
26. R. Meyer, L. Regis, V. Vale, B. Paule, R. Carminati, R. Bahia, L. Moura Costa, R. Schaer, I. Nascimento, S. Freire, Vet. Immunol. and Immunopathol. 107 (2005) 249-54. DOI:10.1016/j.vetimm.2005.05.002.
27. L.G. Pacheco, S.E. Slade , N. Seyffert, A.R. Santos, T.L. Castro, W.M. Silva, A.V. Santos, S.G. Santos, L.M. Farias, M.A. Carvalho, A.M. Pimenta, R. Meyer, A. Silva, J.H. Scrivens, S.C. Oliveira, A. Miyoshi, C.G. Dowson., V. Azevedo, BMC Microbiol. 2011 17;11(1):12. DOI:10.1186/1471-2180-11-12.
28. A. Ravagnani, C. L Finan, M. Young, BMC Gen. 6 (2005) 39. DOI:10.1186/1471-2164-6-39.
29. E.C. Hett, M.C. Chao, L.L. Deng, E.J. Rubin, PLoS Pathog. 4 (2008):e1000001. DOI: 10.1371/ journal.ppat.1000001.
30. V. Anantharaman, L. Aravind, Genome Biol. 4 (2003) R11. DOI:10.1186/gb-2003-4-2-r11.
31. M. Kuhn, W. Goebel, Infect. Immun. 57 (1989) 55-61.
32. T.J. Smith, S.A. Blackman, S.J. Foster, Microbiol. 146 (2000) 249–262.
33. J.M. Aramini, P. Rossi, Y.J. Huang, L. Zhao, M. Jiang, M. Maglaqui, R. Xiao, J. Locke, R. Nair, B. Rost, T.B. Acton, M. Inouye, G.T. Montelione, Biochem. 47 (2008) 9715-9717. DOI: 10.1021/bi8010779.
34. M.H. Jr. Saier, I. T. Paulsen, Semin. Cell Dev. Biol. 12 (2001) 205-213. DOI:10.1006/scdb.2000.0246.
35. T.T. Tseng, K.S. Gratwick, J. Kollman, D. Park, D.H. Nies, A. Goffeau, M.H. Jr. Saier, J. Mol. Microbiol. Biotechnol. 1(1999) 107-125.
36. M. Sara, U.B. Sleytr, J. Bacteriol. 169 (1987) 4092-4098.
37. S. Miyoshi, S. Shinoda, Microbes Infect. 2, (2000) 91–98. DOI:10.1016/S1286-4579(00)00280-X .
38. J.C. Olson, D.E. Ohman, J Bacteriol. 174 (1992) 4140–4147.
39. F. Tonello, S. Morante, O. Rossetto, G. Schiavo, C. Montecucco, Adv Exp Med Biol. 389 (1996) 251–260.
40. P. J. Hanna, Appl Microbiol. 87 (1999) 285–287.
41. J. Raveneau, C. Geoffroy, J. L. Beretti, J. L. Gaillard, J. E. Alouf, P. Berche, Infect Immun. 60 (1992) 916–921.

3.2.5 Densidade de epitopos na porção madura de proteínas exportadas: um filtro adicional para vacinologia reversa

O uso da imunoinformática pode ser uma estratégia para ranquear e reduzir a quantidade de candidatos oriundos da utilização da seleção de proteínas exportadas de um genoma. Entretanto, o uso da imunoinformática demanda a escolha de um parâmetro de difícil escolha. Determinar qual alelo de MHC para o qual uma proteína será analisada frente à possibilidade de gerar uma resposta imunológica, não é tarefa fácil mesmo quando alelos de interesse estão disponíveis para seleção em um certo programa de análises. Como a linfadenite caseosa (LC) acomete caprinos e ovinos, não existiam alelos de MHC disponíveis como opção de predição para esses animais, por exemplo, nos programas NetMHC (Lundegaard e cols., 2008) e NetCTL (Larsen e cols., 2007), entretanto a *C. pseudotuberculosis*, também infecta o ser humano (Peel e cols., 1997; Mills e cols., 1997; Liu e cols., 2005; Join-Lambert e cols., 2006). A constatação da infecção de humanos pela *C. pseudotuberculosis* serve como um argumento favorável à predição de epitopos no genoma da *C. pseudotuberculosis* a partir de alelos humanos, além de alelos de outros hospedeiros. Para esse propósito criou-se uma estatística denominada *Mature Epitope Density* (MED) que quantifica o potencial de uma proteína no tocante à geração de resposta imune. Essa estatística leva em consideração a quantidade de epitopos preditos na porção madura de uma proteína classificada como exportada. Resultados preliminares dessa análise consideravam como bons candidatos a gerar respostas imunes algumas proteínas secretadas e comprovadamente expostas ao meio extracelular da bactéria patogênica *C. pseudotuberculosis.* Dentre elas, uma proteína única do gênero *Corynebacterium* e experimentalmente comprovada como secretada no genoma de duas linhagens da *C. pseudotuberculosis*, 1002 e C231. Até a presente data, quatro proteínas comprovadamente secretadas nas linhagens 1002 e C231 estão sendo analisadas no tocante à sua capacidade de gerar respostas imunológicas. A escolha desses quatro candidatos ocorreu por serem encontradas no exoproteoma destas duas linhagens de *C. pseudotuberculosis* e terem sido indicadas como as mais promissoras pela análise da estatística MED.

3.2.5.1 Metodologia da estatística MED

O artigo científico a seguir foi submetido e mostra que a estatística MED não é exclusiva da *C. pseudotuberculosis*, sendo valorosa no organismo *Mycobacterium tuberculosis*, linhagem H37Rv. A análise de sensibilidade *versus* falsos positivos (curva de *ROC*) mostrou que a maior parte dos genes com alto valor de MED são fatores de virulência envolvidos na patogenicidade do organismo *M. tuberculosis*. Na sequência, são apresentados resultados da aplicação da estatística MED específicos para cinco genomas da *C. pseudotuberculosis.*

Sequence analysis

Mature Epitope Density – A strategy for detecting exported prokaryotic proteins related to antigenicity and pathogenicity

Anderson R. Santos[1], Vanessa Bastos[1], Eudes Barbosa[1], Jan Baumbach[3], Josch Pauling[3], Artur Silva[2], Anderson Miyoshi[1] and Vasco Azevedo[1,*]

[1]Molecular and Cellular Genetics Laboratory, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

[2]DNA Polymorphism Laboratory, Universidade Federal do Pará, Campus do Guamá, Belém, Pará, Brazil.

[3]Computational Systems Biology group, Max Planck Institute for Informatics, Campus E2.1, 66123 Saarbrücken, Germany.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Current immunological bioinformatics approaches focus on the prediction of allele specific epitopes capable of triggering immunogenic activity. Prediction of MHC class I epitopes is well studied. Various software solutions exist for this purpose. However, currently available tools do not account for the concentration of epitope products in the mature protein product, and they also do not provide information concerning the product's relation to immunogenicity and pathogenicity.

Results: We developed a computational strategy based on measuring the epitope's concentration in the mature protein: Mature Epitope Density (MED). Our method, though simple, is capable of identifying pathogenicity-related proteins. Our on line available software implementation provides a computationally light and reliable analysis of bacterial exoproteins and their pathogenic potential. We evaluated our computational approach by using the *Mycobacterium tuberculosis* (*Mtb*) H37Rv exoproteome as gold standard model. Sixty proteins were investigated, out of 553 *Mtb*'s predicted exoproteins, for experimental evidence concerning immunogenicity or pathogenicity; half out the 60 proteins were highest scored while the other half were lowest scored. Among the half lowest scored proteins, 14 (47%) were confirmed as not related to immunogenicity or pathogenicity, and 22 (73%) of the half highest scored proteins were confirmed as related to pathogenicity, respectively. There was also no experimental evidence of pathogenic contributions for three of the highest MED-scored *Mtb* proteins. Hence, these three proteins are novel putative vaccine and drug targets for *Mtb*.

Availability: A web version of MED is publicly available on line at http://lgcm.icb.ufmg.br/medpipe.

Contact: vasco@icb.ufmg.br

Supplementary Information: Supplementary data are available at Bioinformatics on line.

1 INTRODUCTION

Tuberculosis (TB) has been one of the major causes of morbidity and mortality worldwide for centuries, and control of the spread of *Mycobacterium tuberculosis* (*Mtb*) infection remains a public health priority (Rowland and McShane, 2011). More than 9 million new cases of TB in humans arise every year, resulting in nearly 2 million deaths worldwide (Rylance et al., 2010). Bacille Calmette-Guérin (BCG), the current vaccine for TB, has limitations; even though it is protective against severe childhood TB, it does not satisfactorily prevent the pulmonary disease in adults (Thaiss and Kaufmann, 2010). Effective prophylactic and therapeutic immunization is a key strategy for global epidemic control. Novel TB vaccine candidates include BCG or recombinant BCG (rBCG) strains, which are used in heterologous prime-boost strategies as the prime vaccination. Boosting vaccination can include viral vectors that express immunodominant *Mtb* antigens or fusion proteins of these antigens, combined with adjuvanticity to ensure immunogenicity. Many *Mtb* antigens have been tested as vaccine candidates, such as Ag85B, Ag85A, TB10.4 and ESAT-6 (Thaiss and Kaufmann, 2010), but without success. Consequently, discovering new antigens continues to be a crucial factor in the successful development of vaccines against TB.

Exported proteins are currently the main target for Reverse Vaccinology (RV), due to their essential role in host-pathogen interactions. Examples of this interaction include: (i) adherence to host cells, (ii) invasion of the cell to which there was compliance, (iii) damage to host tissues, (iv) resistance to environmental stress from the defense machinery of the cells and (v) mechanisms for subversion of the host immune response. In general, RV leaves a great number of proteins as potential targets to be confirmed via cost-intensive and time-consuming wetlab experiments. However, incorporating immunoinformatic filters which extract highly potent target proteins into the RV process could reduces these costs. Part of the immunoinformatics focus is on small peptides ranging from 8 to 11 residues, called linear epitopes, and particularly on those that strongly bind to MHC class I molecules. Just one epitope per protein can be enough to create an immunological response in a host. Bioinformatics techniques to search for linear epitopes are well understood and available, but they can also lead to high false

© Oxford University Press 2012

positive rates. Despite these false positives, epitope predictors are capable of identifying weak epitope motifs or even strong motifs that have been experimentally neglected (Santos et al., 2011).

Epitope density has been described in research ranking *Mtb* proteins as a function of "hot spots" or regions with enriched MHC class II binding epitopes (Gaseitsiwe et al., 2010). This work reported 544, 609 and 757 15mers peptides binding, respectively, to three, two and just one of the molecules HLA-DR1, -DR2, and -DR4. Analysis of two of the 61 proteins examined in that study showed that Ag85B and MPT63 contain, respectively, 30 and 23 peptides highest binding to MHC molecules; however, there was experimental information available for only 10 peptides which all derived from MPT63.

Asking whether specific defined domains have high epitope densities, were found that signal peptides and trans-membrane domains have exceptionally high epitope densities (Kovjazin et al., 2011). This work computed the high epitope density of signal peptides using *in silico* methods and corroborated by the high percentage of identified signal peptides epitope in the IEDB (immune epitope database). The enhanced immunogenicity of signal peptides was experimentally confirmed using peptides derived from *Mtb* proteins. Were found high antigen specific response rates and population coverage to signal peptides sequences compared with non-signal peptides peptide antigens derived from the same proteins.

Our approach is based on the hypothesis of not only immunogenicity but also pathogenicity being revealed by the overall set of predicted epitopes and their concentrations in mature proteins. This is concept, which we called MED (Mature Epitope Density) is similar to epitope density (Gaseitsiwe et al., 2010; Kovjazin et al., 2011). We report experimental evidence demonstrating a direct relationship between MED and pathogenicity in the *Mycobacterium tuberculosis* (*Mtb*), strain H37Rv.

2 METHODS

Genome data: The complete genome of *Mtb* H37Rv was obtained from the GenBank database under the identifier NCBI: NC_000962. All coding sequences were selected and exported as amino acids in FASTA format, using the annotation software ARTEMIS, from Sanger Institute.

Prediction schema: Our software environment integrates SurfG+, TMHMM and NetMHC for MED predictions (Barinov et al., 2009; Krogh et al., 2001; Lundegaard et al., 2008b). An amino acid MULTIFASTA file is first processed by SurfG+ to filter sequences predicted to be secreted (SEC) and potentially surface exposed (PSE). The SEC sequences have their signal peptide intervals removed from the original sequence, maintaining only the predicted protein mature sequences for further processing (Kovjazin et al., 2011). This also was done for the predicted PSE sequences; but another TMHMM prediction step was used on the sequences because SurfG+ does not store the TMHMM results concerning the mature portions of the sequences. An artificial amino acid sequence is created from each original amino acid sequence predicted as SEC and PSE, containing only the concatenated original amino acid sequence portions that were predicted as the mature portions. Then, the artificial amino acid sequences are submitted to NetMHC, configured to predict all the 55 possible MHC alleles within the software (version 3.0). Only the predicted strongly binding peptides are filtered for further processing. Finally, the MED score is calculated for each amino acid sequence, according to Equation 1.

$$\mathrm{MED}=\frac{\mathit{Predictions}}{\mathit{Chances}}=\frac{\mathit{Predicted\,epitopes}*(50-\mathrm{Average}(\mathit{MHC\ Affinity}))}{\mathit{Aminoacids\,length}-\mathit{Epitope\ length}+1} \qquad (1)$$

Equation (1) divides the number of linear predicted epitopes from each amino acid sequence by the number, for instance, of possible 9mers peptides overlapping windows. In order to ensure qualitative differentiation for this ratio calculation, the epitopes' MHC bind affinity average is multiplied also, after be normalized according to the maximum MHC strong bind affinity (50 nM). This calculation bears the Mature Epitope Density (MED), a number measured in nanomolar per mer (nM/mer) units. All amino acid sequences are ordered by descending MED score and presented as the final result. The prediction schema was implemented using a Linux shell script. The web server is hosted in Ubuntu OS, release 10.10. All of the processing spent about 90 minutes for *Mtb* H37Rv amino acid sequences, using a standard personal desktop computer.

Control datasets: 100 antigens and 100 non-antigens swissprot identifiers were obtained from a previous work (Doytchinova and Flower, 2007). These protein identifiers were retrieved from the Uniprot database (UniProt, 2012), culminating in 107 and 121 amino acid sequences used as positive (ctrl+) and negative (ctrl-) control groups, respectively. To enrich our tests, a set of 38 Mtb's proteins (Mtb+) were similarly retrieved, first from AntigenDB (Ansari et al., 2010) and finally from Uniprot. The Mtb+ control group was obtained selecting the antigenic proteins from *M. tuberculosis* and filtering for those knew as eliciting immune cellular responses.

Evidence dataset: Sixty proteins, out of 553 predicted *in silico* exported, were chosen for detailed investigation of experimental proof concerning their capacity to induce cellular responses. In this regard, 30 out of 60 proteins were lowest scored and the others 30 were highest scored based on MED. An extensive literature search was carried out to look for evidence concerning whether or not a protein is related to immunogenicity or bacterial pathogenicity. Were found supporting evidence for 41 out of 60 proteins, depending on if a protein: induces a cellular response, has evidence of frame shifts, differential expression, is part of a known pathogenic protein family or has a cloning experiment that failed. The complete evidence dataset and corresponding published evidence can be found in the supplementary material.

3 RESULTS

3.1 Allele frequency

Figure 1 shows a MHC allele histogram from the predicted epitopes for the *Mtb* H37Rv exported proteins. The MHC alleles are ordered according to their decreasing number of predicted epitopes. The first five MHC alleles are human and represent 52.32%, while the first 15 represent 80.83% of all predicted epitopes. The last 24 MHC alleles represent only 2.58% of the overall NetMHC epitope prediction.

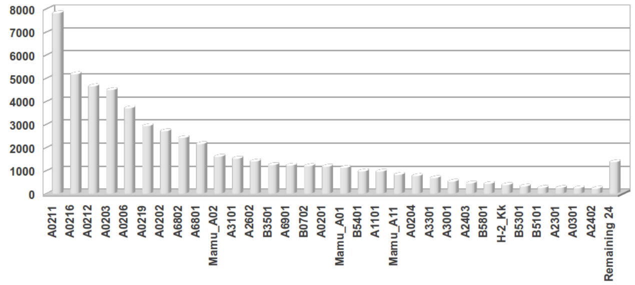


Figure 1: MHC alleles in the software NetMHC and the number of predicted strong binders to epitopes from Mtb H37Rv exported proteins.

3.2 Control datasets

The control groups were analyzed in our pipeline and the main results are presented in the Figure 2. The results were divided in panels exhibiting protein quantities, percentage regarding these quantities and average MED score, all subdivided according to the predicted sub-cellular location. The number of proteins (panel 1) and the percentage (panel 2) predicted as cytoplasmic are the majority for the groups ctrl- and Mtb+, whilst the ctrl+ have more predicted exported proteins. Curiously, the Mtb+ has the majority of proteins predicted as cytoplasmic. It's surprising because is expected the majority of antigenic proteins as exported to extracellular milieu, just like observed in the ctrl+ group that contains several pathogenic organisms.

Two results should be noted in the panel 3. First, the average MED scores were kept very similar in the three control groups, showing that MED is not necessarily a binary statistic classifier for targets but a continuous statistic capable to aim the definition of preferable targets over the others. However, when there are significant differences between MED scores, we can use it just like a binary classifier. This procedure was assessed in the evidence dataset, the next section. Second, the average MED score for proteins predicted as membrane integral is twice greater than the others sub-cellular compartments. This result is corroborated by a work, where signal peptides and trans-membrane domains were found to have exceptionally high CD8+ T cell epitope densities (Kovjazin et al., 2011).

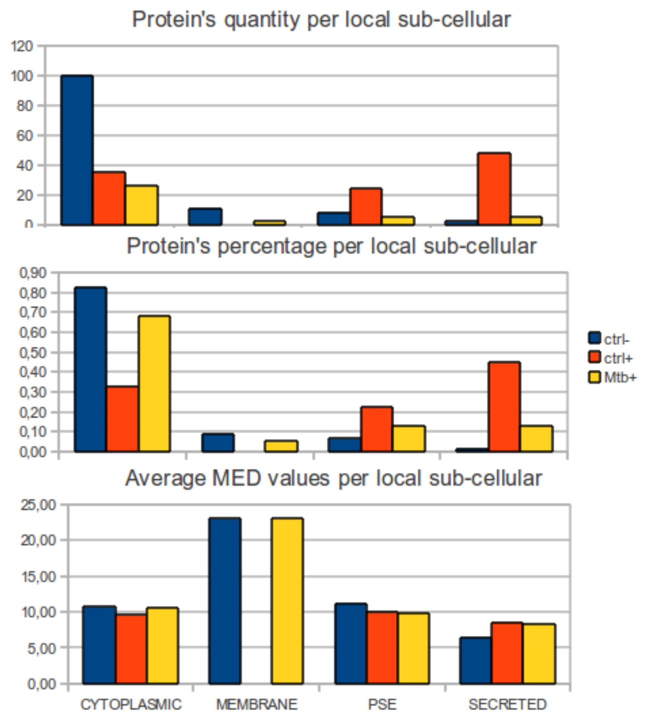


Figure 2: Quantities (panel 1), percentages (panel 2) and average MED scores, per predicted local sub-cellular, for antigenic (+) and non-antigenic (-) protein's control groups. These include *M. tuberculosis* antigenic proteins (Mtb+) observed eliciting immune cellular responses.

3.3 Evidence dataset

In Table 1, MED scores range from 15.67 to 27.00 nM/mer, with the highest MED score data set represented on the far right of Figure 3. These values strongly contrast with MED scores in Table 2, which are between 0.00 and 3.19 nM/mer, with the lowest MED score dataset represented on the far left of Figure 3. Analyses of the proteins scored within these extremely different ranges allow us to develop evidence for the general importance of MED scores.

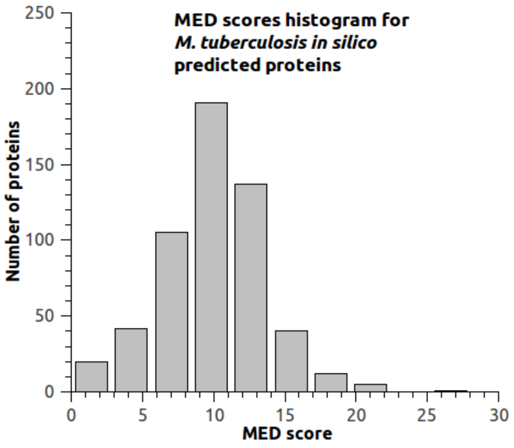


Figure 3: MED scores histogram for *Mtb* H37Rv exported proteins. Data in Tables 1 and 2 are situated in the extremities.

3.4 MED score limitations

Figure 4 shows a box plot from the data used to calculate the MED score in the sample dataset. In Figure 5, 41 out 60 proteins were used to plot the receiver operating characteristic (ROC) curve; the sample of 60 proteins (30 lowest and 30 highest MED scored) was used in . The numerator and denominator for each protein were investigated to determine how protein length can influence the MED score. The number of epitopes predicted in the highest-scored subset is more than twice as high as in the lowest-scored subset. This was expected, since there is evidence that the highest-scored subset is composed of proteins related to pathogenicity while the lowest-scored subset is not. The number of possibilities for linear epitopes in the lowest-scored subset is almost three times as high as compared to the highest-scored subset. This numerical difference in the denominators can be a limitation for the MED score strategy, especially in data above the median. Quartiles Q3 and Q4 among those with lowest chances included half (7/14) of the evidence contrary to our hypothesis of an existent relation between MED and pathogenicity. These quartiles include denominators between 537 and 1,860 (just one greater than 1,498). Thus, according to the data, MED scores tend to indicate false positives when there is a difference factor of at least five between the number of predictions and the number of epitope's possibilities located in the mature amino acid sequence portion. No false positives were observed when this factor was less than two. An interesting result is that the two biggest controls groups from the had average factor (fold) of 3.22 and 2.82 for ctrl+ and ctrl-, respectively.

© Oxford University Press 2012

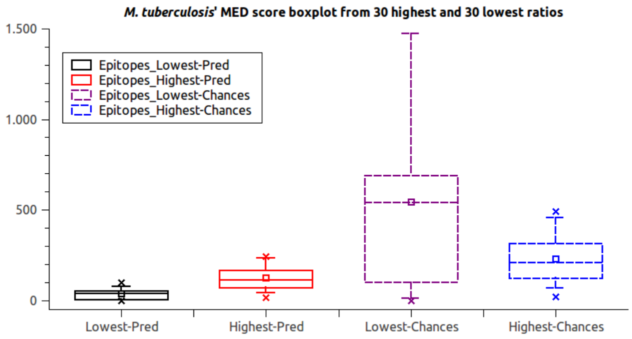


Figure 4: Boxplot of numerators and denominators within the 60 lowest and highest MED scores from the *Mtb* H37Rv exported proteins. "Pred" stands for epitope predictions and "Chances" stands for possible 9mers windows in an amino acid sequence's mature portion; both are used in Equation 1.

3.5 Evidence dataset

Tables 1 and 2 summarize results for 41 MED scores from the *Mtb* H37Rv exported proteins. Table 1 list 22 of the 30 highest MED scored-proteins and Table 2 lists 14 of the 30 lowest-scored proteins. Each protein is accompanied by at least one unique publication identifier. These can be doi, pubmed id or a patent number. A protein can be cited twice or thrice by different publications; some publications cite several proteins. The first columns in Tables 1 and 2 show the protein locus tags, followed by the number of predicted epitopes (n) and epitope probability as a function of its proportion in the mature protein (d). The MED score is calculated as n divided by d. An evidence can be favorable or contrary based on publication results and the expectation indicated by the MED score.

Genome Locus	*n*	*d*	MED (nM/mer)	Local	Evidence	Unique publication identifier
Rv2452c	14	18	27,00	SEC	favorable	10.1046/j.1365-2958.1999.01593.x
Rv1811	66	108	21,34	PSE	favorable	PMID:10760138
Rv3018c	145	234	20,72	PSE	favorable	10.1099/jmm.0.47565-0, 10.1046/j.1365-2958.1999.01593.x, 10.1016/j.vaccine.2004.08.046
Rv1489	37	63	20,36	PSE	favorable	10.1186/1471-2180-10-132, 10.1021/pr0500049, 10.1016/j.tube.2008.01.003
Rv0847	58	98	19,89	SEC	favorable	10.1016/j.tube.2006.01.014, 10.1016/j.tube.2006.01.014
Rv0436c	78	123	19,14	PSE	favorable	10.1074/jbc.M004658200
Rv0116c	117	214	17,61	SEC	favorable	10.1099/mic.0.024802-0
Rv1841c	167	308	17,33	PSE	favorable	10.1128/jb.184.4.1112-1120.2002
Rv2339	224	437	17,25	PSE	favorable	10.1093/molbev/msm111
Rv0589	195	364	17,10	PSE	favorable	10.1007/s11010-011-0733-5
Rv1158c	107	189	17,07	SEC	favorable	10.1016/j.tube.2004.09.005
Rv0286	129	242	17,04	PSE	favorable	10.1128/IAI.70.12.6996–7003.2002
Rv3497c	161	314	16,87	SEC	favorable	10.1073/pnas.1631248100
Rv1967	151	305	16,53	SEC	favorable	10.1111/j.1574-695X.2010.00677.x
Rv1620c	156	311	16,52	PSE	favorable	10.1073/pnas.1003219107, 20090285847
Rv3000	86	167	16,04	PSE	favorable	10.1016/j.tube.2006.01.014
Rv1522c	236	469	16,33	PSE	favorable	10.1073/pnas.0401657101
Rv2690c	64	126	16,03	PSE	favorable	Patent 7393540
Rv0804	87	175	15,85	SEC	favorable	10.1107/S1744309108031679
Rv0598c	58	104	15,85	SEC	favorable	PMID:12657046
Rv3693	203	404	15,69	SEC	favorable	10.4049/jimmunol.1002212, 10.1002/pmic.200600853
Rv2262c	100	206	15,69	PSE	favorable	PMID:12368431

Table 1. Proteins with experimental evidence for their MED scores, including the subset of the highest-scoring proteins. The n column gives the numerator, while the d column indicates the denominator for Equation 1.

Genome Locus	*n*	*d*	MED (nM/mer)	Local	Evidence	Unique publication identifier
Rv0532	59	555	3,19	SEC	contrary	10.1021/pr1005108
Rv0746	77	741	3,11	SEC	contrary	10.1186/1471-2148-6-95, 10.1016/j.micinf.2006.03.015
Rv1468c	37	328	3,03	SEC	contrary	10.1021/pr1005108
Rv3590c	48	542	2,96	SEC	favorable	10.1016/S1672-0229(08)60039-X
Rv3511	66	678	2,91	SEC	favorable	10.1186/1471-2148-6-95
Rv1100	20	160	2,88	PSE	contrary	10.1099/mic.0.27204-0
Rv3312A	4	64	2,69	SEC	contrary	10.1073/pnas.0602304104
Rv3595c	34	400	2,51	SEC	contrary	10.1186/1471-2148-6-95
Rv1091	60	814	2,40	SEC	contrary	10.1186/1471-2148-6-95
Rv3706c	4	50	2,32	PSE	contrary	10.3389/fmicb.2010.00121
Rv1396c	37	537	2,30	SEC	favorable	10.1046/j.1365-2958.2002.02813.x
Rv3345c	98	1498	2,05	SEC	favorable	10.1186/1471-2148-6-95, 10.1099/mic.0.26660-0
Rv0559c	4	78	2,05	SEC	contrary	10.1371/journal.pone.0007615
Rv3388	44	690	2,03	SEC	contrary	10.1016/j.tube.2003.12.014
Rv0833	52	689	1,75	PSE	favorable	10.1186/1471-2148-6-95
Rv2487c	28	655	1,15	SEC	contrary	Patent EP2207035
Rv3514	43	1448	0,93	SEC	contrary	10.1111/j.1365-2567.2010.03383.x
Rv3508	40	1860	0,71	SEC	contrary	10.1371/journal.pone.0002375, 10.1002/prot.10586
Rv3655c	0	0	0	PSE	contrary	10.1371/journal.pone.0010474

Table 2. Proteins with experimental evidence for their MED scores and listing a subset of the lowest scoring proteins. The n column shows the numerator, while the d column indicates the denominators for Equation 1.

3.6 MED score sensitivity

Among the 30 proteins that were lowest scored, 14 had evidence contrary and five had evidence favorable to the MED score concept. Among the 30 highest scored proteins, there was favorable evidence for 22 proteins based on the MED score, and there was no protein with contrary evidence. Among the lowest and highest scored remainders, none had favorable or contrary evidence related to MED scores. These results were used to create

© Oxford University Press 2012

a ROC curve graph & that calculates sensitivities of 81% for MED scores with 7% false positives.

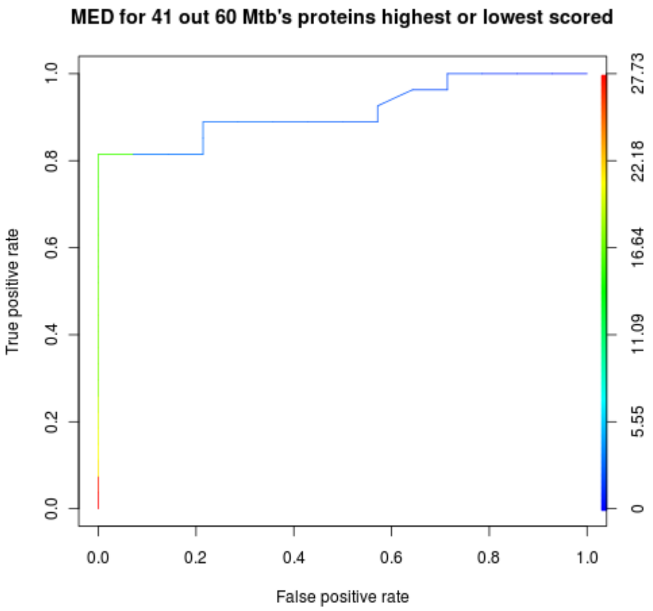


Figure 5: Receiver operating characteristic (ROC) curve from the mature epitope density (MED) score calculated for 41 *Mtb* H37Rv exported proteins.

3.7 Novel probable putative *Mtb* antigens

The *Mtb* H37rv proteins Rv0235c, Rv0492A and Rv1004c were predicted to have some of the highest MED scores, 17.78, 20.31 and 18.58 nM/mer, respectively. The former two are predicted as being potentially exposed on the bacterial surface and the latter is predicted as secreted. Respectively, there are 78, 43 and 228 predicted epitopes against 138, 73 and 386 epitope chances for these proteins. This is the first published indication of their roles in bacterial pathogenicity. MED scoring results suggest these proteins as useful putative targets for future investigations.

4 DISCUSSION

4.1 Allele frequency

The available methods for MHC epitope prediction take into account allele frequency in the selection of potential candidates (Larsen et al., 2007; Lundegaard et al., 2008a). Some alleles are extremely rare; some are specific of some population or are widespread (Gupta et al., 2010). The here applied tools to search for epitopes are not novel but how we read results from standard software tools can be considered a novelty. We here proposed to interpret no only epitope prediction from some specific MHC alleles but from all available alleles. This propose have a rationale: the idea of to assess the immunogenic potential of a protein, independent of alleles, helps avoid to exclude a protein from a list of candidates *in silico* just because the allele suitable for a specific population was not selected. For example, there are pathogenic organisms causing different diseases in different hosts, including humans, caprines, ovines, equines, bovines and buffaloes (Aleman et al., 1996; Baird and Fontaine, 2007; Join-Lambert et al., 2006; Liu et al., 2005; Mills et al., 1997; Peel et al., 1997; Selim, 2001; Williamson, 2001; Yeruham et al., 2003). In such case, it is not reasonable to exclude not a single allele from the current limited number available in software tools. Still, if a specific set of alleles is desired, these could be selected via medpipe.

4.2 Control datasets

Even within the ctrl- group, the average MED scores were similar to those from the ctrl+ and Mtb+ groups. Because of this, we focus on predicted exported proteins to create a priority list of targets for *Mtb* genome. It's a reasonable strategy, since one of the main differences between the ctrl- and the ctrl+ groups are the number of predicted cytoplasmic versus exported proteins; 111 and 10 for ctrl- versus 35 and 72 for ctrl+, respectively. It's more likely that exported proteins interact with the host cells than membrane and cytoplasmic proteins (Santos et al., 2011; Sibbald and van Dij, 2009; Simeone et al., 2009; Stavrinides et al., 2008). However, it's important also do not neglect proteins that could be exported via non-classical mechanisms. This conclusion can also be drew analyzing, the Figure 2, panel 2, where the majority of Mtb+ proteins fits on cytoplasmic ones. Medpipe allows predicting cytoplasmic targets, but this is the major part of any bacterial genome; medpipe still do not allows differentiating between cytoplasmic proteins and those without classical exportation motifs and exported via non-classical pathways.

Besides, it's quite difficult to compare MED score with previous software trained for antigenic features because such programs tend to be binary classifiers (Doytchinova and Flower, 2007; Tung et al., 2011; Wang et al., 2011). For instance, two controls dataset here used were split in to training sets (75 proteins) and test sets (25 proteins). Such division does not make senses for MED score because it does not depend of training step. Instead, the MED technique searches for immunological features based on a probable immunological memory concerning epitopes from knew pathogens. In this regard, the results obtained with the evidence dataset is more informative because they represents experimental evidence about the predictive strengthens or weakens of the method.

4.3 Evidence dataset

An extensive literature search for proteins from the well studied organism *Mtb* gave experimental indication to validate our hypothesis of a protein's relation to pathogenicity being revealed based on the overall set of predicted epitopes. When searching evidence about the proteins within the evidence dataset, were also found experimental results about another 43 proteins but these were not here included, because it is difficult to determine if these proteins should be presented as false or true positives. Such dilemma was less difficult to workaround when considering only 60 proteins: the 30 higher and the 30 lowest MED scored proteins out of 553 *Mtb*'s predicted exported proteins &.

4.4 NetMHC version

The newest NetMHC software (version 3.2) offers the possibility to predicted epitopes for 57 MHC alleles (www.cbs.dtu.dk/services/NetMHC/) but there is not yet a stand-

© Oxford University Press 2012

alone version available for download. The here used NetMHC version (3.0) is the previous one and offers the possibility to predict epitopes for 55 MHC alleles (Lundegaard et al., 2008a). However, the novelties of the version 3.2, compared to the version 3.0, are a small increment in the number of MHC alleles and the possibility to predicted epitopes of lengths from 8 to 14mers. The authors of version 3.2 advise that predictions of peptides longer than 11mers have not been extensively validated. They advise also that caution should be taken for 8mers predictions as some alleles might not bind 8mers to any significant extend (www.cbs.dtu.dk/services/NetMHC/). Besides, most of MHCs prefers peptides of 9mers and the alleles' set from the version 3.0 still are present within the version 3.2 (Lundegaard et al., 2008b). Therefore, epitope predictions based on the version 3.0 still are valid to answer relevant biological queries.

4.5 Pathogenic proteins?

The here presented method was initially conceived to predict immunogenic proteins. As related in the methods section, the *in silico* predicted exoproteins were ordered by decreasing value of their MED scores. After that, we search the literature for evidence proving or denying the immunogenicity from each protein. Surprising, the majority of the here presented true positives (Table 1) had pathogenic evidence instead of immunogenic evidence (18 out 22 true positives), as detailed in the supplementary material. One protein (Rv3018c) has evidence for immunogenicity and pathogenicity, simultaneously. In the same way, this criterion was applied also with the true negatives (Table 2), where seven out 14 contrary cases fit pathogenic class instead immunogenic. Could these apparently unexpected results have a rationale? Could the Pathogenomics give us a rationale about these findings? The Pathogenomics is defined as the analysis at the genomic level of the processes involved in bacterial pathogenesis caused by the interaction of pathogenic microbes and their hosts (Hacker, 2006b). The identification of mutants showing altered pathology may be a useful framework for understand tuberculosis but it is not clear how these phenotypes relate to human tuberculosis (Hacker, 2006a). Here we presented evidence that *Mtb* pathogenic proteins have some of the highest MED scores within the *Mtb* genome.

4.6 Concluding remarks

The search for new vaccine targets against prokaryotic microorganisms has been leveraged by extensive use of software motif recognition in sequences. Nevertheless, considerable experimental effort is necessary to filter out the most promising candidates. The here presented method and software available on line can help minimizing experimental efforts by indicating promising prokaryotic proteins related to immunogenicity and pathogenicity. The proposed method was called MED score, exhibiting a strong relation to important proteins in the M. tuberculosis pathogenesis.

5 IMPLEMENTATION AND AVAILABILITY

The MED pipeline tool is implemented in PHP and Linux shell scripts. It is freely available at http://lgcm.icb.ufmg.br/medpipe.

ACKNOWLEDGEMENTS

We thank our colleagues at the Molecular and Cellular Genetics Laboratory (LGMC) of the Federal University of Minas Gerais (UFMG), Brazil, for sharing research skills and for their support. JB and JP are grateful for financial support from the Cluster of Excellence for Multimodal Computing and Interaction (MMCI) of the German Research Foundation.

Funding: CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil.

Conflict of Interest: none declared.

REFERENCES

Aleman, M., Spier, S.J, Wilson, W.D and Doherr, M. (1996). Corynebacterium pseudotuberculosis infection in horses: 538 cases (1982-1993). *J Am Vet Med Assoc*, **209**, 804-809.

Ansari, H.R., Flower, D.R and Raghava, G.P.S. (2010). AntigenDB: an immunoinformatics database of pathogen antigens. *Nucleic Acids Res*, **38**, D847-53.

Baird, G.J. and Fontaine, M.C. (2007). Corynebacterium pseudotuberculosis and its role in ovine caseous lymphadenitis. *J Comp Pathol*, **137**, 179-210.

Barinov, A., Loux, V, Hammani, A, Nicolas, P, Langella, P, Ehrlich, D, Maguin, E and van de Guchte, M. (2009). Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria. *Proteomics*, **9**, 61-73.

Doytchinova, I.A. and Flower, D.R. (2007). Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine*, **25**, 856-866.

Gaseitsiwe, S., Valentini, D, Mahdavifar, S, Reilly, M, Ehrnst, A and Maeurer, M. (2010). Peptide microarray-based identification of Mycobacterium tuberculosis epitope binding to HLA-DRB1*0101, DRB1*1501, and DRB1*0401. *Clin Vaccine Immunol*, **17**, 168-175.

Gupta, S.K., Smita, S, Sarangi, A.N, Srivastava, M, Akhoon, B.A, Rahman, Q and Gupta, S.K. (2010). In silico CD4+ T-cell epitope prediction and HLA distribution analysis for the potential proteins of Neisseria meningitidis Serogroup B—A clue for vaccine development. *Vaccine*, **28**, 7092-7097.

Hacker, J.. (2006a). Pathogenomics: Insights into Tuberculosis and Related Mycobacterial Diseases. In U. Dobrindt and W. Gobel (ed.), *Pathogenomics: Genome Analysis of Pathogenic Microbes*. Wiley VCH Verlag GmbH & Co. KGaA Weinheim , Vol. 2, pp. 616.

Hacker, J.. (2006b). *Pathogenomics: Insights into Tuberculosis and Related Mycobacterial Diseases*. Wiley VCH Verlag GmbH & Co. KGaA Weinheim .

Join-Lambert, O.F., Ouache, M, Canioni, D, Beretti, J, Blanche, S, Berche, P and Kayal, S. (2006). Corynebacterium pseudotuberculosis necrotizing lymphadenitis in a twelve-year-old patient. *Pediatr Infect Dis J*, **25**, 848-851.

© Oxford University Press 2012

Kovjazin, R., Volovitz, I, Daon, Y, Vider-Shalit, T, Azran, R, Tsaban, L, Carmon, L and Louzoun, Y. (2011). Signal peptides and trans-membrane regions are broadly immunogenic and have high CD8+ T cell epitope densities: Implications for vaccine development. *Mol Immunol*, **48**, 1009-1018.

Krogh, A., Larsson, B, von Heijne, G and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567-580.

Larsen, M.V., Lundegaard, C, Lamberth, K, Buus, S, Lund, O and Nielsen, M. (2007). Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics*, **8**, 424.

Liu, D.T.L., Chan, W, Fan, D.S.P and Lam, D.S.C. (2005). An infected hydrogel buckle with Corynebacterium pseudotuberculosis. *Br J Ophthalmol*, **89**, 245-246.

Lundegaard, C., Lamberth, K, Harndahl, M, Buus, S, Lund, O and Nielsen, M. (2008a). NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. *Nucleic Acids Res*, **36**, W509-12.

Lundegaard, C., Lund, O and Nielsen, M. (2008b). Accurate approximation method for prediction of class I MHC affinities for peptides of length 8, 10 and 11 using prediction tools trained on 9mers. *Bioinformatics*, **24**, 1397-1398.

Mills, A.E., Mitchell, R.D and Lim, E.K. (1997). Corynebacterium pseudotuberculosis is a cause of human necrotising granulomatous lymphadenitis. *Pathology*, **29**, 231-233.

Peel, M.M., Palmer, G.G, Stacpoole, A.M and Kerr, T.G. (1997). Human lymphadenitis due to Corynebacterium pseudotuberculosis: report of ten cases from Australia and review. *Clin Infect Dis*, **24**, 185-191.

Rowland, R. and McShane, H. (2011). Tuberculosis vaccines in clinical trials. *Expert Rev Vaccines*, **10**, 645-658.

Rylance, J., Pai, M, Lienhardt, C and Garner, P. (2010). Priorities for tuberculosis research: a systematic review. *Lancet Infect Dis*, **10**, 886-892.

Santos, A.R., Ali, A, Barbosa, E, Silva, A, Miyoshi, A, Barh, D and Azevedo, V. (2011). The reverse vaccinology – A contextual overview. *IIOABJ*, **2**, 8-15.

Selim, S.A.. (2001). Oedematous skin disease of buffalo in Egypt. *J Vet Med B Infect Dis Vet Public Health*, **48**, 241-258.

Sibbald, M.J.J.B. and van Dij, J.M.L. (2009). Secretome Mapping in Gram-Positive Pathogens. In Karl Wooldridge (ed.), Bacterial Secreted Protein: Secretory Mechanisms and Role in Pathogenesis. *Caister Academic Press*, , 193-225.

Simeone, R., Bottai, D and Brosch, R. (2009). ESX/type VII secretion systems and their role in host-pathogen interaction. *Curr Opin Microbiol*, **12**, 4-10.

Stavrinides, J., McCann, H.C and Guttman, D.S. (2008). Host-pathogen interplay and the evolution of bacterial effectors. *Cell Microbiol*, **10**, 285-292.

Thaiss, C.A. and Kaufmann, S.H.E. (2010). Toward novel vaccines against tuberculosis: current hopes and obstacles. *Yale J Biol Med*, **83**, 209-215.

Tung, C., Ziehm, M, Kämper, A, Kohlbacher, O and Ho, S. (2011). POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics*, **12**, 446.

UniProt, C.. (2012). Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res*, **40**, D71-5.

Wang, H., Lin, Y, Pai, T and Chang, H. (2011). Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. *J Biomed Biotechnol*, **2011**, 432830.

Williamson, L.H.. (2001). Caseous lymphadenitis in small ruminants. *Vet Clin North Am Food Anim Pract*, **17**, 359-71, vii.

Yeruham, I., Elad, D, Friedman, S and Perl, S. (2003). Corynebacterium pseudotuberculosis infection in Israeli dairy cattle. *Epidemiol Infect*, **131**, 947-955.

© Oxford University Press 2012

3.2.5.2 Aplicação da MED em *C. pseudotuberculosis*

Os resultados a seguir foram obtidos por meio da aplicação da metodologia para medir a estatística MED, metodologia apresentada na seção 3.2.5.1 . Essa estatística foi aplicada sobre as proteínas preditas *in silico* de pertencerem ao exoproteoma do pangenoma de *C. pseudotuberculosis,* pangenoma homogeneizado de acordo com a seção 3.2.1.

Apresenta-se as maiores médias da densidade de epitopos na porção madura de proteínas de cinco genomas, linhagens 1002, C231, FRC41 e PAT10. Aproximadamente 90 e 200 proteínas, por genoma, preditas como secretadas e potencialmente expostas na superfície, foram utilizadas para gerar essa tabelas, respectivamente.

Na Tabela 3, são exibidos 15 identificadores do pangenoma (pan *locus*) relativos a proteínas preditas como secretadas. Também apresenta-se a Tabela 4 com 45 identificadores do pangenoma relativos a proteínas preditas como potencialmente expostas na superfície bacteriana. Para essas predições foram utilizados todos os alelos de MHC disponíveis no programa NetMHC (Lundegaard e cols., 2008), em sua maioria alelos de MHC humano. Esse método é apoiado pelo fato da tanto a *M. tuberculosis* quanto a *C. pseudotuberculosis* infectarem humanos e outros animais. O fator de corte para definir a quantidade de elementos nessas duas tabelas foi o valor médio da estatística MED de proteínas secretadas (10.92 nM/mer) e potencialmente expostas na superfície (10.72 nM/mer).

Convém lembrar que cada identificador de um pangenoma corresponde a um número de genes igual à quantidade de genomas analisados no pangenoma. Dessa forma, os 60 identificadores do pangenoma presentes nessas duas tabelas correspondem a 300 (5x60) proteínas que pertencem ao pangenoma central das cinco linhagens de *C. pseudotuberculosis* analisadas. As tabelas completas listando os pan locus (identificador único de genes homologos no pangenoma) e respectivos *locus* do exoproteoma predito *in silico* de *C. pseudotuberculosis* estão na seção de anexos.

Ordem	Pan locus	MED	Local subcelular
1	plcpsec057	21.61	SECRETED
2	plcpsec043	17.22	SECRETED
3	**plcpsec074**	**16.26**	**SECRETED**
4	plcpsec035	14.17	SECRETED
5	plcpsec005	14.16	SECRETED
6	**plcpsec056**	**14.06**	**SECRETED**
7	**plcpsec067**	**13.13**	**SECRETED**
8	plcpsec031	12.62	SECRETED
9	plcpsec093	12.54	SECRETED
10	plcpsec089	12.30	SECRETED
11	plcpsec076	12.04	SECRETED
12	plcpsec010	11.62	SECRETED
13	plcpsec072	11.54	SECRETED
14	plcpsec085	11.31	SECRETED
15	plcpsec059	11.11	SECRETED

Tabela 3: Média das maiores densidades de epitopos maduros (MED) em proteínas exportadas homólogas do pangenoma de *C. pseudotuberculosis* e preditos como secretados. Em negrito estão os candidatos sob testes experimentais e encontrados no exoproteoma.

Ordem	Pan locus	MED	Local subcelular
1	plcppse067	17.72	PSE
2	plcppse118	17.06	PSE
3	plcppse166	16.00	PSE
4	plcppse152	15.38	PSE
5	plcppse144	15.25	PSE
6	plcppse110	15.02	PSE
7	plcppse176	14.58	PSE
8	plcppse124	14.51	PSE
9	plcppse122	13.87	PSE
10	plcppse087	13.63	PSE
11	plcppse021	13.57	PSE
12	plcppse169	12.95	PSE
13	plcppse005	12.84	PSE
14	plcppse072	12.79	PSE
15	plcppse034	12.64	PSE
16	plcppse121	12.53	PSE
17	plcppse160	12.52	PSE
18	plcppse132	12.51	PSE
19	plcppse092	12.38	PSE
20	plcppse017	12.36	PSE
21	plcppse080	12.27	PSE
22	plcppse150	12.22	PSE
23	plcppse009	12.14	PSE
24	plcppse138	11.96	PSE
25	plcppse048	11.89	PSE
26	plcppse125	11.89	PSE
27	plcppse075	11.88	PSE
28	plcppse070	11.76	PSE
29	plcppse126	11.69	PSE
30	plcppse023	11.59	PSE
31	plcppse140	11.58	PSE
32	plcppse093	11.55	PSE
33	plcppse062	11.39	PSE
34	plcppse073	11.38	PSE
35	plcppse097	11.29	PSE
36	plcppse020	11.13	PSE
37	plcppse159	11.08	PSE
38	plcppse010	11.05	PSE
39	plcppse158	11.00	PSE
40	plcppse032	10.95	PSE
41	plcppse079	10.94	PSE
42	plcppse170	10.92	PSE
43	plcppse054	10.88	PSE
44	plcppse085	10.85	PSE
45	plcppse106	10.75	PSE

Tabela 4: Média das maiores densidades de epitopos maduros (MED) em proteínas exportadas homólogas do pangenoma de *C. pseudotuberculosis* e preditos como potencialmente expostos na superfície (PSE).

Proteínas da categoria PSE são classificadas de acordo com a porção exterior à parede celular ser a C-terminal (PSE C) ou N-terminal (PSE N), com tamanhos maiores ou iguais a 50 aa. No caso da porção terminal ser superior a 100 aa, então uma proteína é classificada com PSE L. Proteínas identificadas pelo programa LipoP são classificados como PSE E e sinais de retenção são classificados como PSE R. Estes rótulos também podem ser conjugados para criar outras categorias de proteínas PSE.

Ao listar genes pertencentes ao secretoma pangenômico central e detentores dos maiores valores da estatística MED (Tabela 3), aparecem três dentre os quatro candidatos que estão em fase de validação *in vitro*. Dois desses três genes estão representados no gráfico da Figura 8 como os primeiros e maiores picos da estatística MED. O gráfico da Figura 8 representa os valores da MED aferidos em proteínas do exoproteoma central das linhagens 1002 e C231 e também preditos *in silico* no secretoma central da *C. pseudotuberculosis*. Os genes que aparecem na Tabela 3 identificados pelos pan locus 'plcpsec074' e 'plcpsec067' correspondem, no gráfico da Figura 8, aos *locus* Cp1002_0126a e Cp1002_1957, respectivamente. Os genes identificados pelo pan_id 'plcpsec056' são homólogos ao *locus* Cp1002_1802, da linhagem 1002. Este gene não está presente no gráfico da Figura 8, por ter sido encontrado apenas na linhagem 1002, em ambos os experimentos do exoproteoma (Pacheco e cols., 2011; Silva e cols., 2012).

O motivo pelo qual apenas três, dentre quatro pan locus da Tabela 3, estarem atualmente em testes reside no fato da escolha dos quatro alvos ter ocorrido quando os genomas ainda não estavam homogeneizados. Após a homogeneização dos cinco genomas, dois genes sob o pan_id 'plcpsec100' foram anotados como não sendo secretados. O gene da linhagem FRC41 (cpfrc_00367) foi anotado como citoplasmático e o gene da linhagem 1002 (Cp1002_0369) foi anotado como um possível pseudogene. Considerando que o genoma da FRC41 foi utilizado como referência para uma remontagem do genoma da 1002, existe a possibilidade de que essa região do genoma na linhagem FRC41 possua erros de montagem. Esses prováveis erros podem ter se propagado para a linhagem 1002 durante a última revisão de montagem do genoma da linhagem 1002. Convém salientar que na penúltima versão do genoma da linhagem 1002 esse gene possuía peptídeo sinal anotado e faz parte do proteoma exportado da referida linhagem. Além disso, esse gene na linhagem PAT10 foi anotado como potencialmente exposto na superfície, ao passo que nas linhagens C231 e I19 estão anotados como secretados.

Além da evidência experimental desses quatro conjuntos de genes pertencerem ao proteoma exportado, também existe evidência experimental para um deles, o pan_id 'plcpsec056', de forte hibridização com mimetopos da técnica de *PhageDisplay* (Almeida, 2011).

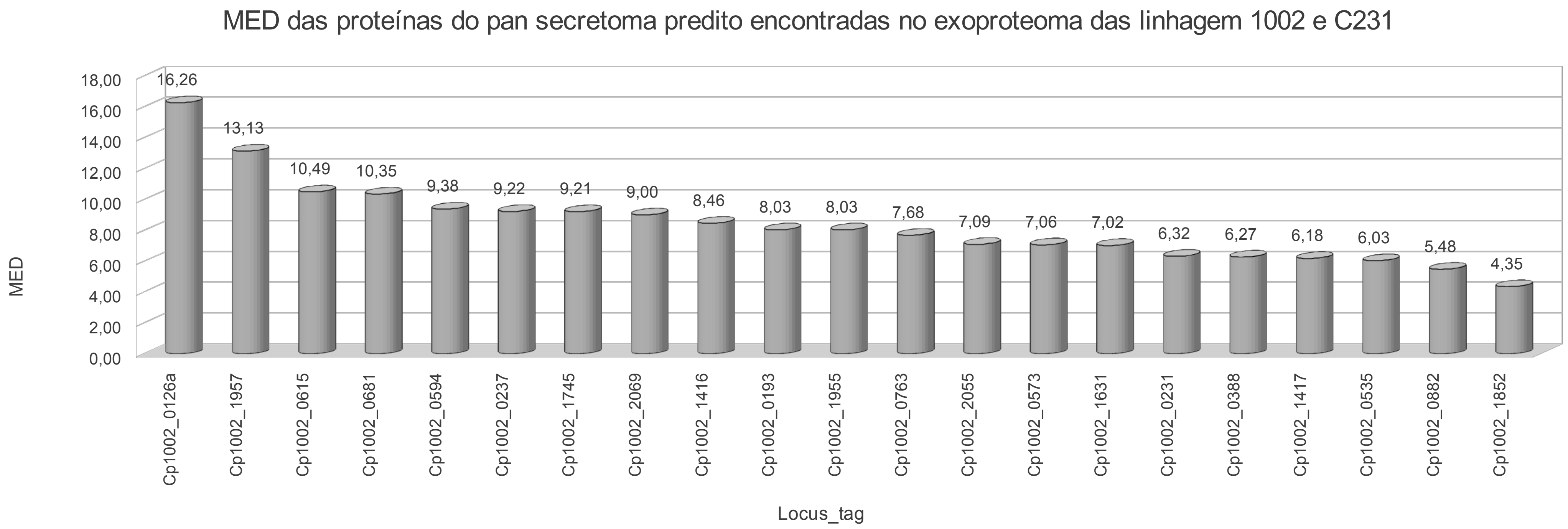


Figura 8: Análise do valor da MED no secretoma da *C. pseudotuberculosis*, linhagens 1002 e C231, obtido nos resultados de Pacheco e cols. (2010) e, ao mesmo tempo, também encontrados no pan secretoma predito *in silico*.

Outra análise relevante referente a aplicação da estatística MED, sobre o genoma de *C. pseudotuberculosis*, diz respeito às seis proteínas mais imunorreativas encontradas nos resultados da análise sorológica preliminar do secretoma (seção 3.2.4). Na Tabela 4, as médias dos valores da estatística MED para essas seis proteínas, nos cinco genomas analisados, ficaram próximas das médias das proteínas secretadas (8.60 nM/mer) e potencialmente expostas na superfície (9.22 nM/mer), conforme exibido na Figura 9. Os valores da MED da Tabela 4 também ficaram próximos dos limites mínimos empiricamente selecionados para a Tabela 3 (10.92 nM/mer) e Tabela 4 (10.72 nM/mer). Na Tabela 5 as quatro proteínas preditas *in silico* como secretadas possuem valores de MED que variam entre 8.35 e 10.35 nM/mer e as duas proteínas preditas como potencialmente expostas na superfície tem os valores 9.19 e 6.59 nM/mer.

Os resultados da estatística MED para essas seis proteínas mais imunoreativas evidenciam que a estatística MED é eficaz para selecionar alvos para a *C. pseudotuberculosis,* além de oferecer oportunidade para um aumento na quantidade de candidatos das tabelas 3 e 4.

Pan locus	**Idenficador do *NCBI***	**MED (nM/mer)**	**Local subcelular**
pgcpsec029	gi|302330380	10.35	SECRETED
pgcpsec012	gi|302329946	9.22	SECRETED
pgcpsec048	gi|302331099	8.46	SECRETED
pgcpsec033	gi|302330462	8.35	SECRETED
pgcppse060	gi|302205783	9.19	PSE
pgcppse143	gi|302331553	6.59	PSE

Tabela 5: Média da estatística MED para seis proteínas homologas encontradas nos genomas de *C. pseudotuberculosis* como as mais imunorreativas em uma análise sorológica preliminar.

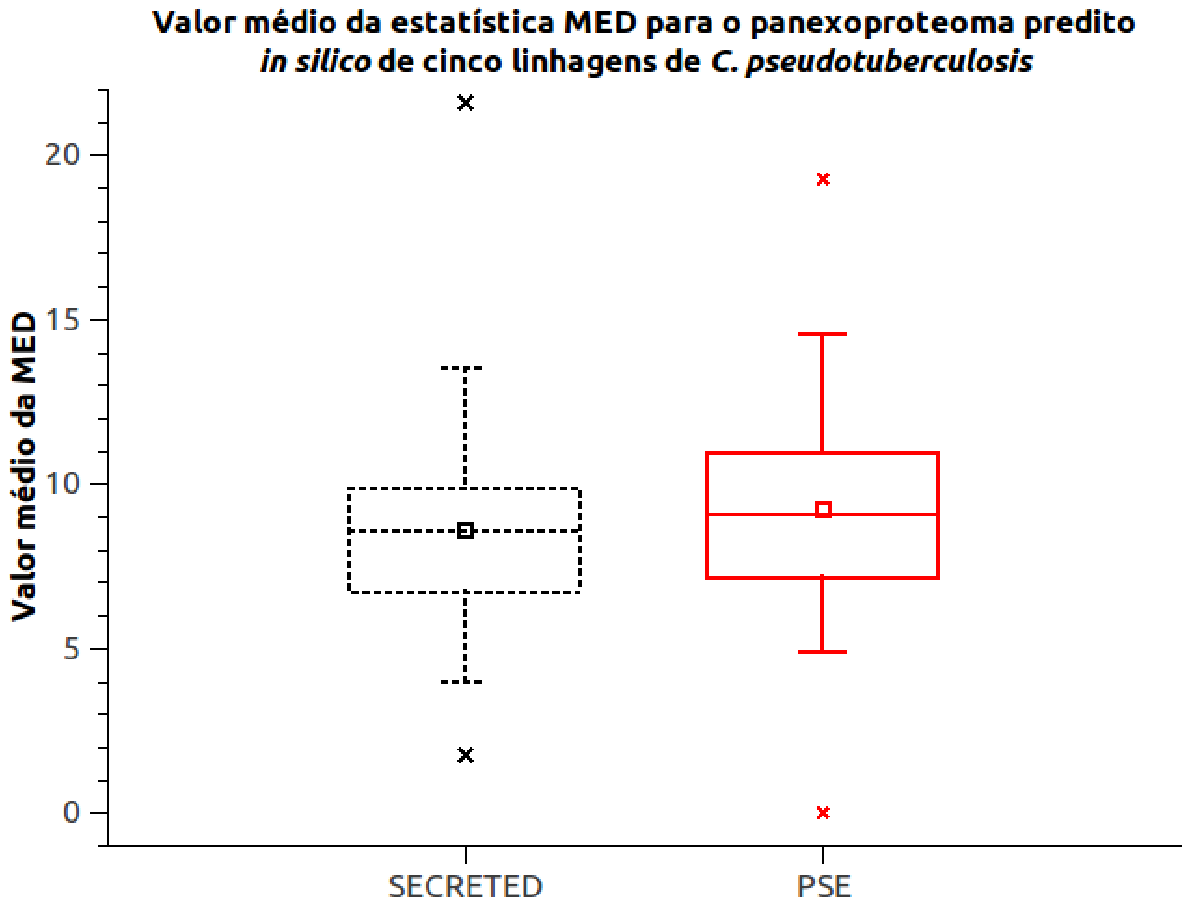


Figura 9: Boxplot da estatística MED para o exoproteoma predito *in silico* da *C. pseudotuberculosis*.

3.2.5.3 Discussão

Apesar do vasto repertório de programas de computadores para predição de epitopos e o crescente aumento na disponibilidade de novos genomas bacterianos completos, a determinação de proteínas consideradas promissoras do ponto de vista imunológico não é uma tarefa simples (Tsurui e Takahashi, 2007). Como citado nas seções anteriores, a predição de epitopos não lineares é uma tarefa difícil por embutir um problema computacional clássico e ainda sem solução ótima para a tecnologia atual da computação. Problemas computacionais como esses são conhecidos na ciência da computação como *NP-Completo*. Tais problemas envolvem uma grande quantidade de possibilidades na busca por uma solução ótima, que nesse caso remonta a descoberta da estrutura tridimensional formada pelo paratopo da molécula de *MHC* conjugada com possíveis epitopos (Sun e cols., 2011).

Como exemplo dos problemas envolvidos na predição de epitopos, vamos supor uma tentativa de realizar essa tarefa buscando uma solução ótima e determinística. Cada passo que é dado na determinação da interação entre dois aminoácidos, um de um epitopo e outro de um paratopo, é uma aproximação da interação real e que possui um erro propagado para as predições dos próximos pares de aminoácidos. Depois de poucos pares de aminoácidos preditos estarem interagindo para formar uma estrutura tridimensional, começa a entrar em cena, por exemplo, a interação entre esses pares por meio de forças não-covalentes e limitações de giro das cadeias carbono alfa de aminoácidos denominadas limitações estéreo químicas (Ramachandran e cols., 1963). Suponha que devido às interações fracas dos pares de aminoácidos já preditos, percebe-se que a interação desses pares não está correta visto que tende a criar forças não covalentes ou ângulos de rotação da cadeia carbônica dos peptídeos que impediriam a conformação tridimensional do epitopo com seu paratopo. Porém, esse não é o fim da análise porque novas possibilidades de interação podem ser consideradas para as etapas anteriores. Várias abordagens são utilizadas para tentar solucionar esse problema, no entanto o método mais intuitivo seria recomeçar o processo de predição de interações entre aminoácidos desde o primeiro par. Dada a quantidade de variáveis que precisam ser recalculadas e a quantidade de possibilidades de conformação desses pares de epitopos, o problema torna-se inviável do ponto de vista computacional, apesar de que em teoria é possível testar todas as possibilidades de conformação desses pares de aminoácidos e chegar à solução ótima. Porém, os tempos de computação esperados mesmo com todo o poderio computacional da humanidade são maiores do que o tempo que os humanos existem no planeta Terra. É por esse motivo que o programa atualmente disponível para a predição de epitopos não lineares utilizando somente a estrutura primária dos aminoácidos tem um desempenho insuficiente e de pouca credibilidade (Blythe e Flower, 2005; Sun e cols. 2011).

O exemplo utilizado no parágrafo anterior referiu-se a uma tentativa de encontrar a interação entre um epitopo não linear e uma molécula de *MHC*. Porém uma proteína pode possuir quase tantas regiões possíveis de serem epitopos quanto a sua extensão em aminoácidos. Os programas atuais testam todas as regiões possíveis pela presença de epitopos (Hoof e cols., 2009; Stranzl e cols., 2010). Suponha, por exemplo, epitopos com o tamanho médio de 15 aminoácidos. Se uma proteína apresentar 500 aminoácidos então existirão pelo menos 486 possíveis regiões (500-15+1) que devem ser avaliadas quanto a presença de epitopos. Se o cálculo ótimo de uma interação entre epitopo e *MHC* consome tempo, realizar tal análise em todas as proteínas de um genoma é no mínimo um esforço de escalas quase que inimagináveis.

Outro problema que é mais percebido em epitopos lineares é justamente a quantidade de epitopos preditos por proteína que é elevado. O tamanho médio de epitopos lineares é de nove aminoácidos, porém existem epitopos lineares comprovados experimentalmente com tamanhos de oito aminoácidos (Stranzl e cols., 2010). Em sequências *fastas* aleatoriamente geradas para simular proteínas é possível obter sequências de pelo menos quatro aminoácidos que são idênticas as sequências de epitopos experimentais, contrariando a baixa probabilidade de tal situação ocorrer que é de 6.25e-06. Considerando-se um epitopo com oito aminoácidos, têm-se então pelo menos 50% de um epitopo experimental que foi aleatoriamente gerado. Além desses 50% idênticos ainda existe a possibilidade de que aleatoriamente também sejam gerados aminoácidos que impliquem em substituições conservativas quando confrontados com os epitopos experimentalmente comprovados. Isso significa que mesmo por acaso, uma proteína fictícia pode gerar sequências de aminoácidos próximas às sequências de epitopos válidos. Isso poderia explicar porque tantos epitopos lineares podem ser encontrados em proteínas não fictícias (Santos e cols., em preparação). A dúvida mais comum que surge quando são utilizados programas de predição de epitopos lineares é: “Qual epitopo utilizar de uma lista de dezenas ou centenas preditos em uma proteína, bem como com predições diferentes por alelos de MHC diferentes?” (Lin e cols., 2008).

Mediante tantos desafios da imunoinformática, nesse trabalho foi proposta uma contribuição no tocante a extrair alguma informação a partir da vasta gama de dados gerados por preditores de epitopos lineares. Foi percebido que ao invés de uma aparente confusão com tantos possíveis epitopos de uma proteína poder-se-ia procurar nessa gama de dados alguma pista sobre como o sistema imunológico de um hospedeiro interage com todos esses possíveis epitopos. Dessa forma, foi criada a estatística MED, assim como descrita na seção de resultados, que mostrou evidências sobre uma possível relação entre a quantidade de epitopos de um proteína e o papel da mesma na patogenicidade do agente etiológico de uma doença. Essa relação foi evidenciada devido ao fato da maioria dos candidatos elencados como os melhores pela estatística MED, possuírem comprovação experimental em *M. tuberculosis* de produzirem respostas imunes celulares ou seus mutantes originarem uma diminuição da virulência do patógeno. É esperado que não seja necessário gerar mais dados, mas apenas lê-los de maneira diferente em uma tentativa de romper com a visão tradicional de lidar com problemas científicos até então sem solução aparente.

3.2.6 Metodologias para pré-processamento e aferição de qualidade de agrupamentos de sequências biológicas

O artigo científico a seguir apresenta uma técnica da álgebra linear denominada decomposição por valores singulares, termo derivado do inglês *Singular Value Decomposition* (*SVD*). Essa técnica é conhecida pela capacidade de eliminar de um conjunto de dados, o conteúdo menos relacionado com a informação pertinente, conteúdo conhecido como "ruído". Uma maneira de entender o ruído é imaginar um imagem digitalizada. Mesmo após removermos vários pixeis dessa imagem, ainda é possível discernir o que a imagem está representando. A SVD é considerada uma técnica de pré-processamento de dados que atua reduzindo a dimensão de representação de dados por intermédio da redução da quantidade de atributos utilizados em uma matriz. Após um pré-processamento, os dados ficam aptos a serem utilizados por técnicas de Aprendizado de Máquina (AM). No caso específico de AM não-supervisionado, para o qual não existe informação prévia a respeito do melhor resultado possível, o pré-processamento para eliminação de ruídos mostra-se fundamental em problemas que envolvem agrupar elementos ou criar *clusters*.

Uma questão em aberto no uso da técnica de SVD é a quantidade de valores singulares utilizados nas matrizes, por exemplo, de distâncias filogenéticas entre espécies. Essa quantidade de valores, conhecida como *rank,* geralmente pode ser menor do que *n*, a quantidade de elementos da matriz de distâncias, mas o valor apropriado depende dos dados sob análise. Não existe uma fórmula matemática para determinar o melhor valor para *rank* de uma SVD, porém percebe-se que experimentar todos os valores singulares é computacionalmente viável desde que seja monitorada por uma métrica de qualidade também calculada em tempos computacionalmente viáveis. O uso de métricas de qualidade em técnicas de AM não-supervisionado pode ser puramente estatístico ou aproveitar algum conhecimento específico do domínio dos dados. Nesse caso optou-se por aproveitar o conhecimento biológico do domínio, na forma de classificação taxonômica de espécies. Utilizou-se uma base de dados de proteínas mitocondriais de 64 espécies classificadas de acordo com a taxonomia de Lineu para mensurar qualitativamente como os agrupamentos de espécies eram afetados pela variação da quantidade de valores singulares e pelo número de agrupamentos previamente especificados. Os resultados foram melhores do que o artigo científico que originalmente tentou criar árvores filogenéticas com essa mesma base de dados de proteínas mitocondriais (Stuart e cols., 2002) e isso deu origem à publicação apresentada a seguir. Após a constatação da viabilidade dessa metodologia que melhora o agrupamento de sequências biológicas, criou-se a possibilidade para utilizá-la, por exemplo, no agrupamento de proteínas cuja função ainda é desconhecida. Ao passo que se consiga agrupar proteínas de função desconhecida juntamente com proteínas de função conhecida, amparados por uma métrica de qualidade, têm-se uma evidência a respeito de sua funcionalidade. O próximo desafio será determinar uma métrica hierárquica de qualidade para proteínas não mitocondriais.

Santos *et al. BMC Genomics* 2011, **12**(Suppl 4):S11
http://www.biomedcentral.com/1471-2164/12/S4/S11

BMC Genomics

PROCEEDINGS Open Access

A singular value decomposition approach for improved taxonomic classification of biological sequences

Anderson R Santos[1†], Marcos A Santos[2†], Jan Baumbach[3], John A McCulloch[1], Guilherme C Oliveira[4], Artur Silva[5], Anderson Miyoshi[1], Vasco Azevedo[1*]

From 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010)
Ouro Preto, Brazil. 15-18 November 2010

Abstract

Background: Singular value decomposition (SVD) is a powerful technique for information retrieval; it helps uncover relationships between elements that are not *prima facie* related. SVD was initially developed to reduce the time needed for information retrieval and analysis of very large data sets in the complex internet environment. Since information retrieval from large-scale genome and proteome data sets has a similar level of complexity, SVD-based methods could also facilitate data analysis in this research area.

Results: We found that SVD applied to amino acid sequences demonstrates relationships and provides a basis for producing clusters and cladograms, demonstrating evolutionary relatedness of species that correlates well with Linnaean taxonomy. The choice of a reasonable number of singular values is crucial for SVD-based studies. We found that fewer singular values are needed to produce biologically significant clusters when SVD is employed. Subsequently, we developed a method to determine the lowest number of singular values and fewest clusters needed to guarantee biological significance; this system was developed and validated by comparison with Linnaean taxonomic classification.

Conclusions: By using SVD, we can reduce uncertainty concerning the appropriate rank value necessary to perform accurate information retrieval analyses. In tests, clusters that we developed with SVD perfectly matched what was expected based on Linnaean taxonomy.

Background

We developed a methodology, based on singular value decomposition (SVD), for improved inference of evolutionary relationships between amino acid sequences of different species [1]. SVD produces a revised distance matrix for a set of related elements. Our SVD-based computations provide results that are close to the internationally accepted scientific gold standard, Linnaean taxonomy.

The reason we chose this methodology is the proven capacity that SVD has to establish non-obvious, relevant relationships among clustered elements [2][3][4][5], providing a deterministic method for grouping related species. A distance matrix derived from SVD can be used by cladogram software to produce a "phylogenetic tree", yielding a visual overview of the relationships. We compared species grouping by this method with Linnaean taxonomy grouping and found that the species clusters were similar.

The rationale behind SVD is that a matrix A can be represented by a set of derived matrices [2], in the same

* Correspondence: vasco@icb.ufmg.br
† Contributed equally
[1] Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Av. Antônio Carlos, 6627, MG, 31.270-901, Brazil
Full list of author information is available at the end of the article

© 2011 Santos et al; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

way that a number can be derived into factors. One can also think of SVD as a set of matrices that provide numerically different representations of data without loss in semantic meaning, as for example representation in different base numbers. To understand the mathematical concept of SVD, suppose that ‘A’ is an array of real numbers or complex numbers composed of m rows by n columns. A matrix with a singular value decomposition of matrix A can be made:

$$A = U \Sigma V^T \tag{1}$$

where U is an orthonormal m x m matrix, and Σ is an m x n matrix, known as the diagonal matrix, with real and non negative numbers. The matrix V^T is known as a conjugate transpose, an n x n unit matrix with real or complex numbers. As the diagonal values of Σ are ordered in descending order, Σ is a direct function of matrix A and characterizes the singular values of this matrix, ordering them from the most significant to the least significant values. Considering a subset of singular values of size k<n, we can obtain A_k an approximate matrix of matrix A:

$$A_k = U_k \Sigma_k V_k{}^T \tag{2}$$

Thus, data approximation depends on how many singular values are used [6]. In this case, the number of singular values k is also known as the rank of matrix A_k, indicating how many lines and columns in matrix A_k are linearly independent. The possibility of extracting information based on less data is part of the reason for this technique’s success, as it allows data compression/decompression, with an execution time that does not increase exponentially with increasing matrix size, making analysis viable [6]. A data set represented by a smaller number of singular values than the original, full-size data set has a tendency to group data items that would not be grouped together if we used the original data set [2]. This could explain why clusters derived from SVD can expose non-trivial relationships among the original data set items [7]. In this paper we do not use the matrix A_k, product’s factorization by SVD to rank k; with only two arrays of SVD, the matrix D_k[3] is represented in the context of the matrix

$$A_k = U_k \Sigma_k V_k{}^T = U_k \left(\Sigma_k V_k{}^T \right) = U_k D_k \tag{3}$$

The justification for using only D_k is that it has k lines instead of m lines from A_k, so D_k is made up of linear combinations from U_k columns, which in turns provides the relationship $A \approx A_k \approx D_k$.

The main data set that we used was obtained from a previous study involving SVD [8], with 832 mitochondrial protein sequences from 13 families of mitochondrial genes, obtained from 64 vertebrate mitochondrial genomes. We organized these 832 sequences into 64 single FASTA sequences, each representing a single Linnaean species, concatenating the sequences of the 13 families of mitochondrial genes of each species. From here on, we will refer to this set of data as dataset1. Dataset1 consists of 64 highly-related species that have at least 8 of 14 Linnaean taxonomy levels in common with each other. As we also wanted to investigate how SVD parameters can influence cluster quality, we added 12 additional species to this data set, creating a second set of data, which we named dataset2 (Figure 1). We chose these 12 new species based on their high diversity, in order to create a less homogeneous data set; our objective was to determine whether SVD would separate non-related and related species into different groups. The species within dataset1 all belong to the same infraphylum (*Gnathostomata*), whereas the 12 new species that were included to increase diversity were selected from other phyla, also from the animal kingdom. The 12 species included in dataset2 were *Aphrocallistes vastus*, *Asterias amurensis*, *Aurelia aurita*, *Balanoglossus carnosus*, *Branchiostoma belcheri*, *Bugula neritina*, *Callyspongia plicifera*, *Candida albicans*, *Metridium senile*, *Ostreococcus tauri*, *Phallusia fumigata*, and *Unionicola foili.*

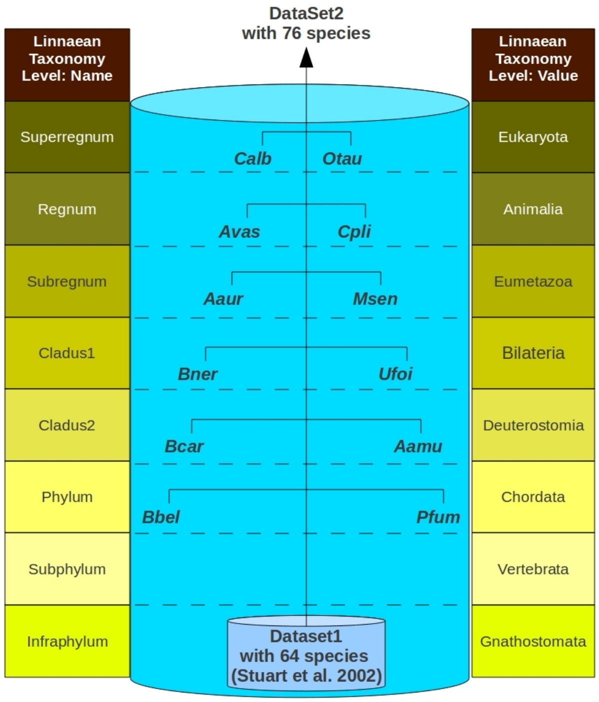


Figure 1 Dataset2 schema. Construction scheme for a set of species that were used as a negative control for the partitioning techniques.

The quality of the clusters that were generated was measured by the number of Linnaean taxonomy levels each species within the cluster bore in common with the other species; this was calculated as a function of an increasing rank value. When certain rank values are reached, larger values do not improve cluster quality, because there is no increase in taxonomic levels that the species have in common; in some cases a decrease is observed. The cluster quality obtained from a certain rank value maintains the number of shared common Linnaean taxonomy levels constant. This is evidence that there is an intrinsic relationship between these species that is mirrored in the distance matrix derived from these clusters; this quality helps build relevant cladograms.

Results and discussion

Singular value decomposition and number of clusters matters

In this study we give support to the hypothesis that choice of an appropriate data representation and a fixed number of clusters, combined with a good algorithm for categorizing this data, is sufficient for the production of biologically significant clusters. An A matrix has rank n, where n indicates the number of distinct species. The rank value (k) defines the degree of resolution of matrix D_k compared to the original matrix D, so k must be less than or equal to n. However, a k value close to n is undesirable, because one obtains a strong approximation to the original matrix D, which is useless to uncover relationships. We need to avoid this so-called 'noise data' [9] and find a smaller number of singular values that adequately represent the original data and thus achieve a reduction in the amount of data that needs to be processed [9][10]. We found that there is an optimal rank value that can be obtained by systematically testing all possible rank values and distances that define whether two species will form part of the same cluster, based on Linnaean taxonomic levels. A maximum distance value defining whether two species belong to a cluster can be experimentally found by increasing and decreasing an initial, empirically-defined distance, for example, the maximum distance between two species in a data set. We tried a systematic search for parameters that could confirm or deny this hypothesis. Working with singular value decomposition, one of the main parameters is the number of singular values necessary to create matrix decomposition sufficient to correctly separate all 76 species. This can be done by an algorithm called kdcSearch, which systematically examines possibilities for variation in singular values, Euclidean distance separation of clusters and number of clusters, a triad that we call kdc values. A systematic search to evaluate these three parameters proved to be computationally viable, independent of human intervention; it separates the target species into groups that represent similarity relationships between protein sequences and thus infer homology between species. The clusters generated through systematic choice of these parameters were biologically significant, demonstrating that we were on the right path in our attempt to determine the smallest number of singular values and the correct Euclidean distance that will correctly represent the original data, giving the correct separation of species groups. Based on these experiments we showed that even an "as simple as possible agglomerative clustering algorithm" (ASAP) can benefit from singular value decomposition to improve the quality of clusters that are generated. The next step was to use the parameters that were optimal [5] according to our methodology in other algorithms that have been thoroughly tested by the scientific community. The choice was made by K-Means [11], Expectation Maximization (EM) [12], Adaptive Quality-based Clustering Algorithm (AQBC) [13], K-Medoids [14], and MakeDensityBasedClusterer (MDBC) [15], since there is a statistically well-founded background, they have been widely used, and they are available as free software packages from R [16], Waikato Environment for Knowledge Analysis (WEKA) [15], and the JAVA Machine Learning Library [17]. The K-Means requires that an array of numbers be processed to calculate distances for the creation of clusters. It also opens the possibility of including a parameter that defines a fixed number of clusters to be created with the elements in the distance matrix. The same number of clusters inferred from the analysis done by ASAP, our in-house agglomerative clustering algorithm, was used by the K-Means algorithm. The K-Means implemented in the R statistical software, from now on called the K-Means-R algorithm, was parameterized for the initial number of elements, but not for specific elements. There is no such parameterizing in the K-Means implemented in the WEKA (K-Means-WEKA) software, making it possible that different results will be obtained with these two programs. We chose as the number of initial elements for calculating the first K-Means-R average half of the items or half of the species. The first run of K-Means-R was done with a matrix regarded as adequate because it had been generated with the parameters of the algorithm systematically observed ASAP, a rank value of nine and eight clusters. The algorithms EM, MBDC, K-Means-WEKA and K-Medoids were configured for eight clusters, without altering the other configuration parameters. The algorithm AQBC does not allow fixing the number of clusters, but we empirically tested parameters till we obtained the same number of clusters (eight). Then we looked for a way to compare the results from the various algorithms. At first glance it seemed that the result of, for example, K-Means-R was as good as the result from ASAP, but the large number of species and the not less considerable number of clusters made the comparison difficult. We needed a measure that would allow us to objectively compare the performances of the algorithms. Then we initiated execution of all algorithms with a number of singular values that represented the original array, without any reduction in the rank of the matrix decomposed into singular values. Despite minor variations in quality in some clusters, the overall quality of the clusters did not differ from the performance of all algorithms on a distance matrix generated with a reduction in rank. Table 1 shows quality calculations of eight clusters using ASAP and K-Means-R algorithms with different numbers of singular values. Clusters shown in this table are from the second round of trying to create smaller clusters, while maintaining correct separation of the *Aves* group (positive control), or the first recursive call of the ASAP algorithm. Both K-Means-R and ASAP were configured to generate eight clusters. Both algorithms used the matrix of the trigrams representing 8,000 combinatorial possibilities of 20 amino acids (20^3), also called N-gram with N=3, with 60 singular values (the original matrix, since all possible singular values were used) and another matrix derived from the former with only nine singular values; these quantities of singular values and clusters gave good SVD results in final clustering. The first column shows the cluster identification. The columns that follow are in groups of four, showing the results of K-Means and ASAP, using a trigrams frequency matrix created by SVD with 60 or nine singular values. The four columns under the label 'Number of species clusters Joined by' show the number of species obtained in each cluster. The four columns under the label 'Linnaean Taxonomy levels in common by clusters' show the number of Linnaean levels in common for each cluster, and the four columns under the label 'common Linnaean taxonomy frequency levels (cLtlf) by cluster' show the results of the metrics that we suggested. These come from multiplication of the column 'Number of species clusters Joined by' by the column 'Linnaean Taxonomy levels in common by clusters'. There were no significant differences (Student t test) in the quality of clusters generated by the algorithms, based on a comparison of the mean number of Linnaean levels in common and cLtlf, even though they used different singular values, as shown in the Additional file 1 (Figure S1). The Chi-square test did not demonstrate any significant relation between these four clustering rounds. However, there were significant differences between the algorithms and cluster data, based on cLtlf, as shown in Table 2. This table shows an alternative to measuring algorithm performance with different calibration parameters, using Linnaean taxonomy to infer cluster quality. The sum of the individual qualities of each cluster is measured by cLtlf. When cLtlf is weighted by the variation of this quality around the mean or standard deviation, the quality of results can be inferred through Linnaean clusters metric quality (Lcq). It is worth noting that clusters whose data are shown in Table 1 possessed a large number of taxonomic levels in common (60% of the levels that we used in this work). It is possible that so many in-common taxonomic levels left little scope for differentiation between the clusters, making the average quality very similar regardless of the method used for clustering. This result in the comparison between K-Means-R and ASAP was also observed in the results produced by the other algorithms that we tested. When there was a set of clusters with homogeneous qualities, it was necessary to find a measure that would discriminate the effectiveness of the algorithms with different numbers of singular values. Therefore, we used a measurement that takes into consideration the sum of the qualities of all clusters provided by a given

Table 1 Using the distance matrix that corrected separated *Aves* cluster: K-Means compared to ASAP

	Number of species joined by clusters				Linnaean Taxonomy levels in common by clusters				common Linnaean taxonomy levels frequency (cLtlf) by cluster			
Cluster	**K-means with rank 60**	**SNJ with rank 60**	**K-means with rank 09**	**SNJ with rank 09**	**K-means with rank 60**	**SNJ with rank 60**	**K-means with rank 09**	**SNJ with rank 09**	**K-means with rank 60**	**SNJ with rank 60**	**K-means with rank 09**	**SNJ with rank 09**
1	10	10	10	10	10	10	10	10	100	100	100	100
2	14	27	14	25	10	9	10	9	140	243	140	225
3	4	1	4	7	12	13	12	8	48	13	48	56
4	7	17	4	7	8	8	10	8	56	136	40	56
5	2	2	9	2	13	11	9	12	26	22	81	24
6	6	1	4	4	10	13	10	10	60	13	40	40
7	5	1	6	4	9	13	10	12	45	13	60	48
8	11	1	8	1	9	13	8	13	99	13	64	13

This table displays the results of K-Means and ASAP on a cluster of 60 species obtained in the first ASAP clustering round, when 76 species were separated into clusters.

Table 2 Inferring quality from clustering methods

Algorithm/ software	Rank	N	Min cLtlf	Max cLtlf	Mean cLtlf	cLtlf clusters sum (ΣcLtlf)	cLtlf standard deviation (σ)	Linnaean clusters quality (ΣcLtlf/σ)	Linnaean clusters quality gain (K09/ K60)%	cLtlf median	Median clusters quality gain (K09/ K60)%
AQBC-javaml	K09	8	32	180	71.25	570	52.27	10.90	49.58%	42.50	26.87%
	K60	8	0	220	64.38	515	70.64	7.29		33.50	
EM-weka	K09	8	40	120	70.12	561	31.53	17.79	48.99%	57.00	1.79%
	K60	8	16	160	70.25	562	47.06	11.94		56.00	
Kmeans-weka	K09	8	30	180	69.38	555	46.70	11.88	9.26%	61.50	-2.38%
	K60	8	16	180	69.88	559	51.39	10.88		63.00	
Kmeans-R	K09	8	40	140	71.62	573	34.48	16.62	9.21%	62.00	6.90%
	K60	8	26	140	71.75	574	37.72	15.22		58.00	
K-Medoids-R	K09	8	24	160	70.12	561	44.37	12.64	15.92%	60.00	13.21%
	K60	8	26	180	68.50	548	50.24	10.91		53.00	
MDBC-weka	K09	8	30	180	69.38	555	46.70	11.88	9.26%	61.50	-2.38%
	K60	8	16	180	69.88	559	51.39	10.88		63.00	
ASAP-in house	K09	8	13	225	70.25	562	67.68	8.30	27.51%	52.00	197.14%
	K60	8	13	243	69.12	553	84.92	6.51		17.50	

All evaluated partitioning's algorithms showed improved performance considering the Linnaean clusters quality when used the optimized distance matrix created by the better kdc parameters tested.

method, weighted by the variation in the quality of clusters. We analyzed this metric to look for significant differences between the two algorithms and the two numbers of singular values. The K-Means-R algorithm performance was two-fold better than that of ASAP. When we used an array of nine decomposed singular values, the number considered optimal for this set of data, in accordance with the methodology suggested here, K-Means and ASAP algorithms had 9 and 28% better performances, respectively, when compared to the original results from these methods, without singular value decomposition. The other algorithms that we tested also gave a significant increase in the quality of clusters in the results of the matrix decomposed into nine singular values and eight clusters, versus the non-decomposed matrix and eight clusters. In decreasing order, the increases in performance for each method were ~50% (AQBC), ~49% (EM), ~27% (ASAP), ~16% (K-Medoids), and ~9% (K-Means-WEKA, MDBC and K-Means-R). Despite the equal percentage increase for the algorithms K-Means-WEKA, MDBC and K-Means-R, the absolute quality values for K-Means-R were approximately 50% higher than those from K-Means-WEKA and MDBC, considering the distance matrices with and without decomposition by singular values. We chose the K-Means-R method for more detailed analysis of the results because this is a widely used algorithm and because in terms of absolute quality, it gave results very close to those from algorithm EM, which was the best in terms of absolute quality. These results have some details that are worthy of note. First, they show that in fact a matrix decomposed into a certain number of singular values, using a certain number of clusters, can create a representation of the original data with better quality than that obtained when we use the original data matrix (full rank). This reinforces the need for decomposition of a matrix into a smaller number of singular values for the removal of so-called 'noise' attributable to a full-rank array [9][10][18]. Second, the clustering algorithm was instrumental in generating good-quality clusters. It can be seen in Table 2 that the performance of K-Means-R and EM algorithms was two-fold better than that of the ASAP algorithm. Third, the method that we suggest here, to systematically explore the parameters needed to obtain the best performance of the K-Means proved essential to allow the K-Means to generate even better quality clusters. Fourth, the representation of a sequence of amino acids as a vector that stores the trigram frequency of 20 amino acids was effective to capture the levels of similarity between the sequences of the protein species that we analyzed, without incurring the problems that classical algorithms have with protein sequence alignments [19]. Fifth and finally, the quality metrics using the Linnaean classification suggested in this study were effective in measuring the quality of the biological significance of clusters constructed from mitochondrial proteins of dozens of species. Consequently, we conclude that when we use a smaller number of singular values to generate clusters, the quality of the clusters is significantly improved when compared with clusters generated with a matrix with all singular values, independently of algorithm. These results show that the combination of correct choice of algorithm, the number of singular values, the number of clusters and a quality metric with biological significance allows separation of species groups that are biologically meaningful. Furthermore, the use of trigrams of amino acids provides an effective way to determine similarity between protein sequences without using sequence alignment algorithms.

In the remainder of this paper, we show preliminary findings and methods that helped us reach our final conclusions, including how we arrived at an adequate number of singular values that allowed us to separate a set of species into groups with biological significance. To this end, we found that using arrays of trigram frequencies of amino acids to determine statistical properties was as good as using 4-gram frequencies [19]. We show that the size of the sequences that are analyzed can affect the separation of elements into clusters. We also present measures that allow us to infer the biological significance of a cluster and measure the quality of the clustering methods compared to Linnaean taxonomic classification of species.

Algorithm kdcSearch: parameterizing rank and number of partitions

The objective of the algorithm kdcSearch (Figure 2) is to identify a 'k' rank value and a quantity 'c' of partitions that promote correct separation of species, based on biological significance according to Linnaean taxonomy. This 'k' rank is responsible for the reducing the dimensions of the data that hide evolutionary relations among species, also known as data noise. A quantity of partitions 'c' should correctly separate the positive control group from the other species and possibly separate the other species into partitions with evolutionarily-significant relationships. In this algorithm, the number of partitions 'c' is a function of 'd', that is c=f(d), with 'd' being the Euclidian distance between elements in a symmetric matrix of distances between the species. The value of 'd', on the other hand, varies according to the distance matrix created with rank 'k', establishing the relations d=f(k) and c=f(f(k)). In this way, we look for the 'k' value that will eliminate data noise and generate a distance 'd' responsible for creating a number of partitions 'c' with the greatest capacity to infer evolutionary relationships between grouped species. In this process, we use an in-house algorithm, ASAP, to partition the species. Considering the random selection of pivotal elements for the creation of partitions by ASAP, it is not possible to estimate a priori what distance 'd' will create a number 'c' of partitions. Consequently, a systematic search is made with a range of X and Y values of Euclidean distances in order to determine which distances give what number of partitions. This algorithm begins with an X value equal to the largest Euclidean distance between species represented in a symmetric matrix of distances with a maximum value; that is, it has not undergone decomposition by singular values and for this reason k is equal to the total 'n' of the number of species in this matrix. Considering the maximum distance of the symmetric matrix, only one partition with 'n' species is formed. The value Y is always zero, the point at which the search for partitions of value 'k' terminates and 'n' partitions of the data are always formed, with only one species per partition. With the values in hand (k1, k2, k3, ..., kn) with their respective values (c1=f(d1) , c2=f(d2), c3=f(d3), ..., cn=f(dn)), in turn with their respective biological significance levels, measured by the function cLtlf, we filtered the configuration that gave the correct separation of the control group and the greatest number of partitions of species sharing the highest possible numbers of Linnaean classification levels. This algorithm is recursive because if no group of these variables provides a partitioning of the control group isolated from the other species, then the algorithm did not yet find a solution with the desired level of Linnaean taxonomic relationship. In order to simplify, it is not necessary to analyze all of the possible numbers of partitions; analysis is made only within well-defined intervals. Taking as an example, dataset2 with 76 species, an alternative is to analyze the number of partitions

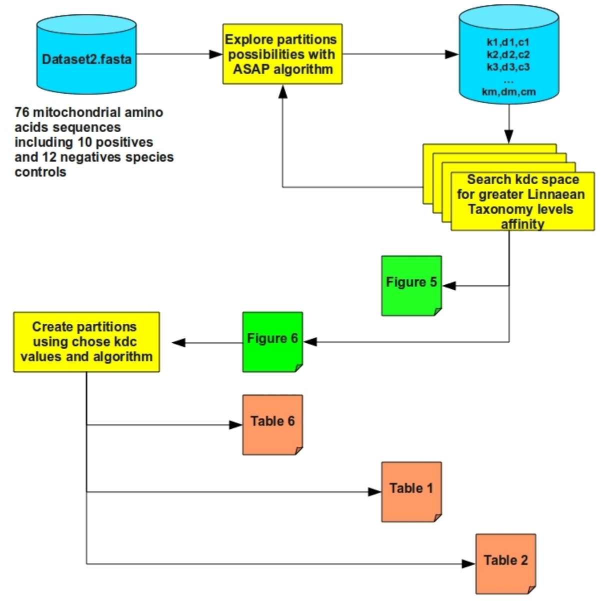

Dataset2.fasta
Explore partitions possibilities with ASAP algorithm
k1,d1,c1
k2,d2,c2
k3,d3,c3
...
km,dm,cm
76 mitochondrial amino acids sequences including 10 positives and 12 negatives species controls
Search kdc space for greater Linnaean Taxonomy levels affinity
Figure 5
Create partitions using chose kdc values and algorithm
Figure 6
Table 6
Table 1
Table 2

Figure 2 kdcSearch algorithm schema. Main procedures, datasets and products. Multiple rectangles mean recurring calls.

containing groups of three (c3, c6, c9, ..., c75). This example can be obtained from the algorithm below through the initialization of a variable that, as it divides the total number of species by 25, permits the creation of an incremental step of three levels between analyses. The number was determined empirically and the algorithm below is adduced by the variable EDRD (Empirical Dimensional Range Division).

Table 3 Linnaean taxonomy levels

Linnaean Taxonomy levels		
Number	**Name**	**Value**
14	*Species*	*Aythya americana*
13	*Genus*	*Aythya*
12	*Familia*	*Anatidae*
11	*Ordo*	*Anseriformes*
10	*Subclassis*	*Carinatae*
9	*Classis*	*Aves*
8	*Infraphylum*	*Gnathostomata*
7	*Subphylum*	*Vertebrata*
6	*Phylum*	*Chordata*
5	*Cladus2*	*Deuterostomia*
4	*Cladus1*	*Bilateria*
3	*Subregnum*	*Eumetazoa*
2	*Regnum*	*Animalia*
1	*Superregnum*	*Eukaryota*

Linnaean taxonomy levels used to classify the species in this paper. The numbers denote an increasing degree of nomenclature specialization.

When one of the recursions of the algorithm kdcSearch finds one or more groups of variables k, d and c that give correct separation of the positive control group, the algorithm recursions are finalized. In this case, there is no reason to continue making recursions, since the desired level of cohesion for the elements of the partitions has reached its limit, measured by the positioning of the positive control. In the case of the data that we analyzed here, this situation occurs after the end of the first recursion by the algorithm kdcSearch, culminating in the plotting of the final graphs and implementing the function ‘Finalize’. The code for the function ‘Finalize’ was left open because at this stage of execution, the algorithm finds various groups of the variables k, d and c (kdc) that promote correct separation of the positive control group in a partition separate from the other species. At this point, the question is which group of values kdc is a good result. What differentiates one group of variables kdc from another is the quality of the partitioning of the other species compared with Linnaean taxonomic classification. We think that it would not be useful to develop an algorithm that one particular kdc group is better than others because they give different levels of separation of species. A researcher can be trying to separate a group of species at the level of ‘Classis’ with nine Linnaean levels in common (Table 3), while another researcher may try to separate this same group at the ‘Ordo’ level, with 11 Linnaean levels in common. Consequently, it would be reasonable to consult the last table generated by the algorithm kdcSearch to adjust the result to the necessities of a specific objective. However, in case the final objective is not well defined, an option to completely automate this process could be to compare the partitioning medians for each kdc group with which it was possible to separate integrally and isolatedly the positive control species group. This comparison creates an estimate of the cohesiveness of the partitionings based on comparison with Linnaean taxonomic classifications. Values of kdc that give larger medians would be chosen as superior, promoting partitionings with greater biological significance. The rationale that explains the use of the median as a parameter for the procedure ‘Finalize’ can be better comprehended by analysis of the data in Tables 4 and 5. These tables show the cLtlf results for nine sets of kdc values that by definition are good results because they can separate integrally and isolatedly the positive control group. In the partitionings produced with these kdc values, there is always a partition of the positive control group with a cLtlf equal 100 (10 species sharing 10 Linnaean levels). Values of kdc that cannot optimally separate the positive control group from the other species were also included. The set of kdc values used as a negative control in this analysis is suffixed with the symbol ‘(-)’. In Table 4, we can see that the kdc sets that have many partitions with only one isolated element (cLtlf=1*13=13 or the minimum cLtlf) reduce the median cLtlf value for all of the partitions produced in this set by the respective set of kdc values. The intention of these partitionings is to demonstrate evolutionary relationships among species; the kdc values that give large numbers of partitions with only one element each do not give much information about such relationships. Consequently, it is understood that the best kdc values are those that have the fewest species isolated in partitions with only one element. Table 5 also shows the application of the measurement ‘Linnaean cluster quality’ to the partitionings based on these kdc values; however, this measure was not effective in indicating how informative the partitionings for each group of kdc values were in terms of the relationships based on Linnaean classification. It can be seen that the kdc values of the negative control had larger ‘Linnaean cluster quality’ values than the various sets of kdc values that adequately separated the positive control group. Apparently, ‘Linnaean cluster quality’ is not efficient at classifying kdc values at this ‘Finalize’ step of the algorithm search, though it is efficient while the positive control group has not been integrally separated in an

Table 4 Function Finalize: sample data

06clusters k03	**06clusters k06**	08clusters k06	**08clusters k09**	*08clusters k12(-)*	08clusters k45	*10clusters k30(-)*	12clusters k12	14clusters k18	14clusters k21	*14clusters k36(-)*	14clusters k60
100	**100**	100	**100**	*248*	100	*144*	100	100	100	*88*	100
243	**243**	200	**225**	*13*	243	*252*	216	240	250	*240*	220
45	**64**	56	**56**	*180*	13	*13*	13	13	13	*13*	13
96	**100**	13	**56**	*30*	136	*13*	64	90	88	*96*	112
13	**13**	100	**24**	*13*	22	*56*	22	22	22	*22*	22
40	**40**	45	**40**	*32*	13	*13*	13	13	13	*20*	20
		24	**48**	*13*	13	*13*	16	16	13	*13*	13
		40	**13**	*13*	13	*13*	24	24	13	*13*	24
						13	40	40	30	*13*	13
						13	48	13	13	*13*	13
							13	13	13	*13*	13
							13	13	13	*13*	13
								13	13	*13*	13
								13	13	*13*	13

The statistic cLtlf for all of the partitionings of species obtained with nine kdc values that separate the positive control group in the function Finalize of the algorithm kdsSearch, along with three kdc values as a negative control (-).

isolated partition. However, based on the median, the sets of kdc values that do not separate the positive control group into isolated clusters were correctly classified as being of low quality based on Linnaean classification, as well as other kdc values that had many partitions with the lowest Ltlf. In Table 5, the kdc values with the largest medians are in bold, and the kdc values that do not adequately separate the positive control group are in italic. It is relevant to point out that though some kdc values can adequately separate the positive control group, many partitions have the minimum cLtlf; these were responsible for the low kdc values, values even lower than some kdc values that do not adequately separate the positive control group. In this study, we decided to analyze in more detail the partitions created by the kdc values with eight partitions and rank nine, which produced the third best median result without separating many species into isolated partitions. This choice is justified by the fact that these kdc parameters make the correct separation of the mammals ‘Hsap’ and ‘Ppya’ in a partition separate from those of the other species. These two species were used as a second positive control group. In the set of kdc values with six partitions and rank six, the configuration classified as having the best median, these two species are in a partition with 25 other species. Another option would be to use the kdc values with six partitions and rank three, which were classified as the second-best median. In this kdc configuration, ‘Hsap’ is isolated in a partition, while ‘Ppya’ is in a partition with 26 other species. Accordingly, the kdc values that give eight partitions with rank nine promote correct separation of the two positive

Table 5 Function Finalize: sample statistics

ASAP/ Clusters	Rank	N	Min cLtlf	Max cLtlf	Mean cLtlf	cLtlf clusters sum (ΣcLtlf)	cLtlf standard deviation (σ)	Linnaean clusters quality (ΣcLtlf/σ)	cLtlf median
06clusters	**K03**	**6**	**13**	**243**	**89.50**	**537**	**82.46**	**6.51**	**70.50**
06clusters	**K06**	**6**	**13**	**243**	**93.33**	**560**	**80.81**	**6.93**	**82.00**
08clusters	K06	8	13	200	72.25	578	60.65	9.53	50.50
08clusters	**K09**	**8**	**13**	**225**	**70.25**	**562**	**67.68**	**8.30**	**52.00**
08clusters(-)	*K12*	*8*	*13*	*248*	*67.75*	*542*	*92.41*	*5.87*	*21.50*
08clusters	K45	8	13	243	69.12	553	84.92	6.51	17.50
10clusters(-)	*K30*	*10*	*13*	*252*	*54.30*	*543*	*81.02*	*6.70*	*13.00*
12clusters	K12	12	13	216	48.50	582	59.10	9.85	23.00
14clusters	K18	14	13	240	44.50	623	63.29	9.84	14.50
14clusters	K21	14	13	250	43.36	607	66.12	9.18	13.00
14clusters(-)	*K36*	*14*	*13*	*240*	*41.64*	*583*	*63.66*	*9.16*	*13.00*
14clusters	K60	14	13	220	43.00	602	60.68	9.92	13.00

Comparison of the Lcq values and the ΣcLtlf medians for partitionings of the species obtained in Table 4.

control groups and were responsible for significantly improving performance in the statistic 'Linnaean cluster quality' and in most of the medians of the partitioning algorithms that were tested (Table 2).

From 76 to 60 species and eight clusters

We decided to use a 76 species data set (dataset2), incorporating 12 species that were less related to the original group, in order to develop relationship trees that included clusters with distantly related species. The 64 species data set (dataset1) from the study by Stuart contains closely related species, as all of them share 8 of the 13 Linnaean taxonomy levels used in our study to differentiate species [5]. When a correct fit was made (15 imposed clusters and rank value of 39), we were able to separate 60 of the 64 species and the additional 12 species using ASAP (Figures 3 and 4). These 12 added species plus four of the original species from dataset1 did not group into a single cluster. Instead, we obtained several different clusters, most of which included only one species.

Analyses were then carried out on only the 60 species from the data set that were joined as a single cluster; the ASAP algorithm was run with 15 clusters and a rank value of 39. When the ASAP algorithm was run with the original 64 species data set, some elements were separated into isolated clusters despite actually sharing several Linnaean taxonomy levels in common with all of the other species.

This could be due to the fact that mitochondrial protein sequences for some species within the data set used in this study were not available. Since our algorithm only uses the frequency of occurrence of amino acid triplets, a lower frequency can affect the quality of the clusters that are generated, as does the presence or absence of a triplet sequence. Presence or absence of amino acid triplets are also responsible for early cluster separation of the 12 additional taxonomically distantly related species, incorporated into the original 64 species data set. Consequently, we worked with this 60 species data subset. To do so, we included a recurrence step prediction in our algorithm in order to develop a species subset. We worked with the concept that a good separation of species in clusters distributes the elements in groups of more than one element, whereas a group with only one element gives no information about species ancestry. When we correctly separated the *Aves* group in an isolated cluster, we assumed that other groups should also be close to divisions that have evolutionary significance. Finally, a good separation involves having *Aves* isolated in a single group, while having the largest possible number of other species together in groups, with a few isolated species in groups of only one element (Figure 5, rank value of 39 and 15 clusters). This definition of good separation between species is applicable only when it is not possible to isolate the positive control group, the *Aves* group. But when we split a homogeneous positive control group, the concept of a

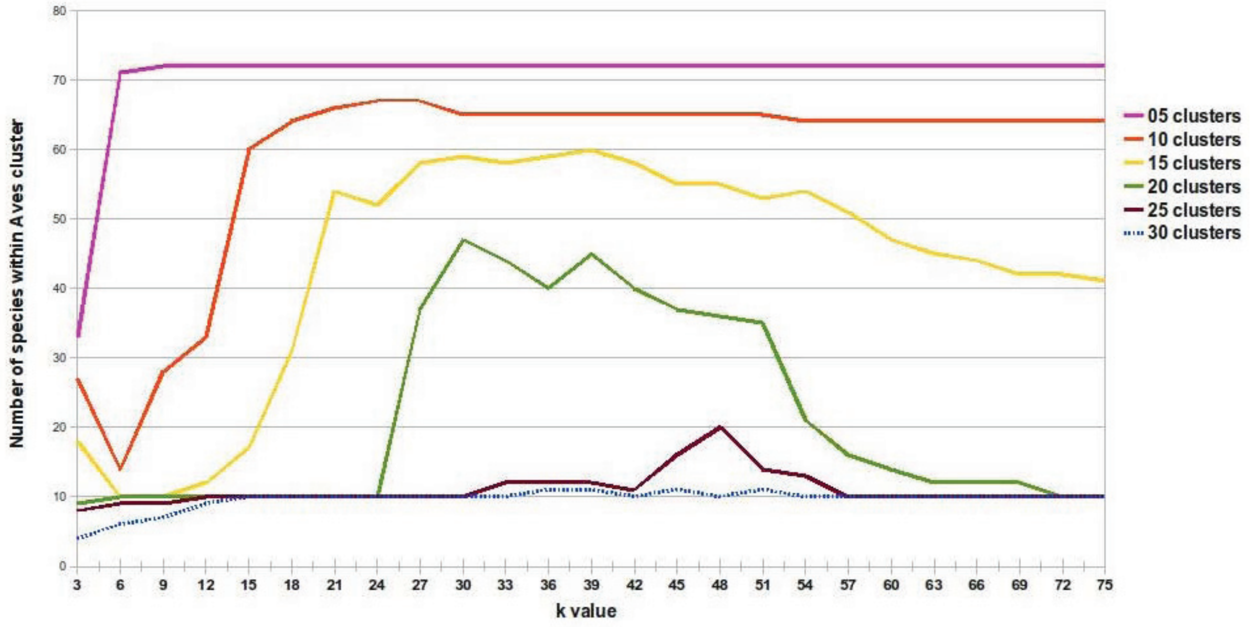


Figure 3 Exploring the number of species in the *Aves* cluster. The number of species grouped into the Aves cluster as a function of rank value and number of clusters. Ordinates are multiplied by the respective maximum Linnaean taxonomy levels shared by species in Figure 5.

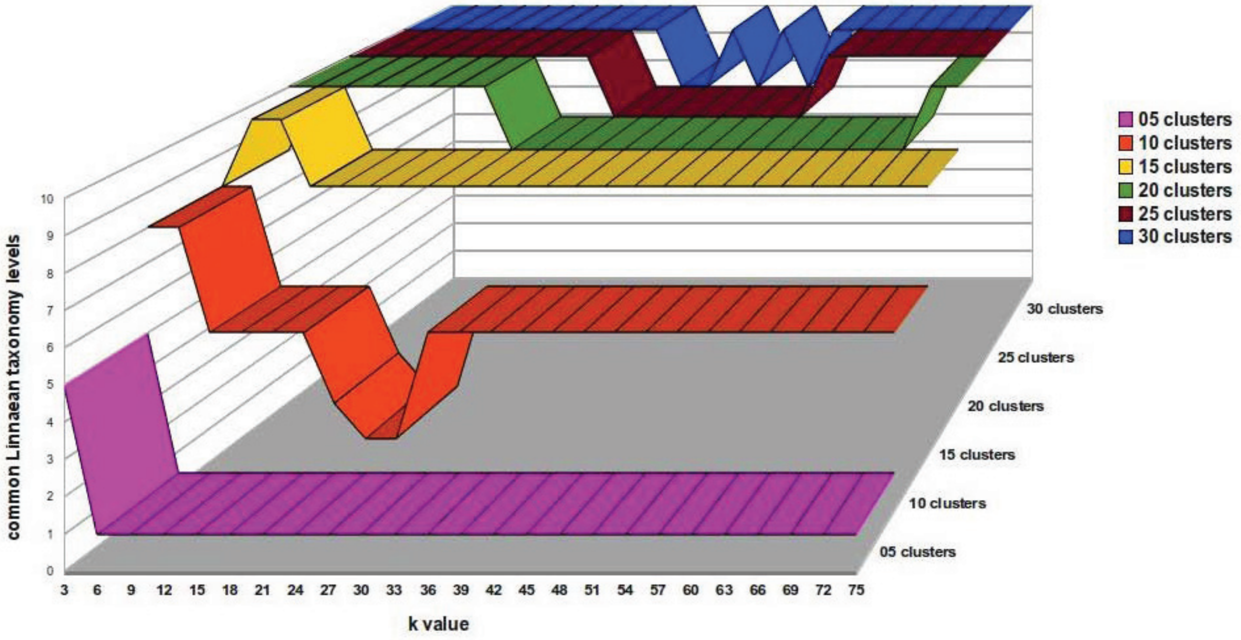


Figure 4 Exploring *Aves* cluster with maximum shared linnaean taxonomy levels. The number of Linnaean levels shared by all species is plotted against rank value and number of imposed clusters. Ordinates are multiplied by the respective number of species that produced Figure 5.

good separation of species is altered and it changes the way we interpret the graph of rank value versus cLtlf in the first recursive call of the algorithm (Figure 6, rank value of nine and eight clusters). In Figure 6, a high cLtlf value means poor cluster quality, because at this level of recursion it is possible to isolate the positive control group in a single cluster and leave few species in isolated groups. Therefore the optimal value for the

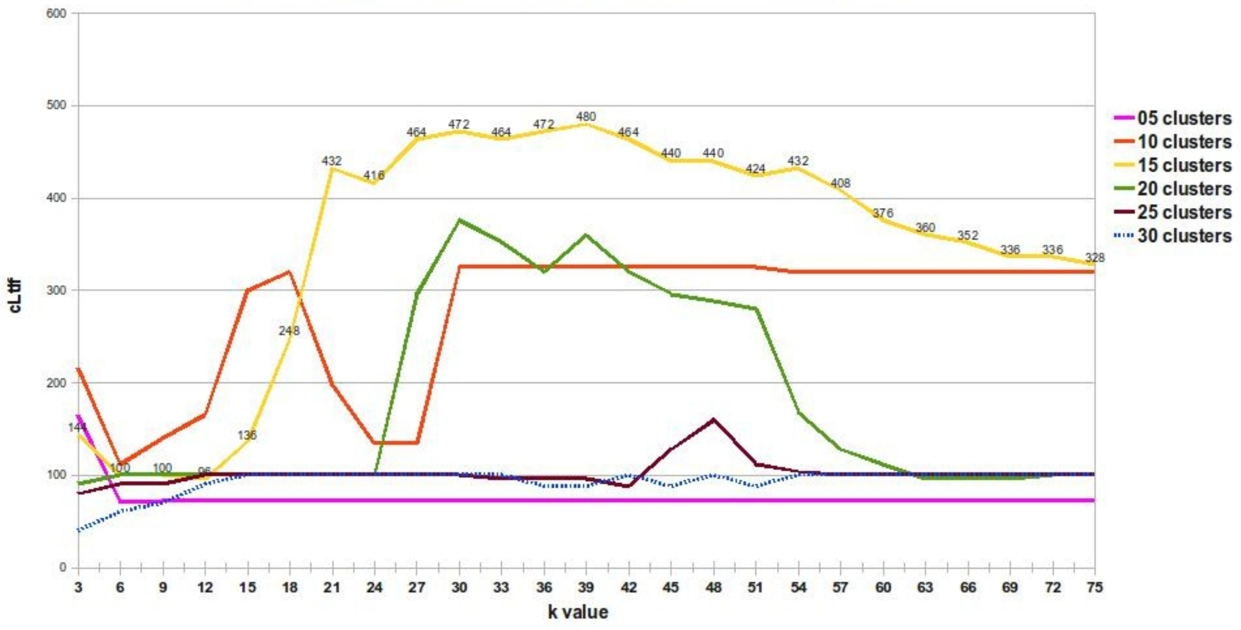


Figure 5 Determining the best algorithm parameters. Aves cluster quality as a function of rank value and different numbers of clusters. The number of clustered elements multiplied by maximum common Linnaean taxonomy levels shared between species gives the quality measure.

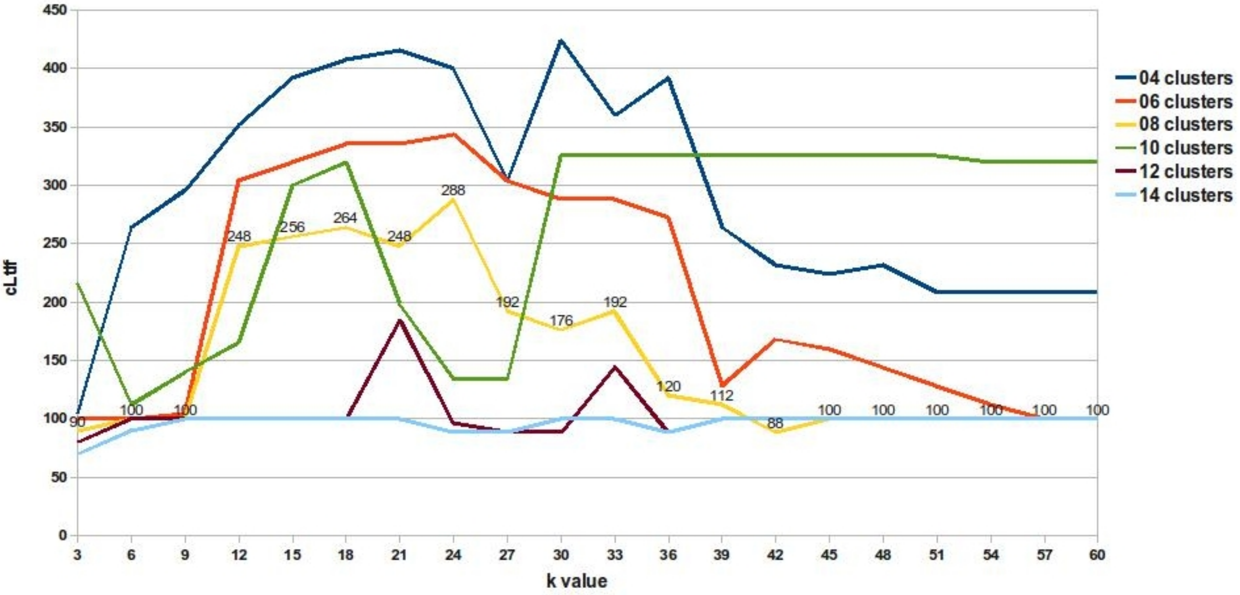


Figure 6 Determining the best algorithm parameters at the first algorithm recurrence step. Aves cluster quality measured with a reduced numbered of species than in dataset2. Now is possible to cluster the Aves species separately and the best algorithm adjustment to this cluster is preferred. Higher curves do not represent better quality.

separation of the group of *Aves* is 100 (10 species * 10 Linnaean levels in common). A value larger or smaller than that gave inappropriate separation, because the positive control is the group of birds with 10 species sharing 10 levels of Linnaeus. Using a second positive control group (’*Hsap*’ and ‘*Ppya*’), we concluded that using eight partitions with rank nine is the best configuration, correctly separating the birds group, creating groups with evolutionary significance and decreasing the number of species in groups of only one element. We used a rank value of nine to create the unrooted tree shown in Figure 7. ASAP was calibrated with a d value that produces eight clusters using the experimented rank value of nine. This choice was made based on obtaining a good separation result, when grouping all species of the *Aves* class into a single cluster, plus a positive control group.

The results of the first execution of our recurrence algorithm based on the 60 species data set can be seen in Table 6. Clusters 2, 5 and 8 are comprised of species of theclass *Mammalia*. Cluster number 5 includes the hominids *Homo sapiens* and *Pongo pygmaeus*, which were together, separated from other mammals due to their mitochondrial protein sequences sharing 12 common Linnaean taxonomy levels. Clusters 2, 5 and 8 were composed of only mammalian species, sharing 9, 12 and 13 common Linnaean taxonomy levels, respectively. It is evident that the number of clusters and rank value used to create distance matrices enables even ASAP to provide adequate clustering based on quality discrimination. All the clusters that were obtained are shown in Figure 7, in which four mnemonic letters represent each species.

Conclusions

Clusters and cladistic trees drawn from distance matrices, which were generated with SVD, showed a good correlation with Linnaean taxonomy. Considering the best estimate, when a difference is found, this does not necessarily mean strong divergence from taxonomic methods, but perhaps a more accurate picture of the relationship between the species that clustered together. This was demonstrated by clusters that were separated from mammalian clusters due to their greater protein sequence relatedness. It also was reinforced by Linnaean taxonomy information.

The similarity between clusters generated by our distance matrix and Linnaean taxonomy is indicative that distance matrices generated by SVD can demonstrate evolutionary relationships of species and construct better quality clusters and phylogenetic trees. These clusters and phylogenetic trees would benefit from amino acid trigrams and the Euclidean distance property of displaying a distance proportional to the number of necessary edits needed to perform a global alignment sequence within a polynomial execution time.

Methods

Datasets

The set of species used in this work is not original [8]. We opted for using a previously known set of data to allow comparisons with other studies that also use this

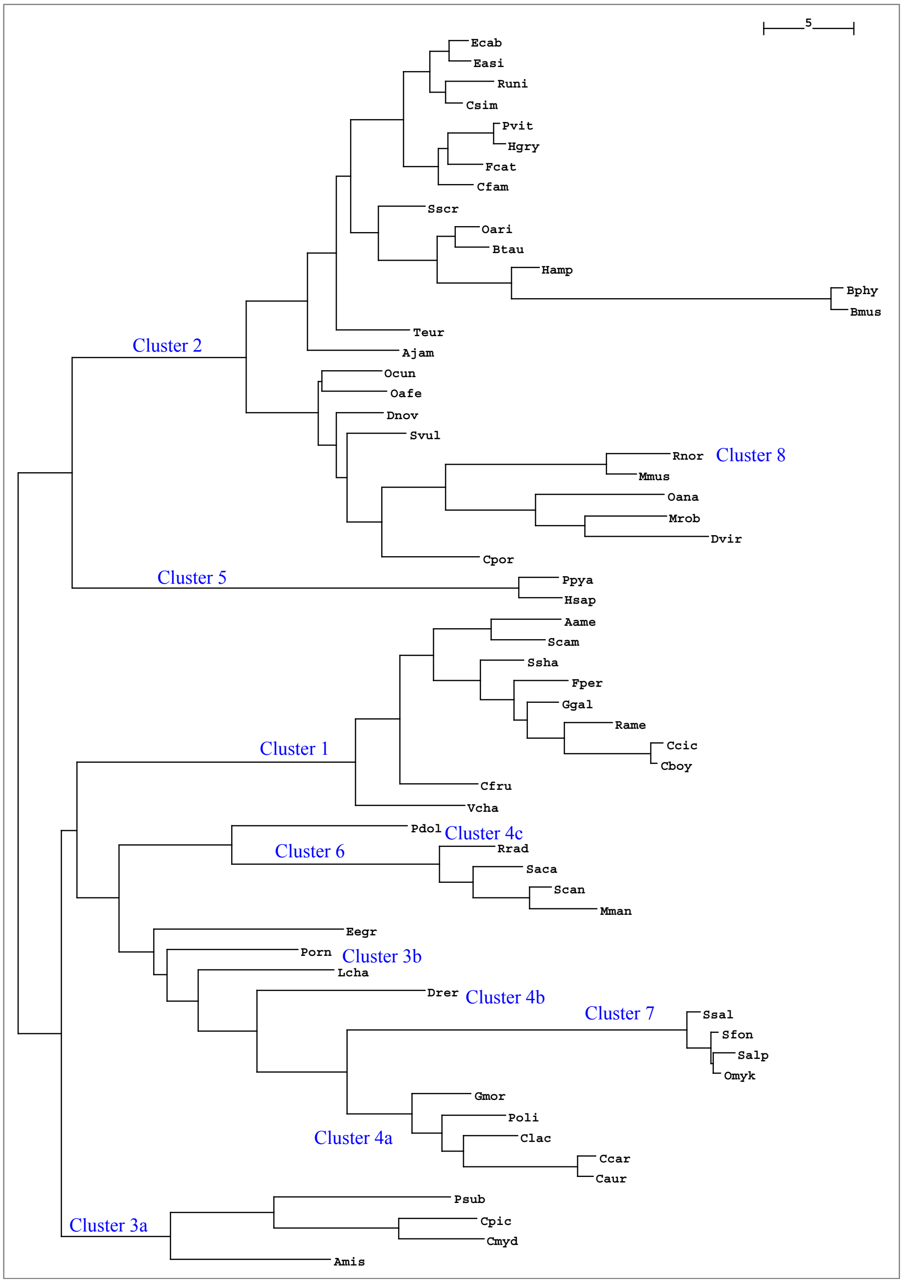


Figure 7 60 species from the Stuart data set. A 60 species data set unrooted tree generated from a distance matrix created with the ASAP algorithm. The original algorithm from this paper provided the distance matrix. Blue labels denote clusters.

data. We named this set of 13 mitochondrial proteins from 64 vertebrate species, dataset1. Within dataset1, a group of 10 species belonging to the class *Aves* was chosen to be the positive control group. We developed a negative control group with mitochondrial protein from 12 other species. Joining the proteins from these 12 species with the 64 in dataset1 gave origin to dataset2. Figure 1 schematically represents dataset2 as a set of data composted of dataset1 and 12 additional species. These 12 additional species were selected based on the criterion of being at least one level above the Linnean level common to all of the species in dataset1. Two species were randomly selected for each Linnaean taxonomic level, from *Phylum* to *Superregnum.* The same 13 mitochondrial proteins from dataset1 were selected for these 12 additional species. The additional amino acid sequences were obtained from the NCBI site. The union of these 13 mitochondrial proteins from the 12 new species with the sequences in dataset1 gave origin to dataset2, which includes positive and negative control groups of species. In order for a partitioning method to be successful, the positive control group needs to stay together in a partition and no other partition can be contaminated by the negative control group.

Positive control group and statistics

In order to show how rank values and the number of imposed clusters affect SVD, we ran ASAP algorithm with different rank values and numbers of clusters. Figure 5 shows the results of these runs for a single cluster, the cluster denominated cluster 1, which contains species belonging to the Linnaean taxon, the *Aves* class. This taxon is ideal for testing our hypothesis, because few and closely related species within the data we used belonged to this taxon. Furthermore, the *Aves* species in our data set tended to mix with less evolutionarily related species when the algorithm was incorrectly calibrated or the number of clusters was too small. For evaluating the quality of the cluster generated, we considered the product of common shared Linnaean taxa among clustered elements multiplied by the number of clustered elements. This indicator gives us a good measure of cluster quality, as it assesses the frequency of commonality within the cluster. Here, we denominated this indicator as "common Linnaean taxonomy level frequency", or cLtlf, and used it to show how cluster quality can vary as a function of the rank value or the maximum number of clusters used. Figure 5 shows the quality of cluster 1 generated by the algorithm, as rank value increases when different numbers of clusters are used to group the entire 76 species data set.

Figure 5 shows that, independent of the maximum number of clusters chosen to represent the 76 species data set, an increase in rank value does not improve cluster quality; consequently, we can safely use a considerably smaller number of singular values than the theoretical maximum. It is possible to roughly estimate an optimal value for rank value from this particular data set. If we consider 15 clusters, a rank value over 39 will not dramatically increase the quality of each cluster (Figure 5).

When we evaluate cluster quality measured by cLtlf, (Figure 5), we see that there is no significant improvement in cluster quality beyond the rank value of 39. This rank is sufficient for a good data representation of our original data set. Also, within cluster 1, the number of elements clustered together and the number of Linnaean taxonomy levels in common as a function of rank value, can be seen, respectively, in Figures 6 and 7. The maximum number of Linnaean taxonomy levels in common within cluster 1 obtained was 10. There is another interpretation for this graph in Table 3, associating these 10 levels in common within the cluster with the 14 Linnaean taxonomy levels considered in our study. This shows that the stringency of the data representation provided with SVD is sufficient to infer Linnaean taxonomy levels. On the other hand, if a less stringent fit is used, such as with an inappropriate number of clusters and rank value, a panoply of unrelated species are included in a cluster. It must be pointed out that our main task in this study was to learn and exemplify the calibration of our algorithm in order to retrieve desirable information. With the data set we used, the desirable information to be retrieved was Linnaean taxa, however, with other data sets this calibration should be tuned to direct the desired objective.

Table 3 characterizes a bird species, *Aythya americana*. The taxonomy levels shared by cluster 1 species in our algorithm executions with 20, 25, and 30 clusters and rank value 24, are levels lower than level number 11, namely the order (Ordo). Levels numbered as 11 (order) and 12 (family) were not shared among the 10 bird species in the data set. As more non-Aves species are added to this bird set, there is a decrease in cluster quality.

Euclidean distance

We can produce a distance matrix that contains a measure of how each species is related to each other. To construct this matrix, each species rank values set is treated as a vector in a k-dimension space. One can

Table 6 Eight clusters from 60 data set

Cluster	Number of species joined	Linnaean taxonomy levels in common	Deepest Linnaean taxonomy level
1	10	10	*Carinatae*
2	25	9	*Mammalia*
3	7	8	*Gnathostomata*
4	7	8	*Gnathostomata*
5	2	12	*Hominidae*
6	4	10	*Elasmobranchii*
7	4	12	*Salmonidae*
8	1	13	*Rattus*

Eight clusters created from the first recurrence algorithm execution calibrated with a rank value of nine. Species were grouped according to their deepest evolutionary relatedness based on Linnaean taxonomy levels. Clusters 2, 5 and 8 belong to the mammalian class.

choose the best measure to calculate the distance among vectors, depending on the particular characteristics in a data set. We decided to use Euclidean distance instead of the cosine distance used by Stuart [8]. This is because there is data indicating that Euclidean distance produces better cluster quality results than cosine distance. There is evidence [20], using the same 64 species data set that we present here, that Euclidean distance is proportional to the number of editions needed to perform a global sequence alignment. Consequently, it gives a more accurate measure of evolutionary relatedness than cosine distance, without the need for a global alignment sequence. There is evidence that the superiority of this Euclidean distance calculation is due to intrinsic evolutionary differences that affect the size of vectors. This is easy to see when one considers two vectors with the same cosine distance but with significant differences in length.

ASAP algorithm: in house agglomerative clustering

We implemented a clustering algorithm that was called ASAP (As Simple As Possible) and showed that even a naive algorithm can benefit from data adequately treated by SVD. Thus, it is not our intention to demonstrate it's worth using this clustering algorithm, but we want to leave the message that regardless of the algorithm, it is worth using SVD conjugated with positive controls in information retrieval, as an initial filter against noise [10][18].

ASAP is an algorithm designed to facilitate the work of measuring the impact of using SVD in clustering algorithms. This algorithm somewhat resembles single-linkage clustering; the differences are that no clustering starts from the two elements with the lowest Euclidean distance. Clustering starts with a random element; also, a new entry is not inserted in the matrix of Euclidean distances for each cluster created between the algorithm interactions.

The idea is quite simple; randomly select a species from the distance matrix, cluster together with other species according to a fixed 'd' distance and remove the clustered species from the distance matrix. Do it again randomly selecting other species, and so on.

(1) Repeat as long as the number of columns in the distance matrix is greater than one:

1.1. Fix the first column as the pivotal element;

1.2. Create a cluster of elements so that the Euclidean distance is smaller than a 'd' value for the pivotal element;

1.3. Remove elements from the novel cluster (lines and columns) from the distance matrix;

1.4 End repeat.

This algorithm was implemented using Scilab1 5.2.1 run on GNU linux Ubuntu, core 2.6.22-16. This implementation is available in the Additional file 2, accompanied with data and raw results.

Clustering algorithms evaluated

K-Means-R

The K-Means algorithm implemented [11] in the R statistical software aims to partition points into k groups such that the sum of squares from points to the assigned cluster centers is minimized. At the minimum, all cluster centers are at the mean of the set of data points which are nearest to the cluster center [16].

K-Means-WEKA

The K-Means algorithm implemented in the WEKA software is denominated SimpleKMeans. This implementation can use either the Euclidean distance or the Manhattan distance. If the Manhattan distance is used, then centroids are computed as the component-wise median rather than mean [15].

Expectation Maximization (EM)

The EM algorithm [12] creates partitions assigning a probability distribution to each instance. EM can decide how many clusters to create by cross validation, or is possible to specify apriori how many clusters to generate [15].

Adaptive Quality-based Clustering Algorithm (AQBC)

It's a heuristic iterative two-step algorithm with computational complexity approximately linear. The first step consists in finding a sphere in the high-dimensional representation of the data where the density of expression profiles is locally maximal. In a second step, an optimal radius of the cluster is calculated based only on the significantly coexpressed items which are included in the cluster. By inferring the radius from the data itself, there is no need to find manually an optimal value for this radius by trial-and-error [13].

K-Medoids

It's an exact algorithm based on a binary linear programming formulation of the optimization problem [21], using 'lp' from package 'lpSolve' as solver [16]. Probably is not possible to obtain clustering solutions depending on available hardware resources due to the quadratic order of the program. The K-Medoids R implementation is an NP-hard optimization problem. Partitioning Around Medoids (PAM) [14] is a very popular heuristic for obtaining optimal K-Medoids partitions [16].

MakeDensityBasedClusterer (MDBC)

It's an algorithm wrapping the SimpleKMeans and possibly others clusterers algorithms. Makes SimpleKmeans return a distribution and density. Fits normal distributions and discrete distributions within each cluster produced by the wrapped clusterer. For the SimpleKMeans supports the number of clusters requestable [15].

Cladograms

The clustering operations were made by calculating the Euclidean distance from the first alphabetically ordered species, defined as the pivotal species, to all the other species. Therefore, when ASAP created the clusters, it already had a symmetric distance matrix containing a data set with all the species. All we needed to do was to create a phylogenetic tree expressed as a Newick phylogenetic tree. We developed an unrooted tree created by the software NEIGHBOR from the PHYLIP package. We drew the unrooted tree in Figure 7, representing the eight clusters of the 60 species from dataset2. All default parameters were used.

Additional material

Additional file 1: Qualitative cluster measures. In this document, we elaborate on aspects of the qualitative cluster measures that are not discussed in this paper, such as the demand for specific metrics for clusters based on Linnaean taxonomic classification, how sequences size influence kdcSearch, a proof that amino acid trigams do not occur by chance, how to make a graphic cluster approximation by cladograms, how the evaluated algorithms were executed and the kdcSearch algorithm pseudo-code.

Additional file 2: Scilab algorithms and raw data. In this file, we elaborate on aspects of the algorithms and data used in this research. Algorithms were written in Scilab version “5.2.0.1266391513”, scilab-5.2.1.

Acknowledgements

Funding: FAPEMIG - Fundação de Amparo à Pesquisa de Minas Gerais, Brazil and CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico - Brazil.
This article has been published as part of *BMC Genomics* Volume 12 Supplement 4, 2011: Proceedings of the 6th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-meeting 2010). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2164/12?issue=S4

Author details

[1]Department of General Biology, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Av. Antônio Carlos, 6627, MG, 31.270-901, Brazil. [2]Computer Science Departament, Instituto de Ciências Exatas, Universidade Federal de Minas Gerais, Belo Horizonte, Av. Antonio Carlos, 6627, 31.270-901, MG, Brazil. [3]Max Planck Institute for Informatics, Campus E2 1, Saarbrücken, Germany. [4]CEBio and Laboratory of Cellular and Molecular Parasitology, Instituto René Rachou, Oswaldo Cruz Foundation, Belo Horizonte, Av. Augusto de Lima 1715, 30190-002, MG, Brazil. [5]Genome and Proteome Network of the State of Pará, Universidade Federal do Pará, Belém, R. Augusto Corrêa, 66.075-110, PA, Brazil.

Authors’ contributions

MAS encouraged the research and writing, BMC application, provided references and applied mathematical knowledge and gave final approval of the version to be published. ARS downloaded all the data and conducted all the tests, decided to use Linnaean taxonomy as a measure of cluster quality, developed the algorithm and wrote the paper.
JB made substantial contributions to conception and design, analysis and interpretation of data. VAA encouraged submission to BMC and gave final approval of the version to be published. JAM, GCO, AM and AS have given final approval of the version to be published.

Competing interests

The authors declare that they have no competing interests.

Published: 22 December 2011

References

1. Golub G, Kahan W: **Calculating the Singular Values and Pseudo-Inverse of a Matrix.** *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis* 1965, **2**:205-224.
2. Berry MW, Dumais ST, OBrien GW: **Using Linear Algebra for Intelligent Information Retrieval.** *SIAM Review* 1995, **37**:573-595.
3. Élden L: **Numerical linear algebra in data mining.** *Acta Numerica* 2006, **15**:327-384.
4. Élden L: **Matrix Methods in Data Mining and Pattern Recognition.** Society for Industrial and Applied Mathematics; 2007.
5. Fogolari F, Tessari S, Molinari H: **Singular value decomposition analysis of protein sequence alignment score data.** *Proteins* 2002, **46**:161-170.
6. Del-Castillo-Negrete D, Hirshman SP, Spong DA, DAzevedo EF: **Compression of magnetohydrodynamic simulation data using singular value decomposition.** *Journal of Computational Physics* 2007, **222**:265-286.
7. Deerwester SC, Dumais ST, Furnas GW, Harshman RA, Landauer TK, Lochbaum KE, Streeter LA: **Computer information retrieval using latent semantic structure.** *U. S. Patent: 4839853* 1989.
8. Stuart GW, Moffett K, Leader JJ: **A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes.** *Mol Biol Evol* 2002, **19**:554-562.
9. Vries JK, Liu X: **Subfamily specific conservation profiles for proteins based on n-gram patterns.** *BMC Bioinformatics* 2008, **9**:72.
10. Ider YZ, Onart S: **Algebraic reconstruction for 3D magnetic resonance-electrical impedance tomography (MREIT) using one component of magnetic flux density.** *Physiol Meas* 2004, **25**:281-294.
11. Hartigan JA, W MA: **Algorithm AS 136: A K-Means Clustering Algorithm.** *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 1979, **28**:100-108.
12. Dempster AP, Laird NM, Rubin DB: **Maximum Likelihood from Incomplete Data via the EM Algorithm.** *Journal of the Royal Statistical Society* 1977, **39**:1-38.
13. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y: **Adaptive quality-based clustering of gene expression profiles.** *Bioinformatics* 2002, **18**:735-746.
14. Kaufman L, Rousseeuw P: **Finding Groups in Data An Introduction to Cluster Analysis.** Wiley Interscience; 1990.
15. Witten IH, Frank E, Hall MA: **Data Mining: Practical Machine Learning Tools and Techniques.** Morgan Kaufmann; 2011.
16. Team RDC: **R: A Language and Environment for Statistical Computing.** 2006.
17. Abeel T, de Peer YV, Saeys Y: **Java-ML: A Machine Learning Library.** *Journal of Machine Learning Research* 2009, **10**:931-934.
18. Liu Q, Zhang Y, Xu Y, Ye X: **Fuzzy kernel clustering of RNA secondary structure ensemble using a novel similarity metric.** *J Biomol Struct Dyn* 2008, **25**:685-696.
19. Vries JK, Munshi R, Tobi D, Klein-Seetharaman J, Benos PV, Bahar I: **A sequence alignment-independent method for protein classification.** *Appl Bioinformatics* 2004, **3**:137-148.
20. Couto BRGM, Ladeira AP, Santos MA: **Application of latent semantic indexing to evaluate the similarity of sets of sequences without multiple alignments character-by-character.** *Genet Mol Res* 2007, **6**:983-999.
21. Gordon AD, Vichi M: **Partitions of Partitions.** *Journal of Classification* 1998, **15**:265-285.

doi:10.1186/1471-2164-12-S4-S11
Cite this article as: Santos *et al.*: A singular value decomposition approach for improved taxonomic classification of biological sequences. *BMC Genomics* 2011 **12**(Suppl 4):S11.

4. Conclusões, contribuições e perspectivas

4.1 Montagem de diversos genomas de microrganismos

Este trabalho de tese está inserido dentro de uma associação entre a Rede de Proteômica e Genômica do Pará (UFPA) e o Laboratório de Genética Celular e Molecular (UFMG), que tem como foco sequenciar genomas de microrganismos. Durante este trabalho de tese participamos dos projetos genoma de 15 linhagens *C. pseudotuberculosis*, um genoma de *Campylobacter fetus*, um *Methylobacter*, um genoma de Archaea, dois de *Streptococcus* e um de *Leptospira*.

O grupo de pesquisa conseguiu depositar 15 genomas de *C. pseudotuberculosis*, abrindo caminho para que uma nova metodologia de análise de dados possa ser utilizada no combate a esse patógeno, a Patogenômica. Partiu-se das tecnologias de sequenciamento Sanger e 454, no primeiro genoma (*C. pseudotuberculosis*, linhagem 1002), passando por duas versões da tecnologia de sequenciamento SOLiD, até alcançar os sequenciadores *Ion Torrent* e *Solid 5500*. O grupo tornou-se pioneiro ao acompanhar o lançamento e desenvolvimento dessas tecnologias de sequenciamento de nova geração, empregando-as em suas pesquisas. Porém, o pioneirismo veio acompanhado de desafios tanto de montagem quanto de anotação e análise de dados. À medida que as tecnologias de sequenciamento melhoravam, ficava mais fácil gerar dados com maior qualidade. Nesse contexto, a análise conjunta de milhares de genes também passou a ser um desafio. Dessa forma, análises *in silico* de genes promissores para vacinas, diagnóstico e drogas foram alguns dos primeiros passos na pesquisa patogenômica.

4.2 Um banco de dados e ferramentas para gerenciar genomas

Um das contribuições dessa tese foi a criação e utilização de um SGBD e um *parser* para a compilação de dados do pangenoma de *C. pseudotuberculosis* e de outros organismos bacterianos. O esquema relacional criado permite realizar análises com rapidez por meio da geração de relatórios em linguagem SQL. De outra forma, essas análises demandariam a escrita de vários programas de computador específicos para junção de dados e avaliação de resultados. Esse banco de dados permitiu agilizar o processo de anotação de um genoma recém montado, bem como garantir uma anotação uniforme entre todas as linhagens.

4.3 Predição do exoproteoma *in silico* da *C. pseudotuberculosis*

Nesse trabalho de tese, a análise *in silico* de proteínas exportadas foi aplicada em cinco linhagens da *C. pseudotuberculosis*. Essa predição permitiu a filtragem de mais de 17 mil proteínas da bactéria *C. pseudotuberculosis*, linhagens 1002, C231, I19, PAT10 e FRC41, culminando numa lista de 122 proteínas preditas *in silico* como secretadas para o meio extracelular. Dessa lista *in silico,* 21 proteínas foram comprovadas experimentalmente como secretadas para o meio extracelular nas linhagens 1002 e C231, linhagens adotadas como modelo. A comprovação de exportação dessas proteínas ocorreu no trabalho publicado em 2011 (Seção 3.2.2), com 44 proteínas comuns às duas linhagens modelo, contendo proteínas secretadas e potencialmente expostas na superfície. A comparação do proteoma predito das cinco linhagens (Seção 3.2.1), mostrou evidências, *in silico,* sobre a possibilidade dessa lista ser maior, com pelo menos 27 proteínas exportadas. A constatação dessa hipótese foi feita parcialmente no trabalho publicado em 2012 (Seção 3.2.3), no qual três proteínas encontradas no exoproteoma variante da linhagem 1002, no ano de 2011, foram encontradas no exoproteoma central. Esses experimentos mostraram que análises *in silico,* em várias linhagens de um organismo, podem servir como estimativa inicial sobre resultados esperados nos experimentos *in vitro*.

4.4 Predição do imunoproteoma *in silico* da *C. pseudotuberculosis*

Ainda como foco em definir candidatos, a imunoinformática foi utilizada para predizer o potencial imunogênico de proteínas exportadas do pangenoma por meio da ligação de peptídeos ao MHC de classe I. Essa abordagem deu origem a uma contribuição original desse trabalho: um novo método de predição de proteínas possíveis indutoras de respostas imunes por meio de uma nova estatística. Essa estatística considera a concentração de epitopos preditos por extensão da porção proteica exposta ao meio extracelular. Esse método foi utilizado sob alta sensibilidade ao predizer proteínas do genoma de *M. tuberculosis* relacionadas com antigenicidade e patogenicidade. Essa medida foi denominada densidade de epitopos maduros, da sigla em inglês *Mature Epitope Density (MED),* permitindo selecionar quatro alvos vacinais de *C. pseudotuberculosis* preditos *in silico* como expostos ao meio extracelular e comprovados experimentalmente como exportados para o meio extracelular nas linhagens modelo. A análise do potencial imunogênico por meio da estatística *MED* também identificou outras 32 proteínas de *C. pseudotuberculosis,* preditas de estarem expostas no meio extracelular, que apresentam um valor de MED elevado. Essa lista tem 27 proteínas ancoradas na membrana celular e cinco outras proteínas preditas como expostas ao meio extracelular.

Quatro candidatos a alvos vacinais preditos como secretados pela *C. pseudotuberculosis,* detentores de algumas das maiores pontuações da estatística MED, estão sendo avaliados no tocante à capacidade de gerar respostas imunes. Esses alvos foram clonados, expressos e purificados, seguindo agora para a etapa de testes de imunoproteômica. Um alvo é foco da dissertação de mestrado da aluna do LGCM Renata Faria Silva; os outros três alvos estão sendo testados em uma parceria com a Universidade Federal de Pelotas, por intermédio da professora Sibele Borsuk. Esses trabalhos serão convertidos em artigos científicos nos quais o estudante também será coautor.

4.5 Ferramenta web para predição de alvos com potencial imunogênico

Criou-se o sítio MEDPIPE, um *pipeline* computacional para expandir as análises aqui apresentadas para os demais genomas de *C. pseudotuberculosis* e outras bactérias. Dessa forma, foi possível aumentar as evidências a respeito do potencial vacinal de listas de proteínas e incluir novos candidatos de linhagens específicas da *C. pseudotuberculosis*.

4.6 A Álgebra Linear para tratamento de dados em Aprendizado de Máquina

Outra contribuição original dessa tese foi o artigo científico mostrando a viabilidade em utilizarmos técnicas de álgebra linear aliadas e métricas de qualidade biológicas para melhorar a qualidade de agrupamentos de sequências biológicas e cladogramas. Essa metodologia permite o agrupamento de proteínas de função desconhecida juntamente com proteínas de função conhecida e, desse modo, poderá viabilizar conclusões a respeito da função de proteínas agrupadas por meios computacionais. Essa técnica poderá ser útil para os genomas de *C. pseudotuberculosis* que possuem em média 27% de proteínas preditas com função desconhecida. Um melhor entendimento a respeito da provável função de proteínas hipotéticas poderá nos fornecer análises mais significativas do trabalho desenvolvido pelo nosso grupo de pesquisa a respeito, por exemplo, da predição de ilhas de patogenicidade, conforme descrito na sessão 6.2.4 (Plasticidade Genômica e Evolução Bacteriana).

4.7 Perspectivas

Como perspectiva desse trabalho está a automação total do tutorial de transferência de anotação funcional entre genomas. Esse trabalho está sendo realizado em uma parceria entre pesquisadores brasileiros, indianos e americanos. Está sendo construído um sítio na internet para que os usuários possam fornecer como entrada uma fita de DNA e receber como resultado um arquivo EMBL ou *GenBank* com a transferência de anotação oriunda de um ou mais genomas indicados pelo usuário.

Por fim, as análises concluídas até a presente data serão estendidas aos demais genomas de *C. pseudotuberculosis*, buscando contribuir com um menor número de candidatos a serem testados em experimentos *in vitro*, descobrir os genes exportados compartilhados entre todos os genomas e entender possivelmente a preferência por determinado hospedeiro.

5. Referências

Almeida, S.S. (2011). Identificação e caracterização de peptídios bioativos através de *phage display* usando genoma completo de *Corynebacterium pseudotuberculosis*. Tese de doutorado em Genética, Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais, Belo Horizonte-MG.

Alves, F.S.F., Pinheiro, R.R and Pires, P.C. (1997). Linfadenite caseosa: patogeniadiagnóstico-controle. *Embrapa Sobral CE*, **27** .

Arroio, A.. (2006). Louis Pasteur: um cientista humanista. *Revista Eletrônica de Ciências*, , http://www.cdcc.sc.usp.br/ciencia/artigos/art_31/EraUmaVez.html.

Arsenault, J., Girard, C, Dubreuil, P, Daignault, D, Galarneau, J.R, Boisclair, J, Simard, C and Bélanger, D. (2003). Prevalence of and carcass condemnation from maedi-visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. *Prev. Vet. Med.*, **59**, 67–81.

Bambini, S. and Rappuoli, R. (2009). The use of genomics in microbial vaccine development. *Drug Discov Today*, **14**, 252-260.

Barakat A.A., Selim S.A., Atep A., Saber M.S., Napih E.K., Elebeedy A.A. (1984) Two serotypes of *Corynebacterium pseudotuberculosis* isolated from different animal species. *Rev Sci Tech Off Int Eptz1,* 151-163.

Barinov, A., Loux, V, Hammani, A, Nicolas, P, Langella, P, Ehrlich, D, Maguin, E and van de Guchte, M. (2009). Prediction of surface exposed proteins in Streptococcus pyogenes, with a potential application to other Gram-positive bacteria. *Proteomics*, **9**, 61-73.

Barksdale, L., Linder, R, Sulea, I.T and Pollice, M. (1981). Phospholipase D activity of *Corynebacterium pseudotuberculosis* (*Corynebacterium* ovis) and *Corynebacterium* ulcerans, a distinctive marker within the genus *Corynebacterium*. *J Clin Microbiol*, **13**, 335-343.

Beláková, J., Horynová, M, Krupka, M, Weigl, E and Raska, M. (2007). DNA vaccines: are they still just a powerful tool for the future?. *Arch Immunol Ther Exp (Warsz)*, **55**, 387-398.

Bendtsen, J.D., Jensen, L.J, Blom, N, Von Heijne, G and Brunak, S. (2004). Feature-based prediction of non-classical and leaderless protein secretion. *Protein Eng Des Sel*, **17**, 349-356.

Bendtsen, J.D., Nielsen, H, von Heijne, G and Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**, 783-795.

Bendtsen, J.D., Nielsen, H, Widdick, D, Palmer, T and Brunak, S. (2005). Prediction of twin-arginine signal peptides. *BMC Bioinformatics*, **6**, 167.

Bernheimer, A.W., Campbell, B.J and Forrester, L.J. (1985). Comparative toxinology of Loxosceles reclusa and *Corynebacterium pseudotuberculosis*. *Science*, **228**, 590-591.

Bhavsar, A.P., Guttman, J.A and Finlay, B.B. (2007). Manipulation of host-cell pathways by bacterial pathogens. *Nature*, **449**, 827-834.

Billington, S.J., Esmay, P.A, Songer, J.G and Jost, B.H. (2002). Identification and role in virulence of putative iron acquisition genes from *Corynebacterium pseudotuberculosis*. *FEMS Microbiol*, , Lett. 208, 41–45.

Binnewies, T.T., Motro, Y, Hallin, P.F, Lund, O, Dunn, D, La, T, Hampson, D.J, Bellgard, M, Wassenaar, T.M and Ussery, D.W. (2006). Ten years of bacterial genome sequencing: comparative-genomics-based discoveries. *Funct Integr Genomics*, **6**, 165-185.

Blythe, M.J. and Flower, D.R. (2005). Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci*, **14**, 246-248.

Briles, D.E., Hollingshead, S.K, Paton, J.C, Ades, E.W, Novak, L, van Ginkel, F.W and Benjamin, W.H.J. (2003). Immunizations with pneumococcal surface protein A and pneumolysin are protective against pneumonia in a murine model of pulmonary infection with Streptococcus pneumoniae. *J Infect Dis*, **188**, 339-348.

Brogden, K.A., Chedid, L, Cutlip, R.C, Lehmkuhl, H.D and Sacks, J. (1990). Effect of muramyl dipeptide on immunogenicity of *Corynebacterium pseudotuberculosis* whole-cell vaccines in mice and lambs. *Am J Vet Res*, **51**, 200-202.

Brown, C.C., Olander, H.J and Alves, S.F. (1987). Synergistic hemolysis-inhibition titers associated with caseous lymphadenitis in a slaughterhouse survey of goats and sheep in Northeastern Brazil. *Can J Vet Res*, **51**, 46-49.

Brown, C.C., Olander, H.J, Biberstein, E.L and Morse, S.M. (1986). Use of a toxoid vaccine to protect goats against intradermal challenge exposure to *Corynebacterium pseudotuberculosis*. *Am J Vet Res*, **47**, 1116-1119.

Bucher, P., Kar*plus*, K, Moeri, N and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Comput Chem*, **20**, 3-23.

Buck, G.A., Cross, R.E., Wong, T.P., Loera, J., Groman, N. (1985) DNA relationships among some tox-bearing corynebacteriphages. *Infect. Immun.*, **49**, 679–684.

Carne, H.R. and Onon, E.O. (1978). Action of *Corynebacterium* ovis exotoxin on endothelial cells of blood vessels. *Nature*, **271**, 246-248.

Carne, H.R., Kater, J.C and Wickham, N. (1956). A toxic lipid from the surface of *Corynebacterium* ovis. *Nature*, **178**, 701-702.

Cerdeira, L.T., Carneiro, A.R, Ramos, R.T.J, de Almeida, S.S, D'Afonseca, V, Schneider, M.P.C, Baumbach, J, Tauch, A, McCulloch, J.A, Azevedo, V.A.C and Silva, A. (2011). Rapid hybrid de novo assembly of a microbial genome using only short reads: *Corynebacterium pseudotuberculosis* I19 as a case study. *J Microbiol Methods*, **86**, 218-223.

Cerdeño-Tárraga, A.M., Efstratiou, A, Dover, L.G, Holden, M.T.G, Pallen, M, Bentley, S.D, Besra, G.S, Churcher, C, James, K.D, De Zoysa, A, Chillingworth, T, Cronin, A, Dowd, L, Feltwell, T, Hamlin, N, Holroyd, S, Jagels, K, Moule, S, Quail, M.A, Rabbinowitsch, E, Rutherford, K.M, Thomson, N.R, Unwin, L, Whitehead, S, Barrell, B.G and Parkhill, J. (2003). The complete genome sequence and analysis of *Corynebacterium* diphtheriae NCTC13129. *Nucleic Acids Res*, **31**, 6516-6523.

Chaplin, P.J., De Rose, R, Boyle, J.S, McWaters, P, Kelly, J, Tennent, J.M, Lew, A.M and Scheerlinck, J.P. (1999). Targeting improves the efficacy of a DNA vaccine against *Corynebacterium pseudotuberculosis* in sheep. *Infect Immun*, **67**, 6434-6438.

Coelho, K.S. (2007). Isolamento, clonagem e caracterização molecular do gene hsp60 de *Corynebacterium pseudotuberculosis* e sua utilização na construção de uma vacina de dna e de subunidade protéica. Dissertação de mestrado em Genética, Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais, Belo Horizonte-MG, 183 pp.

Collett, M.G., Bath, G.F., Cameron, C.M. (1994) *Corynebacterium pseudotuberculosis* infections. in Infectious diseases of livestock with special reference to Southern Africa. eds Coetzer J., Thomson G. R., Justin R. C. (*Oxford University Press, Cape Town, South Africa*), 1387–1395.

Conesa, A., Götz, S, García-Gómez, J.M, Terol, J, Talón, M and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, **21**, 3674-3676.

Coyle, M.B. and Lipsky, B.A. (1990). Coryneform bacteria in infectious diseases: clinical and laboratory aspects. *Clin Microbiol Rev*, **3**, 227-246.

D'Afonseca, V., Prosdocimi, F, Dorella, F.A, Pacheco, L.G.C, Moraes, P.M, Pena, I, Ortega, J.M, Teixeira, S, Oliveira, S.C, Coser, E.M, Oliveira, L.M, Corrêa de Oliveira, G, Meyer, R, Miyoshi, A and Azevedo, V. (2009). Survey of genome organization and gene content of *Corynebacterium pseudotuberculosis*. *Microbiol Res*, , .

Desvaux, M., Hébraud, M, Talon, R and Henderson, I.R. (2009). Secretion and subcellular localizations of bacterial proteins: a semantic awareness issue. *Trends Microbiol*, **17**, 139-145.

Donnelly, J.. (2003). *DNA vaccines. In: New bacterial vaccines*. Kluwer Academic/Plenum Publishers.

Dorella, F.A. (2009). Análise do potencial vacinal de linhagens recombinantes e selvagens inativadas de *Corynebacterium pseudotuberculosis*. Tese de doutorado em Genética, Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais, Belo Horizonte-MG, 191 pp.

Dorella, F.A., Estevam, E.M, Pacheco, L.G.C, Guimarães, C.T, Lana, U.G.P, Gomes, E.A, Barsante, M.M, Oliveira, S.C, Meyer, R, Miyoshi, A and Azevedo, V. (2006). *in vivo* insertional mutagenesis in *Corynebacterium pseudotuberculosis*: an efficient means to identify DNA sequences encoding exported proteins. *Appl Environ Microbiol*, **72**, 7368-7372.

Dorella, F.A., Pacheco, L.G, Seyffert, N, Portela, R.W, Meyer, R, Miyoshi, A and Azevedo, V. (2009). Antigens of *Corynebacterium pseudotuberculosis* and prospects for vaccine development. *Expert Rev Vaccines*, **8**, 205-213.

Dorella, F.A., Pacheco, L.G.C, Oliveira, S.C, Miyoshi, A and Azevedo, V. (2006). *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. *Vet Res*, **37**, 201-218.

Dorrell, N., Mangan, J.A, Laing, K.G, Hinds, J, Linton, D, Al-Ghusein, H, Barrell, B.G, Parkhill, J, Stoker, N.G, Karlyshev, A.V, Butcher, P.D and Wren, B.W. (2001). Whole genome comparison of Campylobacter jejuni human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res*, **11**, 1706-1715.

Dougan, G.. (1994). 1993 Colworth Prize Lecture. The molecular basis for the virulence of bacterial pathogens: implications for oral vaccine development. *Microbiology*, **140 (Pt 2)**, 215-224.

Dunachie, S.J. and Hill, A.V.S. (2003). Prime-boost strategies for malaria vaccine development. *J Exp Biol*, **206**, 3771-3779.

Earl, A.M., Losick R. and Kolter R. (2007). Bacillus subtilis genome diversity. *J Bacteriol*, **189**, 1163-1170.

Eddy, S.R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755-763.

Egen, N.B., Cuevas, W.A, McNamara, P.J, Sammons, D.W, Humphreys, R and Songer, J.G. (1989). Purification of the phospholipase D of *Corynebacterium pseudotuberculosis* by recycling isoelectric focusing. *Am J Vet Res*, **50**, 1319-1322.

Eggleton, D.G., Doidge, C.V, Middleton, H.D and Minty, D.W. (1991). Immunisation against ovine caseous lymphadenitis: efficacy of monocomponent *Corynebacterium pseudotuberculosis* toxoid vaccine and combined clostridial-corynebacterial vaccines. *Aust Vet J*, **68**, 320-321.

Ellis, J.A., Hawk, D.A, Mills, K.W and Pratt, D.L. (1991). Antigen specificity and activity of ovine antibodies induced by immunization with *Corynebacterium pseudotuberculosis* culture filtrate. *Vet Immunol Immunopathol*, **28**, 303-316.

Foss, D.L. and Murtaugh, M.P. (2000). Mechanisms of vaccine adjuvanticity at mucosal surfaces. *Anim Health Res Rev*, **1**, 3-24.

Fukiya, S., Mizoguchi, H, Tobe, T and Mori, H. (2004). Extensive genomic diversity in pathogenic Escherichia coli and Shigella Strains revealed by comparative genomic hybridization microarray. *J Bacteriol*, **186**, 3911-3921.

Gay, C.G., Zuerner, R, Bannantine, J.P, Lillehoj, H.S, Zhu, J.J, Green, R and Pastoret, P.P. (2007). Genomics and vaccine development. *Rev Sci Tech*, **26**, 49-67.

Geromanos, S.J., Vissers, J.P.C, Silva, J.C, Dorschel, C.A, Li, G, Gorenstein, M.V, Bateman, R.H and Langridge, J.I. (2009). The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS. *Proteomics*, **9**, 1683-1695.

Guimaraes, A., Carmo, F, Heinemann, M, Portela, R, Meyer, R, Lage, A, Seyffert, N, Miyoshi, A, Azevedo, V and Gouveia, A. (2011). High sero-prevalence of caseous lymphadenitis identified in slaughterhouse samples as a consequence of deficiencies in sheep farm management in the state of Minas Gerais, Brazil. *BMC Veterinary Research*, **7**, 68.

Hard, G.C.. (1972). Examination by electron microscopy of the interaction between peritoneal phagocytes and *Corynebacterium* ovis. *J Med Microbiol*, **5**, 483-491.

Hard, G.C.. (1975). Comparative toxic effect of the surface lipid of *Corynebacterium* ovis on peritoneal macrophages. *Infect Immun*, **12**, 1439-1449.

Hodgson, A.L., Krywult, J, Corner, L.A, Rothel, J.S and Radford, A.J. (1992). Rational attenuation of *Corynebacterium pseudotuberculosis*: potential cheesy gland vaccine and live delivery vehicle. *Infect Immun*, **60**, 2900-2905.

Hodgson, A.L., Tachedjian, M, Corner, L.A and Radford, A.J. (1994). Protection of sheep against caseous lymphadenitis by use of a single oral dose of live recombinant *Corynebacterium pseudotuberculosis*. *Infect Immun*, **62**, 5275-5280.

Holstad, G.. (1989). *Corynebacterium pseudotuberculosis* infection in goats. IX. The effect of vaccination against natural infection. *Acta Vet Scand*, **30**, 285-293.

Holt, J.G., Krieg, N.R, Sneath P.H.A., Staley J.T., Williiams S.T. (1994) Bergey´s manual of determinative bacteriology. 9th.ed. Baltimore: Willians and Wilkins, p. 189.

Hoof, I., Peters, B, Sidney, J, Pedersen, L.E, Sette, A, Lund, O, Buus, S and Nielsen, M. (2009). NetMHCpan, a method for MHC class I binding prediction beyond humans. *Immunogenetics*, **61**, 1-13.

Join-Lambert, O.F., Ouache, M, Canioni, D, Beretti, J, Blanche, S, Berche, P and Kayal, S. (2006). *Corynebacterium pseudotuberculosis* necrotizing lymphadenitis in a twelve-year-old patient. *Pediatr Infect Dis J*, **25**, 848-851.

Juncker, A.S., Willenbrock, H, Von Heijne, G, Brunak, S, Nielsen, H and Krogh, A. (2003). Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*, **12**, 1652-1662.

Khamis, A., Raoult, D and La Scola, B. (2004). rpoB gene sequencing for identification of *Corynebacterium* species. *J Clin Microbiol*, **42**, 3925-3931.

Kleinstein, S.H.. (2008). Getting Started in Computational Immunology. *PLoS Comput Biol*, **4**, e1000128.

Krogh, A., Larsson, B, von Heijne, G and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567-580.

Larsen, M.V., Lundegaard, C., Lamberth, K., Buus, S., Lund, O. & Nielsen, M. (2007). *BMC Bioinformatics* **8**, 424.

LeaMaster, B.R., Shen, D.T, Gorham, J.R, Leathers, C.W and Wells, H.D. (1987). Efficacy of *Corynebacterium pseudotuberculosis* bacterin for the immunologic protection of sheep against development of caseous lymphadenitis. *Am J Vet Res*, **48**, 869-872.

Lin, H.H., Ray, S, Tongchusak, S, Reinherz, E.L and Brusic, V. (2008). Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research. *BMC Immunol*, **9**, 8.

Linder, R. (1997). Rhodococcus equi and Arcanobacterium haemolyticum: two "coryneform" bacteria increasingly recognized as agents of human infection. *Emerg Infect Dis*, **3**, 145-153.

Lipsky, B.A., Goldberger, A.C, Tompkins, L.S and Plorde, J.J. (1982). Infections caused by nondiphtheria corynebacteria. *Rev Infect Dis*, **4**, 1220-1235.

Liu, D.T.L., Chan, W, Fan, D.S.P and Lam, D.S.C. (2005). An infected hydrogel buckle with *Corynebacterium pseudotuberculosis*. *Br J Ophthalmol*, **89**, 245-246.

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O. & Nielsen, M. (2008). *Nucleic Acids Res* **36**, W509-12.

Mahmouda, A. and Levin, M. (2007). Vaccines at the turn of the 21st century: a new era for immunization in public health. *International Journal of Infectious Diseases*, **11 (Supplement 2)**, S1- S2.

Maione, D., Margarit, I, Rinaudo, C.D, Masignani, V, Mora, M, Scarselli, M, Tettelin, H, Brettoni, C, Iacobini, E.T, Rosini, R, D'Agostino, N, Miorin, L, Buccato, S, Mariani, M, Galli, G, Nogarotto, R, Nardi Dei, V, Vegni, F, Fraser, C, Mancuso, G, Teti, G, Madoff, L.C, Paoletti, L.C, Rappuoli, R, Kasper, D.L, Telford, J.L and Grandi, G. (2005). Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science*, **309**, 148-150.

Manickan, E., Karem, K.L and Rouse, B.T. (1997). DNA vaccines -- a modern gimmick or a boon to vaccinology?. *Crit Rev Immunol*, **17**, 139-154.

McNamara, P.J., Cuevas, W.A and Songer, J.G. (1995). Toxic phospholipases D of *Corynebacterium pseudotuberculosis*, C. ulcerans and Arcanobacterium haemolyticum: cloning and sequence homology. *Gene*, **156**, 113-118.

Mills, A.E., Mitchell, R.D and Lim, E.K. (1997). *Corynebacterium pseudotuberculosis* is a cause of human necrotising granulomatous lymphadenitis. *Pathology*, **29**, 231-233.

Morgan, A.J. and Parker, S. (2007). Translational mini-review series on vaccines: The Edward Jenner Museum and the history of vaccination. *Clin Exp Immunol*, **147**, 389-394.

Movahedi, A.R. and Hampson, D.J. (2008). New ways to identify novel bacterial antigens for vaccine development. *Vet Microbiol*, **131**, 1-13.

Muckle, C.A. and Gyles, C.L. (1983). Relation of lipid content and exotoxin production to virulence of *Corynebacterium pseudotuberculosis* in mice. *Am J Vet Res*, **44**, 1149-1153.

Mulder, N.J., Apweiler, R, Attwood, T.K, Bairoch, A, Barrell, D, Bateman, A, Binns, D, Biswas, M, Bradley, P, Bork, P, Bucher, P, Copley, R.R, Courcelle, E, Das, U, Durbin, R, Falquet, L, Fleischmann, W, Griffiths-Jones, S, Haft, D, Harte, N, Hulo, N, Kahn, D, Kanapin, A, Krestyaninova, M, Lopez, R, Letunic, I, Lonsdale, D, Silventoinen, V, Orchard, S.E, Pagni, M, Peyruc, D, Ponting, C.P, Selengut, J.D, Servant, F, Sigrist, C.J.A, Vaughan, R and Zdobnov, E.M. (2003). The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, **31**, 315-318.

Oliveira, S.C.. (2004). *Vacinas de DNA. In: Livro Genômica*. Editora Atheneu, São Paulo.

Pacheco, L.G.C., Pena, R.R, Castro, T.L.P, Dorella, F.A, Bahia, R.C, Carminati, R, Frota, M.N.L, Oliveira, S.C, Meyer, R, Alves, F.S.F, Miyoshi, A and Azevedo, V. (2007). Multiplex PCR assay for identification of *Corynebacterium pseudotuberculosis* from pure cultures and for rapid detection of this pathogen in clinical samples. *J Med Microbiol*, **56**, 480-486.

Pacheco, L.G.C., Slade, S.E, Seyffert, N, Santos, A.R, Castro, T.L.P, Valadares, A, Santos, S.G, Farias, L.M, Pimenta, A, Meyer, R, Oliveira, S.C, Miyosh, I.A, Dowson, C.G and Azevedo, V. (2010). Three-Phase Partitioning Combined with LC-MSE for Comparative Analysis of the *Corynebacterium pseudotuberculosis* Exoproteome. *Unpublished*.

Paton, M.W., Rose, I.R, Hart, R.A, Sutherland, S.S, Mercy, A.R, Ellis, T.M and Dhaliwal, J.A. (1994). New infection with *Corynebacterium pseudotuberculosis* reduces wool production. *Aust Vet J*, **71**, 47-49.

Paule, B.J.A., Azevedo, V, Regis, L.F, Carminati, R, Bahia, C.R, Vale, V.L.C, Moura-Costa, L.F, Freire, S.M, Nascimento, I, Schaer, R, Goes, A.M and Meyer, R. (2003). Experimental *Corynebacterium* pseudotuberculosis primary infection in goats: kinetics of IgG and interferon-gamma production, IgG avidity and antigen recognition by Western blotting. *Vet Immunol Immunopathol*, **96**, 129-139.

Peel, M.M., Palmer, G.G, Stacpoole, A.M and Kerr, T.G. (1997). Human lymphadenitis due to *Corynebacterium pseudotuberculosis*: report of ten cases from Australia and review. *Clin Infect Dis*, **24**, 185-191.

Pépin, M., Pardon, P, Marly, J, Lantier, F and Arrigo, J.L. (1993). Acquired immunity after primary caseous lymphadenitis in sheep. *Am J Vet Res*, **54**, 873-877.

Piontkowski, M.D. and Shivvers, D.W. (1998). Evaluation of a commercially available vaccine against *Corynebacterium pseudotuberculosis* for use in sheep. *J Am Vet Med Assoc*, **212**, 1765-1768.

Pogson, C.A., Simmons, C.P, Strugnell, R.A and Hodgson, A.L. (1996). Cloning and manipulation of the *Corynebacterium pseudotuberculosis* recA gene for live vaccine vector development. *FEMS Microbiol Lett*, **142**, 139-145.

POSTGRES (1993). **The POSTGRES Group: The POSTGRES Reference Manual, Berkeley**. 1993.

Ramachandran, G.N., Ramakrishnan, C and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol*, **7**, 95-99.

Rappuoli, R.. (2000). Reverse vaccinology. *Curr Opin Microbiol*, **3**, 445-450.

Rappuoli, R.. (2001). Reverse vaccinology, a genome-based approach to vaccine development. *Vaccine*, **19**, 2688-2691.

Restrepo-Montoya, D., Pino, C, Nino, L.F, Patarroyo, M.E and Patarroyo, M.A. (2011). NClassG+: A classifier for non-classically secreted Gram-positive bacterial proteins. *BMC Bioinformatics*, **12**, 21.

Ribeiro, D. (2009-). Avaliação da capacidade protetora de linhagens atenuadas atraves de transposição de *Corynebacterium pseudotuberculosis* contra linfadenite Caseosa em ovinos e caprinos. Tese de doutorado em Microbiologia, Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais, Belo Horizonte-MG.

Rinaudo, C.D., Telford, J.L, Rappuoli, R and Seib, K.L. (2009). Vaccinology in the genome era. *J Clin Invest*, **119**, 2515-2525.

Ruiz, J.C., D'Afonseca, V, Silva, A, Ali, A, Pinto, A.C, Santos, A.R, Rocha, A.A.M.C, Lopes, D.O, Dorella, F.A, Pacheco, L.G.C, Costa, M.P, Turk, M.Z, Seyffert, N, Moraes, P.M.R.O, Soares, S.C, Almeida, S.S, Castro, T.L.P, Abreu, V.A.C, Trost, E, Baumbach, J, Tauch, A, Schneider, M.P.C, McCulloch, J, Cerdeira, L.T, Ramos, R.T.J, Zerlotini, A, Dominitini, A, Resende, D.M, Coser, E.M, Oliveira, L.M, Pedrosa, A.L, Vieira, C.U, Guimarães, C.T, Bartholomeu, D.C, Oliveira, D.M, Santos, F.R, Rabelo, É.M, Lobo, F.P, Franco, G.R, Costa, A.F, Castro, I.M, Dias, S.R.C, Ferro, J.A, Ortega, J.M, Paiva, L.V, Goulart, L.R, Almeida, J.F, Ferro, M.I.T, Carneiro, N.P, Falcão, P.R.K, Grynberg, P, Teixeira, S.M.R, Brommonschenkel, S, Oliveira, S.C, Meyer, R, Moore, R.J, Miyoshi, A, Oliveira, G.C and Azevedo, V. (2011). Evidence for reductive genome evolution and lateral acquisition of virulence functions in two *Corynebacterium pseudotuberculosis* strains. *PLoS One*, **6**, e18551.

Rutherford, K., Parkhill, J, Crook, J, Horsnell, T, Rice, P, Rajandream, M.A and Barrell, B. (2000). Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944-945.

Schatzmayr, H.G.. (2003). [New perspectives in viral vaccines]. *Hist Cienc Saude Manguinhos*, **10**, 655-669.

Scordis, P., Flower, D.R and Attwood, T.K. (1999). FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics*, **15**, 799-806.

Seyffert, N., Guimarães, A.S, Pacheco, L.G.C, Portela, R.W, Bastos, B.L, Dorella, F.A, Heinemann, M.B, Lage, A.P, Gouveia, A.M.G, Meyer, R, Miyoshi, A and Azevedo, V. (2010). High seroprevalence of caseous lymphadenitis in Brazilian goat herds revealed by *Corynebacterium pseudotuberculosis* secreted proteins-based ELISA. *Res Vet Sci*, **88**, 50-55.

Shpigel, N.Y., Elad, D, Yeruham, I, Winkler, M and Saran, A. (1993). An outbreak of *Corynebacterium pseudotuberculosis* infection in an Israeli dairy herd. *Vet Rec*, **133**, 89-94.

Sibbald, M.J.J.B. and van Dij, J.M.L. (2009). Secretome Mapping in Gram-Positive Pathogens. In Karl Wooldridge (ed.), Bacterial Secreted Protein: Secretory Mechanisms and Role in Pathogenesis. *Caister Academic Press*, , 193-225.

Silva, C.L., Bonato, V.L.D, Castelo, A.A.M.C, Lima, K.M and Rodrigues Júnior, J.M. (2004). *Vacinas Gênicas in Livro Genômica*. Editora Atheneu, São Paulo.

Silva, R.F. (2011-) Clonagem e expressão da ORF Cp1002_0126 de *Corynebacterium pseudotuberculosis* para o diagnóstico subclínico da linfadenite caseosa em pequenos ruminantes. Dissertação de mestrado em Genética, Instituto de Ciências Biológicas - Universidade Federal de Minas Gerais, Belo Horizonte-MG.

Silva, W.M., Seyffert, N, Castro, T.L.P, Santos, A.V, Pacheco, L.G.C, Santos, A.R, Ciprandi, A, Zurita-Turk, M, Dorella, F.A, Andrade, H.M, Pimenta, A.M.C, Silva, A, Miyoshi, A and Azevedo, V. (2012). Identification of 11 new exoproteins of *Corynebacterium pseudotuberculosis* through comparative analysis of the secretome. *Advances in Integrative Omics and Applied Biotechnology*, **1**, 22.

Simeone, R., Bottai, D and Brosch, R. (2009). ESX/type VII secretion systems and their role in host-pathogen interaction. *Curr Opin Microbiol*, **12**, 4-10.

Simmons, C.P., Dunstan, S.J, Tachedjian, M, Krywult, J, Hodgson, A.L and Strugnell, R.A. (1998). Vaccine potential of attenuated mutants of *Corynebacterium pseudotuberculosis* in sheep. *Infect Immun*, **66**, 474-479.

Simmons, C.P., Hodgson, A.L and Strugnell, R.A. (1997). Attenuation and vaccine potential of aroQ mutants of *Corynebacterium pseudotuberculosis*. *Infect Immun*, **65**, 3048-3056.

Songer, J.G.. (1997). Bacterial phospholipases and their role in virulence. *Trends Microbiol*, **5**, 156-161.

Stanford, K., Brogden, K.A, McClelland, L.A, Kozub, G.C and Audibert, F. (1998). The incidence of caseous lymphadenitis in Alberta sheep and assessment of impact by vaccination with commercial and experimental vaccines. *Can J Vet Res*, **62**, 38-43.

Stavrinides, J., McCann, H.C and Guttman, D.S. (2008). Host-pathogen interplay and the evolution of bacterial effectors. *Cell Microbiol*, **10**, 285-292.

Stranzl, T., Larsen, M.V, Lundegaard, C and Nielsen, M. (2010). NetCTLpan: pan-specific MHC class I pathway epitope predictions. *Immunogenetics*, **62**, 357-368.

Stuart, G.W., Moffett, K and Leader, J.J. (2002). A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol*, **19**, 554-562.

Sun, J., Xu, T, Wang, S, Li, G, Wu, D and Cao, Z. (2011). Does difference exist between epitope and non-epitope residues? Analysis of the physicochemical and structural properties on conformational epitopes from B-cell protein antigens. *Immunome Res*, **7**, 1-11.

Tambourgi, D.V., De Sousa Da Silva, M, Billington, S.J, Gonçalves De Andrade, R.M, Magnoli, F.C, Songer, J.G and Van Den Berg, C.W. (2002). Mechanism of induction of complement susceptibility of erythrocytes by spider and bacterial sphingomyelinases. *Immunology*, **107**, 93-101.

Tashjian, J.J. and Campbell, S.G. (1983). Interaction between caprine macrophages and *Corynebacterium pseudotuberculosis*: an electron microscopic study. *Am J Vet Res*, **44**, 690-693.

Tettelin, H., Masignani, V, Cieslewicz, M.J, Donati, C, Medini, D, Ward, N.L, Angiuoli, S.V, Crabtree, J, Jones, A.L, Durkin, A.S, Deboy, R.T, Davidsen, T.M, Mora, M, Scarselli, M, Margarit y Ros, I, Peterson, J.D, Hauser, C.R, Sundaram, J.P, Nelson, W.C, Madupu, R, Brinkac, L.M, Dodson, R.J, Rosovitz, M.J, Sullivan, S.A, Daugherty, S.C, Haft, D.H, Selengut, J, Gwinn, M.L, Zhou, L, Zafar, N, Khouri, H, Radune, D, Dimitrov, G, Watkins, K, O'Connor, K.J.B, Smith, S, Utterback, T.R, White, O, Rubens, C.E, Grandi, G, Madoff, L.C, Kasper, D.L, Telford, J.L, Wessels, M.R, Rappuoli, R and Fraser, C.M. (2005). Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A*, **102**, 13950-13955.

Trost, E., Ott, L, Schneider, J, Schröder, J, Jaenicke, S, Goesmann, A, Husemann, P, Stoye, J, Dorella, F.A, Rocha, F.S, Soares, S.D.C, D'Afonseca, V, Miyoshi, A, Ruiz, J, Silva, A, Azevedo, V, Burkovski, A, Guiso, N, Join-Lambert, O.F, Kayal, S and Tauch, A. (2010). The complete genome sequence of *Corynebacterium pseudotuberculosis* FRC41 isolated from a 12-year-old girl with necrotizing lymphadenitis reveals insights into gene-regulatory networks contributing to virulence. *BMC Genomics*, **11**, 728.

Tsurui, H. and Takahashi, T. (2007). Prediction of T-cell epitope. *J Pharmacol Sci*, **105**, 299-316.

Uchijima, M., Yoshida, A, Nagata, T and Koide, Y. (1998). Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T cell responses against an intracellular bacterium. *J Immunol*, **161**, 5594-5599.

Vaneechoutte, M., Dijkshoorn, L, Tjernberg, I, Elaichouni, A, de Vos, P, Claeys, G and Verschraegen, G. (1995). Identification of Acinetobacter genomic species by amplified ribosomal DNA restriction analysis. *J Clin Microbiol*, **33**, 11-15.

Williamson, L.H.. (2001). Caseous lymphadenitis in small ruminants. *Vet Clin North Am Food Anim Pract*, **17**, 359-71, vii.

Yeruham, I., Elad, D, Friedman, S and Perl, S. (2003). *Corynebacterium pseudotuberculosis* infection in Israeli dairy cattle. *Epidemiol Infect*, **131**, 947-955.

Yeruham, I., Elad, D, Van-Ham, M, Shpigel, N.Y and Perl, S. (1997). *Corynebacterium pseudotuberculosis* infection in Israeli cattle: clinical and epidemiological studies. *Vet Rec*, **140**, 423-427.

6. Anexos

6.1 Programas de computador utilizados ou desenvolvidos para a tese

6.1.1 Programas utilizados

O conjunto de programas listado abaixo é de livre utilização para fins acadêmicos tendo sido obtido indiretamente pela instalação do sistema operacional Linux, por acesso a repositórios públicos ou por requisição direta aos respectivos autores.

- **Sistema Operacional Linux Ubuntu versão 10.10**
- **Sistema Gerenciador de Banco de Dados PostgreSQL versão 8.4**
- **Compilador C++ gcc (Ubuntu 4.4.1-4ubuntu9) 4.4.1:** utilizado para gerar um código executável a partir do código fonte de um programa em linguagem C++.
- **flex 2.5.35 (similar o LEX)**: programa para implementar analisadores léxicos de uma linguagem qualquer. Nesta tese, a linguagem que avaliada com base em sua correção léxica são estruturas de dados armazenadas nos formatos de arquivos EMBL e *GenBank*. Se a estrutura léxica estiver correta então é avaliada a estrutura semântica da linguagem.
- **bison (GNU bison) 2.4.1 (Similar ao YACC)**: programa para implementar analisadores semânticos de uma linguagem qualquer. Nesta tese, a linguagem avaliada com base em sua correção semântica são estruturas de dados armazenadas nos formatos de arquivos EMBL e *GenBank*. Se a estrutura semântica estiver correta então os dados contidos na estrutura do arquivo EMBL ou *GenBank* são processados e convertidos como formato de entrada para o esquema relacional CpDB.
- **SignalP 3.0**: prediz a exportação de uma proteína para a membrana celular por intermédio da predição do peptídeo sinal, motivo de exportação clássico da via de secreção Sec.
- **TMHMM 2.0**: pesquisa por motivos hidrofóbicos em sequências proteicas que possam indicar que uma proteína possui porções de sua extensão que seriam exportadas e integralizadas à membrana celular.
- **LipoP**: programa para predição de motivos característicos de lipoproteínas. Inicialmente concebido para bactérias Gram-negativas, mas com evidências de desempenho satisfatório também em bactérias monodérmicas.
- **HMMER**: Conjunto de programas para identificar motivos conservados por meio de redes neurais artificiais. Inclui programas para criação de treinamento, criação de perfis de identificação de motivos conservados e utilização de perfis para busca de motivos em outras fontes de dados. Como exemplo, o programa Surfg possui mais de uma dezena de perfis desenvolvidos pelos autores do programa

para identificar motivos de retenção conhecidos que implicam no ancoramento de uma proteína à parede celular.

- **NetMHC 3.0**: programa que analisa peptídeos buscando por epitopos ligantes ao MHC de classe I de até 55 alelos. Avalia a possibilidade de ligação forte ou fraca de epitopos à molécula de MHC.
- **SecretomeP**: prediz a exportação de proteínas por vias não clássicas.
- **NClassG+**: prediz a exportação de proteínas por vias não clássicas.
- **Scilab**: ambiente de programação para implementação de problemas matemáticos. Possui inúmeras bibliotecas de programas de computadores que facilitam a tarefa de resolver um problema. Por ser um ambiente de programação, não possui uma interface com muitas janelas para entrada de dados. Para utilização ótima é necessário dispender algum tempo para aprender a sua linguagem de programação. Um exemplo de função é a decomposição de valores singulares acionada por uma função chamada *svd.* Os ambientes de programação R e Matlab são similares ao Scilab possuindo diversas funções em comum, porém cada um utiliza uma linguagem de programação específica.
- Programa **Surfg** plus **1.0:** A predição do compartimento celular no qual estariam localizadas a maior parte das proteínas dos genomas de *C. pseudotuberculosis* foi feita utilizando-se o programa Surfg *plus* 1.0 (Barinov e cols., 2009), escrito na linguagem de desenvolvimento JAVA e configurado de acordo com instruções dos autores. Esse programa é específico para a predição do compartimento celular de proteínas oriundas de genomas de procariotos, não existindo ainda uma versão para organismos eucarióticos. O Surfg *plus* é de fato o coordenador da execução de outros programas de computadores largamente utilizados para predição de localização celular: LipoP (Juncker e cols., 2003) que faz a predição de lipoproteínas; SignalP (Bendtsen e cols., 2004a) que prediz motivos de peptídeo sinal; TMHMM (Krogh e cols., 2001) que prediz motivos transmembrânicos; HMMER (Eddy, 1998) para buscas de motivos por meio de Modelos Ocultos de Markov (*HMM*). No programa Surfg foram criados oito *profiles HMM* para a pesquisa por proteínas que pudessem ser retidas na membrana e que são executados pelo Surfg *plus* como, por exemplo, o motivo de ancoramento LPxTG. Essas ferramentas foram instaladas e testadas individualmente em um hardware convencional utilizando o sistema operacional Linux Ubuntu versão 10.10. Em seguida o Surfg *plus* 1.0 foi configurado para instanciar essas ferramentas e utilizá-las em um processo coordenado no qual o

resultado do processamento de uma ferramenta poderia ser utilizado como entrada para outra ferramenta, uma clássica aplicação do conceito de *pipelines* computacionais. Considerando que *C. pseudotuberculosis* é uma bactéria Gram positiva, essa configuração foi explicitamente feita no Surfg *plus* 1.0 para que esse pudesse orquestrar a correta execução do SignalP. O outro parâmetro do Surfg *plus* 1.0 que foi alterado diz respeito ao comprimento esperado da parede celular de cada linhagem dessa bactéria. Esse dado foi aferido por meio de microscopia eletrônica e utilizado na predição do compartimento celular de proteínas de todas as linhagens. O dado do comprimento da membrana não é utilizado pelos programas acionados pelo Surfg *plus*, mas é utilizado no processo de combinação dos resultados dos programas acionados. Os demais parâmetros do Surfg *plus* 1.0 não foram alterados. Apesar desse ser o principal preditor de local subcelular utilizado nesse trabalho, esse incorpora um único programa de predição de proteínas translocadas pela via Sec, o SignalP. Essa limitação pode ser parcialmente resolvida pela utilização de outros programas de predição de proteínas translocadas por vias não clássicas. Estes programas atuam em proteínas classificadas pelo SurfG como citoplasmáticas, a saber, o TatP 1.0 (Bendtsen e cols., 2005), SecretomeP 2.0 (Bendtsen e cols., 2005b) e NClassG+ (Restrepo-Montoya e cols., 2011).

6.1.2 Programas desenvolvidos

O conjunto de programas listado a seguir foi criado pelo nosso grupo especificamente para viabilizar o armazenamento; tratamento e análise de dados do genoma, predição de proteínas exportadas e imunoinformática. Este conjunto de programas está disponível na *internet*, por meio de uma licença *GNU General Public License* (Licença Pública Geral) através do sítio http://sourceforge.net/projects/cpdb/. As versões disponíveis para download imediato são as versões estáveis, no entanto também existem as versões de desenvolvimento disponíveis por meio do programa de controle de versões *subversion*. As versões de desenvolvimento podem conter melhoria de funcionalidades, contudo podem não estar totalmente testadas.

6.1.2.1 Esquema Relacional de banco de dados CpDB para montagem e anotação dos genomas de *C. pseudotuberculosis*

O meio escolhido para armazenar os resultados desse trabalho foi um banco de dados relacional. A preferência foi dada para o Sistema Gerenciador de Bancos de Dados (SGBD) PostgreSQL, versão 8.4, nativo de diversas distribuições do sistema operacional Linux. A escolha foi baseada no fato desse SGBD possuir uma interface similar e não menos robusta quando comparada com o melhor SGBD comercial, o ORACLE™. Dentre as vantagens em utilizar um SGBD, podem ser citadas a manutenção: da integridade dos dados; dos níveis de acesso customizáveis; da automação de relacionamentos entre dados; da unicidade de identificadores e a geração de um número infinito de relatórios em linguagem de alto nível *Structed Query Language* (SQL). Tal linguagem permite extrair relatórios para os quais seria necessário a escrita de muitas linhas de código em uma linguagem de programação de computadores convencional. Para gerar tais relatórios, ao invés da escrita de extensos códigos de programação, somente é necessário passar para o computador "quais dados" se deseja exportar, ao invés de também dizer ao computador "como" exportar esses dados.

Para armazenar dados de genomas da *C. pseudotuberculosis,* foi criado o esquema relacional (ER) para o SGBD PostgreSQL denominado por *C. pseudotuberculosis Data Base* (CpDB) (Figura 10). Fazendo-se uma analogia com um documento de texto, o CpDB pode ser entendido como um documento formatado para ser preenchido, possuindo campos de dados específicos que representam categorias de dados que se deseja representar, mas sem dados. O preenchimento de dados ainda precisa ser feito e, para esse, foi criado um programa na linguagem "C" denominado parseEMBLtoCpDB.

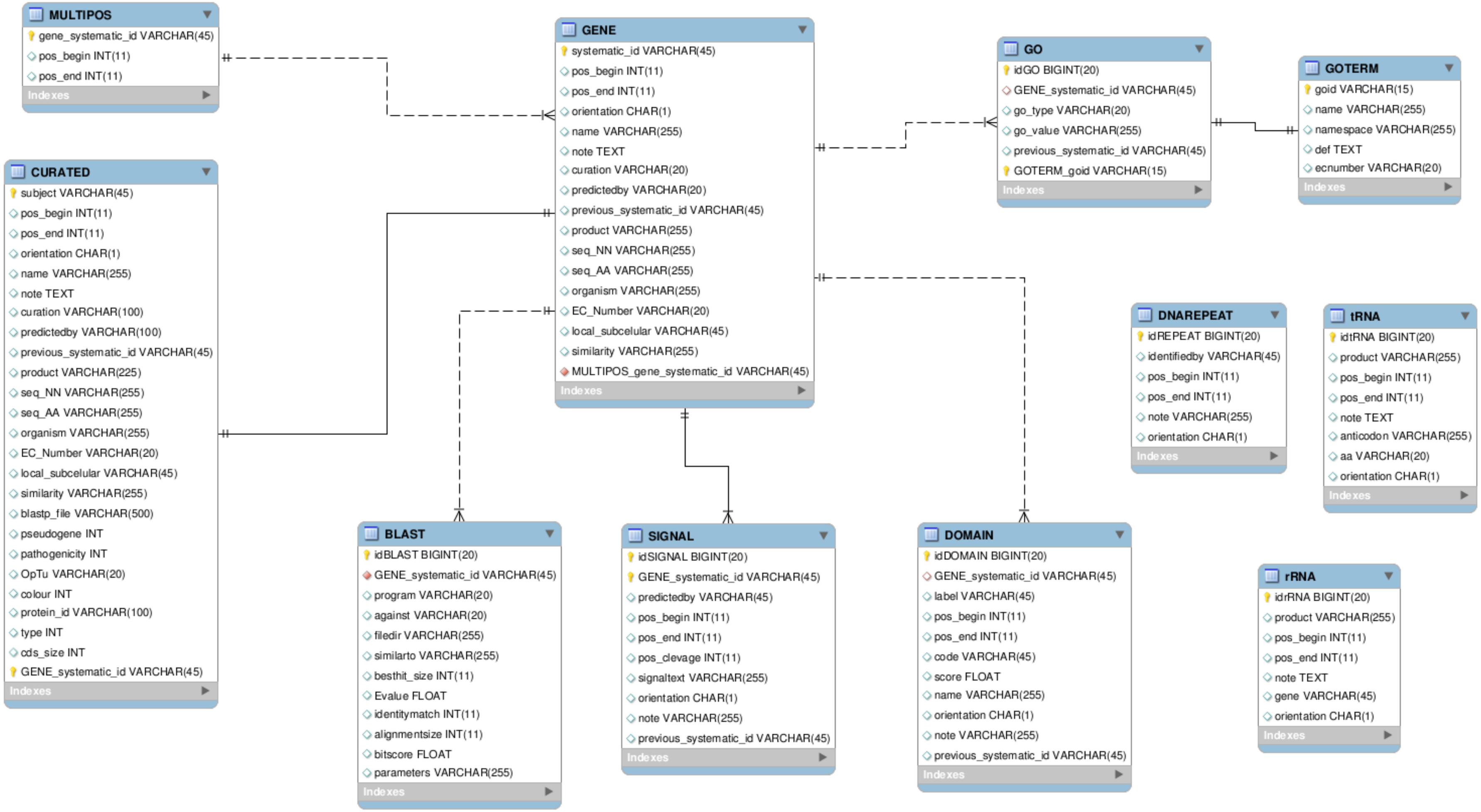


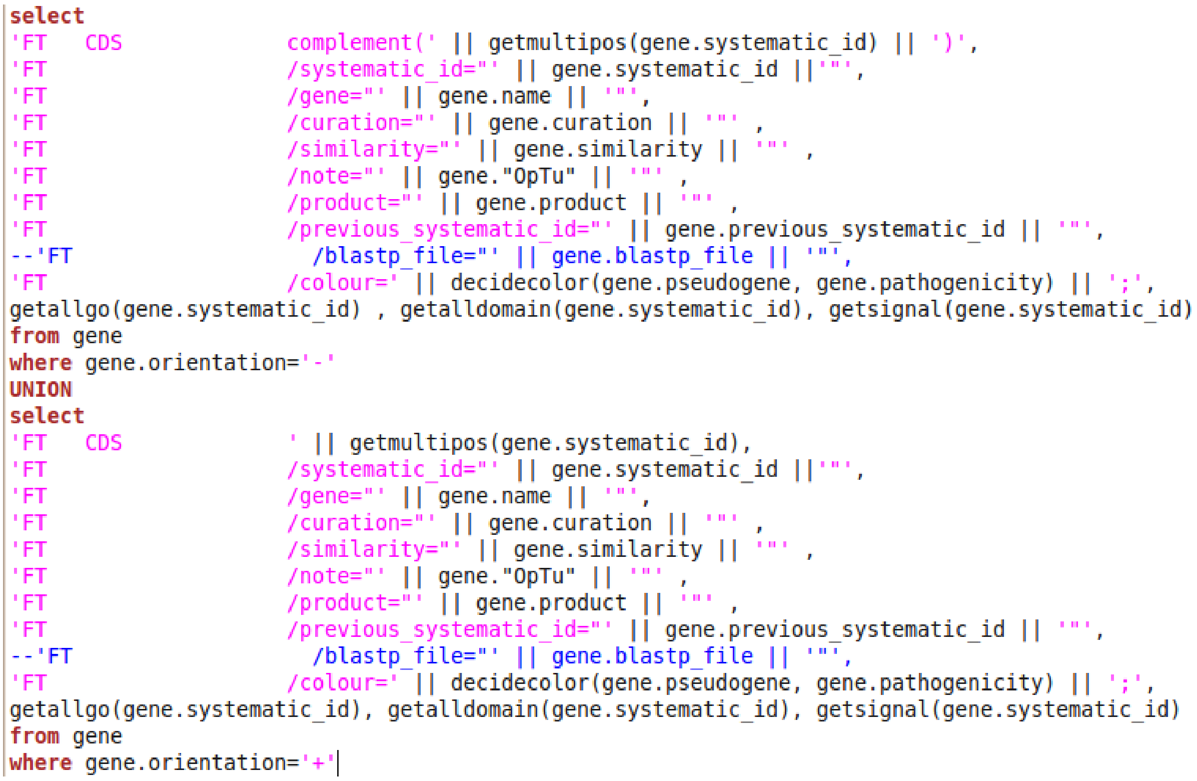
Figura 10: Esquema Relacional (ER) do sistema gerenciador de banco de dados relacionais PostgreSQL, versão 8.4.

As caixas retangulares representam entidades que são transformadas em tabelas de dados. Os losangos representam os elos entre as entidades.

Nesse banco de dados a entidade central é GENE. Todos os dados de um gene que existem em quantidades maiores que uma unidade possuem suas próprias entidades, são eles: domínios proteicos (DOMAIN), resultados de alinhamentos de sequências (BLAST), coordenadas de ORF's que constituem um possível pseudogene (MULTIPOS), resultados de pesquisas feitas ao banco de dados Gene Ontology (GO). A entidade que representa o peptídeo sinal de proteínas preditas como translocadas para a membrana citoplasmática (SIGNALP) é um caso que existe em quantidades unitárias para cada proteína. É de fato um tipo de domínio proteico, mas pensando na transformação de um modelo conceitual em um modelo prático, convertido em tabelas de bancos de dados, separação dessa entidade, da entidade DOMAIN, foi intencional. Nesse ER, não está sendo feita distinção entre GENE e CDS. Quando uma CDS possui um gene conhecido simplesmente é inserido o nome do gene no campo de dados criado para esse propósito denominado "*name*" e o campo de dados "*product*" sempre recebe o resultado da anotação funcional de uma proteína.

As entidades DNAREPEAT, tRNA e rRNA não estão relacionadas com a entidade GENE visto que essas três entidades representam respectivamente, sequências de repetição encontradas numa fita de DNA, ácidos ribonucleicos transportadores e ácidos ribonucleicos ribossomais, entidades que estão relacionadas apenas com uma fita de DNA. Por último há a entidade GOTERM, uma cópia de termos anotadores do *Gene Ontology, c*olocada no banco de dados para enriquecer relatórios com dados sobre local subcelular, processos bioquímicos e funções associadas às proteínas anotadas. Um exemplo de dado oriundo dessa tabela é o dado que identifica uma ação enzimática de uma proteína, o EC_NUMBER, utilizado como pré-requisito para análises da participação de uma proteína em vias metabólicas. As primeiras análises de vias metabólicas realizadas sobre os genomas de *C. pseudotuberculosis* foram feitas com base na exportação de relatórios do banco com esse dado incorporado e oriundo da tabela do banco correspondente a essa entidade.

Uma das consultas em linguagem SQL mais complexas produzidas nesse banco de dados é a exportação da anotação para o formato EMBL utilizado pelo programa ARTEMIS. Essa consulta está exibida na Figura 11, e faz o cruzamento de dados de grande parte das tabelas armazenadas no CpDB. O resultado é um arquivo em formato texto composto de toda a anotação manual e automática com domínios proteicos e termos GO de todo o genoma conforme exemplificado na Figura 12. Esse arquivo texto, quando concatenado com um arquivo EMBL, que possui apenas a fita de DNA do genoma, permite ao programa ARTEMIS exibir esses dados em formato gráfico usando como referencia de posicionamento a fita de DNA, conforme exibido na Figura 13.



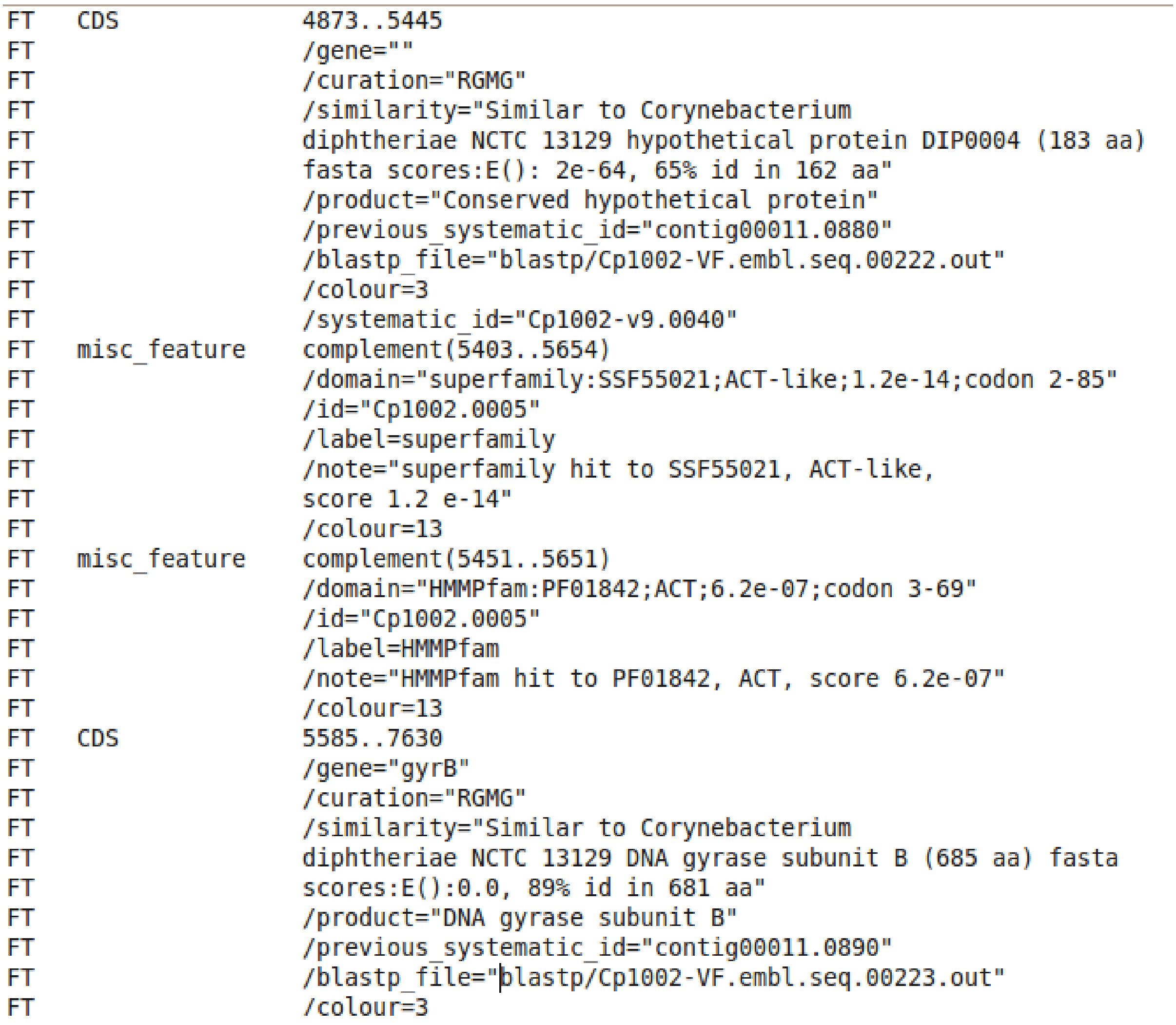
```
select
'FT   CDS               complement(' || getmultipos(gene.systematic_id) || ')',
'FT                     /systematic_id="' || gene.systematic_id ||'"',
'FT                     /gene="' || gene.name || '"',
'FT                     /curation="' || gene.curation || '"' ,
'FT                     /similarity="' || gene.similarity || '"' ,
'FT                     /note="' || gene."OpTu" || '"' ,
'FT                     /product="' || gene.product || '"' ,
'FT                     /previous_systematic_id="' || gene.previous_systematic_id || '"',
--'FT                     /blastp_file="' || gene.blastp_file || '"',
'FT                     /colour=' || decidecolor(gene.pseudogene, gene.pathogenicity) || ';',
getallgo(gene.systematic_id) , getalldomain(gene.systematic_id), getsignal(gene.systematic_id)
from gene
where gene.orientation='-'
UNION
select
'FT   CDS               ' || getmultipos(gene.systematic_id),
'FT                     /systematic_id="' || gene.systematic_id ||'"',
'FT                     /gene="' || gene.name || '"',
'FT                     /curation="' || gene.curation || '"' ,
'FT                     /similarity="' || gene.similarity || '"' ,
'FT                     /note="' || gene."OpTu" || '"' ,
'FT                     /product="' || gene.product || '"' ,
'FT                     /previous_systematic_id="' || gene.previous_systematic_id || '"',
--'FT                     /blastp_file="' || gene.blastp_file || '"',
'FT                     /colour=' || decidecolor(gene.pseudogene, gene.pathogenicity) || ';',
getallgo(gene.systematic_id), getalldomain(gene.systematic_id), getsignal(gene.systematic_id)
from gene
where gene.orientation='+'
```

Figura 11: Consulta em linguagem SQL para exportação de dados do CpDB para o formato EMBL.

Percebe-se que na Figura 11 os textos em cor rosa são constantes e serão exibidos sempre dessa forma para todas as linhas de resultados retornados por essa consulta. Os textos rodeados pelo sinal de pipe duplo (||) são termos variáveis que assumem o valor do resultado retornado do banco. Nessa consulta existe referência visível apenas à tabela "gene", porém as tabelas de múltiplas posições, termos GO, domínios proteicos e peptídeo sinal são acessados para cada gene retornado da consulta. Quem se encarrega das consultas às demais tabelas são funções escritas na linguagem PL/SQL cujos nomes começam com o prefixo "*get*". Essas funções podem retornar múltiplos resultados para cada gene, assim como exemplificado na Figura 12. O comando SQL da Figura 11 tem duas metades unidas pela cláusula UNION da linguagem SQL. Perceba que o comando acima da cláusula UNION é quase idêntico ao comando abaixo da mesma cláusula. A diferença está no formato dos dados da fita reversa que usa o termo "*complement*" precedendo as coordenadas de todas as CDS, além é claro da condição específica para retornar dados na fita direta ou reversa, explicitada pela condição imposta ao dado "*orientation*" igual a "+" ou "-".

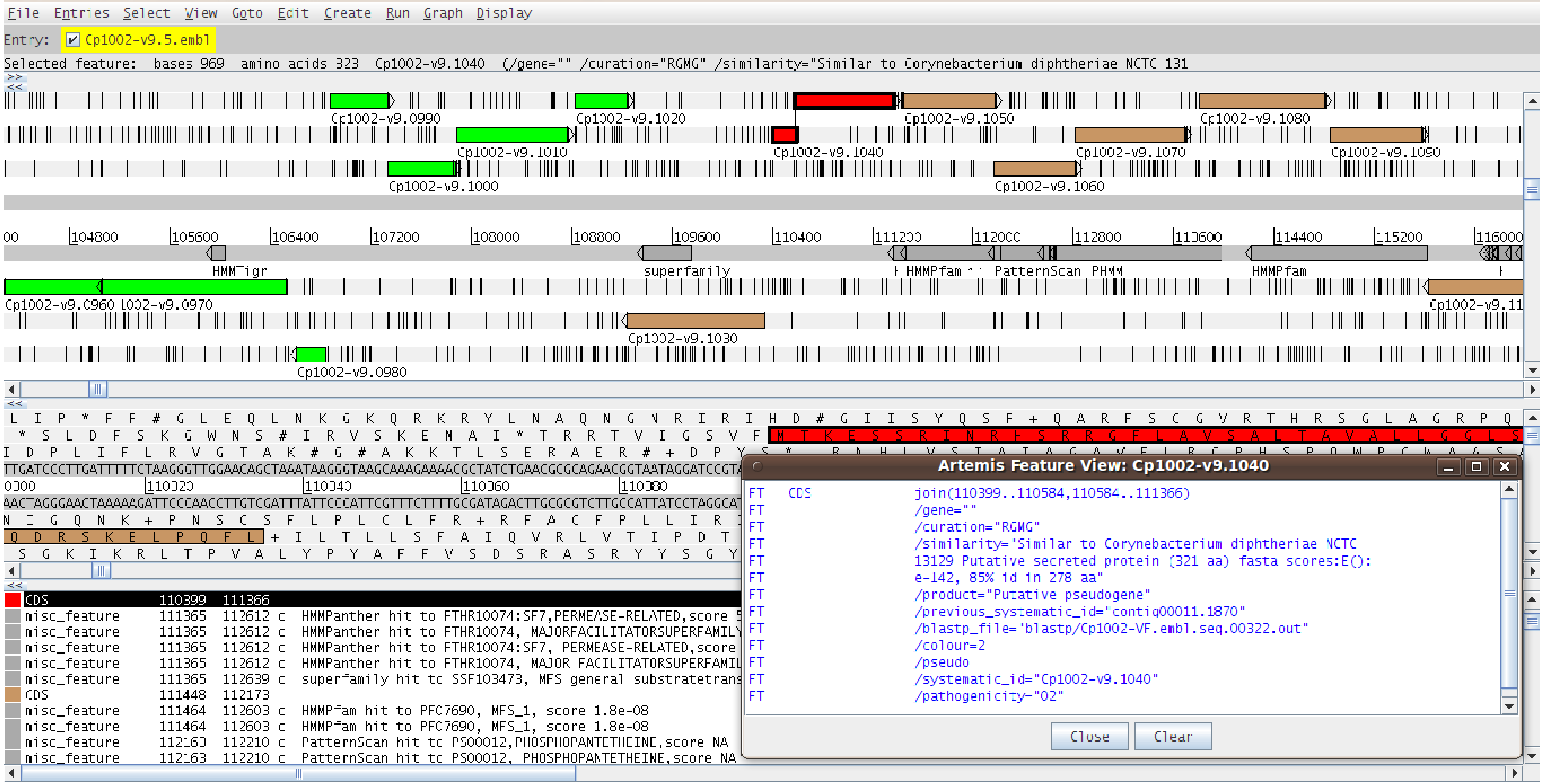
O formato de arquivo representado na Figura 12, é um texto sem formatação especial; sem caracteres especiais; também referenciado como formato de texto simples. Esse arquivo pode ser manipulado por qualquer programa editor de textos. Nesse exemplo têm-se duas CDS's (entidade GENE), sendo que a primeira possui dois domínios proteicos associados (entidade DOMAIN), referentes a porções da CDS que pertenceriam à famílias de proteínas. Esses dois domínios proteicos foram retornados pela função *getalldomain*, chamada no SQL da Figura 11.

Quando o arquivo da Figura 12 é concatenado com um EMBL possuindo apenas uma fita de DNA, gera um novo arquivo que ao ser aberto com o programa ARTEMIS propicia a visualização de dados de acordo com a Figura 13; em verde, vermelho e marrom as proteínas curadas manualmente. Proteínas em vermelho são pseudogenes e em marrom são proteínas identificadas como pertencentes a uma ilha de patogenicidade (Seção 6.2.4). Os domínios proteicos estão sobre as duas fitas centrais em cinza, também em cor cinza, porém com tonalidade mais escura. A janela no canto inferior direito exibe os dados diversos exportados do banco para o arquivo EMBL.



```
FT   CDS             4873..5445
FT                   /gene=""
FT                   /curation="RGMG"
FT                   /similarity="Similar to Corynebacterium
FT                   diphtheriae NCTC 13129 hypothetical protein DIP0004 (183 aa)
FT                   fasta scores:E(): 2e-64, 65% id in 162 aa"
FT                   /product="Conserved hypothetical protein"
FT                   /previous_systematic_id="contig00011.0880"
FT                   /blastp_file="blastp/Cp1002-VF.embl.seq.00222.out"
FT                   /colour=3
FT                   /systematic_id="Cp1002-v9.0040"
FT   misc_feature    complement(5403..5654)
FT                   /domain="superfamily:SSF55021;ACT-like;1.2e-14;codon 2-85"
FT                   /id="Cp1002.0005"
FT                   /label=superfamily
FT                   /note="superfamily hit to SSF55021, ACT-like,
FT                   score 1.2 e-14"
FT                   /colour=13
FT   misc_feature    complement(5451..5651)
FT                   /domain="HMMPfam:PF01842;ACT;6.2e-07;codon 3-69"
FT                   /id="Cp1002.0005"
FT                   /label=HMMPfam
FT                   /note="HMMPfam hit to PF01842, ACT, score 6.2e-07"
FT                   /colour=13
FT   CDS             5585..7630
FT                   /gene="gyrB"
FT                   /curation="RGMG"
FT                   /similarity="Similar to Corynebacterium
FT                   diphtheriae NCTC 13129 DNA gyrase subunit B (685 aa) fasta
FT                   scores:E():0.0, 89% id in 681 aa"
FT                   /product="DNA gyrase subunit B"
FT                   /previous_systematic_id="contig00011.0890"
FT                   /blastp_file="blastp/Cp1002-VF.embl.seq.00223.out"
FT                   /colour=3
```

Figura 12: Exemplo de execução da consulta mostrada no SQL da Figura 11.



```
Selected feature:  bases 969  amino acids 323  Cp1002-v9.1040  (/gene="" /curation="RGMG" /similarity="Similar to Corynebacterium diphtheriae NCTC 131
```

```
 L  I  P  *  F  F  #  G  L  E  Q  L  N  K  G  K  Q  R  K  R  Y  L  N  A  Q  N  G  N  R  I  R  I  H  D  #  G  I  I  S  Y  Q  S  P  +  Q  A  R  F  S  C  G  V  R  T  H  R  S  G  L  A  G  R  P  Q
  *  S  L  D  F  S  K  G  W  N  S  #  I  R  V  S  K  E  N  A  I  *  T  R  R  T  V  I  G  S  V  F  M  T  K  E  S  S  R  I  N  R  H  S  R  R  G  F  L  A  V  S  A  L  T  A  V  A  L  L  G  G  L  S
I  D  P  L  I  F  L  R  V  G  T  A  K  #  G  #  A  K  K  T  L  S  E  R  A  E  R  #  +  D  P  Y
TTGATCCCTTGATTTTTCTAAGGGTTGGAACAGCTAAATAAGGGTAAGCAAAGAAAACGCTATCTGAACGCGCAGAACGGTAATAGGATCCGTA
0300                |110320                |110340                |110360                |110380
AACTAGGGAACTAAAAAGATTCCCAACCTTGTCGATTTATTCCCATTCGTTTCTTTTGCGATAGACTTGCGCGTCTTGCCATTATCCTAGGCAT
N  I  G  Q  N  K  +  P  N  S  C  S  F  L  P  L  C  L  F  R  +  R  F  A  C  F  P  L  L  I  R
 Q  D  R  S  K  E  L  P  Q  F  L  +  I  L  T  L  L  S  F  A  I  Q  V  R  L  V  T  I  P  D  T
  S  G  K  I  K  R  L  T  P  V  A  L  Y  P  Y  A  F  F  V  S  D  S  R  A  S  R  Y  Y  S  G  Y
```

```
CDS               110399  111366
misc_feature      111365  112612 c  HMMPanther hit to PTHR10074:SF7,PERMEASE-RELATED,score 5
misc_feature      111365  112612 c  HMMPanther hit to PTHR10074, MAJORFACILITATORSUPERFAMILY
misc_feature      111365  112612 c  HMMPanther hit to PTHR10074:SF7, PERMEASE-RELATED,score
misc_feature      111365  112612 c  HMMPanther hit to PTHR10074, MAJOR FACILITATORSUPERFAMIL
misc_feature      111365  112639 c  superfamily hit to SSF103473, MFS general substratetrans
CDS               111448  112173
misc_feature      111464  112603 c  HMMPfam hit to PF07690, MFS_1, score 1.8e-08
misc_feature      111464  112603 c  HMMPfam hit to PF07690, MFS_1, score 1.8e-08
misc_feature      112163  112210 c  PatternScan hit to PS00012,PHOSPHOPANTETHEINE,score NA
misc_feature      112163  112210 c  PatternScan hit to PS00012, PHOSPHOPANTETHEINE,score NA
```

```
FT   CDS             join(110399..110584,110584..111366)
FT                   /gene=""
FT                   /curation="RGMG"
FT                   /similarity="Similar to Corynebacterium diphtheriae NCTC
FT                   13129 Putative secreted protein (321 aa) fasta scores:E():
FT                   e-142, 85% id in 278 aa"
FT                   /product="Putative pseudogene"
FT                   /previous_systematic_id="contig00011.1870"
FT                   /blastp_file="blastp/Cp1002-VF.embl.seq.00322.out"
FT                   /colour=2
FT                   /pseudo
FT                   /systematic_id="Cp1002-v9.1040"
FT                   /pathogenicity="02"
```

Figura 13: Um exemplo de exibição de dados exportados do banco de dados criado com o ER CpDB.

6.1.2.2 Utilização do esquema relacional de banco de dados CpDB

O código fonte do esquema relacional de banco de dados CpDB foi extraído do banco de dados POSTGRESQL, versão 8.4, executando no sistema operacional Linux. Para recriar o esquema relacional de banco de dados CpDB, é necessário:

1) Criar manualmente um banco de dados por meio de qualquer interface de administração de banco de dados postgres, como por exemplo PGADMIN3, com o nome "CpDataBase", na porta de acesso padrão, geralmente a 5432. Usando a mesma interface de administração do banco de dados, crie um esquema relacional de nome "CpDB". O esquema relacional não precisa ser preenchido manualmente com nenhuma tabela visto que o código fonte aqui explicitado vai criar todas as tabelas e funções necessárias;

2) Em uma janela de terminal de comandos do Linux, digitar:

```
psql -f CpDB_schema.sql -U postgres -p 5432 -h localhost CpDB
```

O comando *psql* receberá com o parâmetro -f o arquivo com o *script* de criação do banco de dados. O parâmetro -U específica o usuário que tem permissão de escrever no banco de dados de nome CpDataBase, referenciado indiretamente pela porta do parâmetro -p. O parâmetro -h explicita a máquina onde o banco vazio foi criado e o último parâmetro explicita qual esquema relacional será o destino de criação das entidades presentes no arquivo "CpDB_schema.sql".

3) Agora é necessário gerar os comandos para popular o esquema relacional CpDB, visto que o mesmo foi criado sem dado algum. Essa tarefa é de responsabilidade do programa parseEMBLtoCPDB descrito na sequência.

6.1.2.3 parseEMBLtoCpDB: analisador léxico e semântico para o CpDB

A seção anterior mostrou como criar um banco de dados com o esquema relacional CpDB, porém nenhum dado é inserido no esquema relacional CpDB após a sua criação. Para popular esse banco de dados, é necessário obter os dados de um arquivo no formato EMBL ou GBK. Além de obter os dados, também é preciso formatá-los para inserção no esquema CpDB. Para essa finalidade, criou-se um *parser* que realiza análises léxicas e semânticas sobre a "gramática" do arquivo EMBL. Sempre que um regra dessa gramática EMBL é validade pelo *parser*, então os dados dessa construção gramatical é extraída e escrita em um arquivo texto que pode ser utilizado pelo comando *psql* para inserir dados no banco de dados CpDB. A seguir são listados os códigos léxico, semântico e o arquivo *make* que cria o *parser*. Ao final dessa etapa o resultado obtido será uma arquivo executável, compilado a partir de um código em linguagem de programação C, que ao ser executado sobre um arquivo EMBL vai gerar arquivos em formato texto para popular o banco de dados CpDB. Esse programa foi criado por meio dos criadores de compiladores *flex* e *bison*, similares ao compiladores *lex* e *yacc*, respectivamente.

6.1.2.3.1 Populando o banco de dados CpDB com o *parser*

Após compilar com sucesso o *parser*, basta executá-lo passando como parâmetro o arquivo EMBL com a anotação funcional. Por exemplo, supondo que existem dois arquivos EMBL no mesmo diretório do *parser*, um arquivo para cada *contig* anotado, para criarmos um único arquivo de saída com os resultados da execução do *parser* basta digitar:

```
./parseEMBLtoCpDB ./SC-01.embl > ./out

./parseEMBLtoCpDB ./SC-02.embl >> ./out
```

Esse artifício de redirecionamento de saída do processamento serve para criar um histórico de toda a execução do *parser*, porém o arquivo a ser utilizado para popular o banco de dados não é o arquivo “out”, e sim um arquivo com extensão “sql” para cada entidade existente no banco. A execução do *parser* sempre concatena os resultados da execução atual com as execuções anteriores e por isso o arquivo *make* tem um procedimento específico para apagar os arquivos com extensão “sql” antes de uma próxima execução do *parser*.

Para popular o banco de dados, foi necessário que as instruções dos comandos SQL fossem executadas no banco de dados. São centenas de comandos no formato INSERT INTO <*table*> (<*values*>) que vão popular o banco de dados. A quantidade de arquivos com extensão “sql” gerada pelo *parser* depende da quantidade de entidades EMBL presentes nos arquivos da anotação funcional. Se houverem apenas CDS's, os únicos arquivos que devem ser criados são “insert.gene.sql“ e “insert.multipos.sql” caso hajam pseudogenes anotados. Para cada um dos arquivos criados basta digitar:

```
psql -f insert.gene.sql     -U postgres -p 5432 -h localhost CpDB
psql -f insert.multipos.sql -U postgres -p 5432 -h localhost CpDB
     ...
```

Ao término da execução bem sucedida desses comandos, as informações do arquivo EMBL terão sido cadastradas no CpDB e estarão prontas para serem manipulados por meio da linguagem SQL, utilizando qualquer programa de interface com o banco de dados, seja por meio de interface gráfica (PGADMIN3) ou apenas por linhas de comando (psql).

6.1.2.3.2 Extraindo o formato EMBL do banco de dados

Estando o banco de dados CpDataBase com o esquema relacional CpDB_v1 populado, com dados oriundos do formato EMBL, uma possível pesquisa em SQL é a exportação de todas as CDS's com coordenadas adaptadas para outro genoma. Um exemplo da utilidade dessa exportação de dados com coordenadas adaptadas é a correção de *frame-shifts* de regiões codificadores de um genoma. Considere que a anotação funcional desse genoma está correta, mas pequenas inserções ou deleções manuais de bases ocorreram na fita de DNA, após a verificação das leituras do(s) sequenciamento(s) que deram origem à fita consenso de DNA do genoma. Nesse caso pode ser feita a exportação de todas as CDS's do banco com suas coordenadas movidas por uma quantidade fixa de bases, quantidades essas mapeadas pela diferença entra a fita anterior e a posterior às correções manuais de *frame-shifts*.

Outra situação de exportação de dados do banco em formato EMBL ocorre quando, por exemplo, o genoma de uma espécie ou linhagem de um organismo é utilizada para criar uma primeira versão da anotação funcional de outro genoma. Nesse caso, é necessário mapear as proteínas anotadas do genoma de origem para as CDS's do genoma de destino. Tal procedimento pode ser feito por intermédio de um alinhamento de sequências no qual os resultados de alinhamento de cada proteína do genoma de origem é mapeado no genoma de destino. As coordenadas das proteínas do genoma de origem podem ser trocadas pelas coordenadas mapeadas em um alinhamento feito no genoma de destino, durante a exportação de dados por uma consulta SQL submetida ao banco de dados representado pelo esquema relacional CpDB. Para esse propósito o esquema CpDB possui uma tabela denominada *blasthits*, capaz de armazenar o resultado de um alinhamento de sequências assim como ele é gerado pelo programa *blastall* quando utilizado o parâmetro "*-m 8*" ou m8. O parâmetro m8 formata a saída do programa *blastall* em 12 colunas de dados separadas por tabulações contendo principalmente, nesse caso, a identificação da proteína da qual será reaproveitada a anotação funcional (origem), a identificação da proteína que vai receber a anotação funcional (destino) e as coordenadas inicial e final da proteína de destino na fita de DNA do genoma de destino, coordenadas essas que vão ser as da proteína que será importada no genoma de destino.

A última frase do parágrafo anterior subentende um detalhe importante para o processo de transferência da anotação funcional entre dois genomas. A opção por realizar um *blastp* (alinhamento proteico), entre o proteoma de origem para o proteoma de destino, subentende que os resultados apresentados no arquivo tabular m8 do blast não terão as coordenadas iniciais e finais da proteína de destino relativas à proteína na fita de DNA de destino. Para exemplificar, suponha um genoma de origem com toda a sua anotação funcional finalizada e curada manualmente. Espera-se que o genoma de origem possua uma alta similaridade proteica com o genoma de destino que possui apenas uma predição de ORF's realizada. Desse modo, espera-se também que grande parte do proteoma de origem seja quase que idêntico ao do proteoma de destino. Para aproveitar a anotação funcional entre os dois genomas, um banco de dados de proteínas no formato *blast* deve ser criado com o proteoma de destino. Ambos os proteomas são alinhados utilizando como *query* o proteoma de origem e como *subject* o banco de dados do proteoma de destino. Agora considere que uma proteína denominada ORIGEM_0145 possui o tamanho de 153 aminoácidos. As coordenadas dessa proteína no arquivo tabular m8 serão o intervalo de 1 a 153. Considere também que essa proteína de origem teve um alinhamento de similaridade perfeito com a proteína DESTINO_0678. Como o *subject* do alinhamento *blast* foi um arquivo com as proteínas do genoma de destino, então as coordenadas armazenadas no resultado tabular m8 do *blast*, para o *subject* DESTINO_0689 também são de 1 a 153. Essa situação impediria a utilização da anotação entre os dois genomas. Para que a anotação seja reaproveitada entre os genoma de origem e destino, é necessário que as coordenadas do *subject* representem a posição da proteína DESTINO_0689 na fita de DNA do genoma de destino. Dessa forma, ter-se-ia as coordenadas ORIGEM_0145 mapeadas na proteína DESTINO_0689, por exemplo, nas coordenadas 1152323 até 1152476 da fita de destino, independente das coordenadas da proteína do genoma de origem.

Existem duas possibilidades para resolver o problema relatado no parágrafo anterior. A primeira opção é fazer o alinhamento assim como demonstrado nesse exemplo, mas no momento da exportação dos dados incluir em uma cláusula JOIN do SQL, uma tabela que contêm as coordenadas de todas as proteínas do genoma de destino mapeadas na fita de DNA de destino. Assim, ao invés de utilizar os dados de posição inicial e final da proteína alinhada da tabela *blasthits*, esses dados são obtidos de uma terceira tabela contendo as posições corretas da proteína alinhada na fita de destino. Aqui não é documentado como produzir os dados dessa terceira tabela, mas uma possibilidade é uma conjugação dos programas *grep* e *cut* do Linux, o primeiro para isolar os cabeçalhos *fasta* do arquivo de proteínas do genoma de destino e o segundo para filtrar apenas as colunas de identificação posição inicial e final da proteína na fita de DNA, informação sempre disponível quando uma arquivo *fasta* de proteínas é criado a partir de um EMBL.

A segunda opção para resolver o problema apresentado no exemplo de reaproveitamento de anotação funcional entre dois genomas, é a realização de um alinhamento entre as sequências proteicas do genoma de origem e a fita de DNA do genoma de destino. Ao utilizarmos o *tblastn* como programa de alinhamento, as sequências proteicas da origem serão alinhadas contra a fita de DNA de destino e traduzidas em suas seis possibilidades de codificação proteicas. Dessa forma, o arquivo tabular gerado com o formato m8 do *blastall* vai conter as coordenadas corretas da fita de DNA do destino e uma terceira tabela de dados não é necessária para exportar a anotação funcional entre os genomas.

O código a seguir foi explicado na Figura 11, sendo uma consulta em linguagem SQL utilizada na exportação dos dados da anotação funcional entre duas versões de montagem do genoma da *C. pseudotuberculosis*, linhagem C231. Nessa situação não foi preciso realizar alinhamento de sequências para determinar novas posições das CDS's de origem, visto que para serem exportadas para o destino todas as coordenadas das CDS's de origem eram incrementadas ou decrementadas por uma quantidade fixa de bases com consequência da translocação de grandes blocos do genoma entre as duas versões da montagem. Para que esse comando SQL produza a exportação da anotação funcional com as coordenadas alteradas, é necessário que as funções GETMULTIPOS e TRANSFER_COORD, listadas após o comando de exportação, sejam substituídas no esquema original CpDB. Comentários a respeito das estruturas presentes na consulta estão explicados na Figura 11:

```
select
'FT   CDS             complement(' || getmultipos(gene.systematic_id) || ')',
'FT                   /systematic_id="' || gene.systematic_id ||'"',
'FT                   /gene="' || gene.name || '"',
'FT                   /curation="' || gene.curation || '"' ,
'FT                   /similarity="' || gene.similarity || '"' ,
'FT                   /note="' || gene."OpTu" || '"' ,
'FT                   /product="' || gene.product || '"' ,
'FT                   /previous_systematic_id="' || gene.previous_systematic_id || '"',
'FT                   /colour=' || decidecolor(gene.pseudogene, gene.pathogenicity) || ';',
getallgo(gene.systematic_id), getalldomain(gene.systematic_id) ,getsignal(gene.systematic_id)
from gene
where gene.orientation='-'
UNION
select
'FT   CDS             ' || getmultipos(gene.systematic_id),
'FT                   /systematic_id="' || gene.systematic_id ||'"',
'FT                   /gene="' || gene.name || '"',
'FT                   /curation="' || gene.curation || '"' ,
'FT                   /similarity="' || gene.similarity || '"' ,
'FT                   /note="' || gene."OpTu" || '"' ,
'FT                   /product="' || gene.product || '"' ,
'FT                   /previous_systematic_id="' || gene.previous_systematic_id || '"',
'FT                   /colour=' || decidecolor(gene.pseudogene, gene.pathogenicity) || ';',
getallgo(gene.systematic_id), getalldomain(gene.systematic_id), getsignal(gene.systematic_id)
from gene
where gene.orientation='+'
```

Funções utilizadas pela consulta de exportação de dados modificadas para converter coordenadas de CDS's:

```
CREATE OR REPLACE FUNCTION getmultipos(nome character varying) RETURNS text
    AS $$
DECLARE
    inicio integer;
    fim integer;
    cursor_multipos CURSOR FOR  SELECT pos_begin, pos_end FROM MULTIPOS where       gene_systematic_id = nome;
    cursor_pos CURSOR FOR  SELECT pos_begin, pos_end FROM GENE where systematic_id = nome;
    resultado text;
```

```
BEGIN
    resultado = '';
    OPEN cursor_multipos;
    LOOP
      FETCH cursor_multipos INTO inicio, fim;
      IF FOUND THEN
            IF resultado <> '' THEN
                  resultado = resultado || ',';
            END IF;
            resultado = resultado || transfer_coord(inicio) || '..' || transfer_coord(fim) ;
      END IF;
      IF not FOUND THEN EXIT;
      END IF;
    END LOOP;
    CLOSE cursor_multipos;
    IF resultado = '' THEN
      OPEN cursor_pos;
        FETCH cursor_pos INTO inicio, fim;
        resultado = resultado || transfer_coord(inicio) || '..' || transfer_coord(fim) ;
      CLOSE cursor_pos;
    END IF;
    RETURN resultado;
END
$$
    LANGUAGE plpgsql;
CREATE OR REPLACE FUNCTION transfer_coord(par_position integer) RETURNS integer
    AS $$
DECLARE
      new_position int;
BEGIN
      IF par_position > 622761 THEN
            new_position = par_position - 622761;
      ELSE
            new_position = par_position + 1697840;
      END IF;
      return new_position;
END
$$
    LANGUAGE plpgsql;
```

Após executar uma consulta, solicitando ao PGADMIN3 que exporte o resultado para um arquivo texto, bastará formatar o arquivo resultante da execução do SQL para que esse possa ser anexado a um EMBL. O arquivo texto exportado pelo PGADMIN3 poderá possuir pontos e vírgulas separando todas as colunas de dados da consulta SQL esses devem ser removidos por marcas de quebra de linha de modo a separar as várias *features* do arquivo EMBL em linhas distintas. Após garantir que todos as *features* estão devidamente representadas no arquivo texto resultado da exportação dos dados do banco de dados, bastará concatenar esse arquivo com um arquivo EMBL contendo somente a fita de DNA do genoma de destino, gerando assim um novo arquivo EMBL que, além fita de DNA, conterá toda a anotação funcional possível de ser mapeada entre os genomas de origem e destino.

Em caso de exportação de dados utilizando resultados de alinhamentos de sequências, uma decisão importante que deve ser tomada diz respeito a qual o limite de corte para importar uma anotação de uma genoma de origem para outro de destino. Pode ser decidido, por exemplo, que apenas os alinhamentos com mais de 90% de identidade e com mais de 90% do tamanho da proteína de origem sejam aptos a serem migrados para o genoma de destino. Tal consideração pode ser explicitada no corpo da consulta SQL, onde na cláusula WHERE tenha restrições quanto aos valores dos campos da tabela *blasthits,* tabela essa que estaria em uma cláusula JOIN com a tabela GENE.

6.1.2.4 MEDPIPE 1.0

O *pipeline* web exibido na Figura 14, foi criado para ser um meio prático de aferição da estatística *Mature Epitope Density* (MED) sobre genomas bacterianos que possuam ao menos uma predição gênica inicial. O *pipeline* aceita como entrada de dados sequências *fasta* de aminoácidos, seja por meio da janela de entrada de dados ou por meio da seleção de um arquivo *multifasta*. O usuário pode selecionar os parâmetros de seu organismo como o tamanho estimado da parede celular, e se o organismo é Gram positivo ou negativo. O *pipeline*, cuja interface HTML possui código PHP, fará uma verificação a respeito do tamanho máximo permitido para envio de dados ao servidor e a conformidade do arquivo com um padrão *fasta*. Passados essas verificações, o *pipeline* exibido na Figura 14 aciona primeiramente o programa SurfG *plus* 1.0 que, por vez, acionará suas dependências como o SignalP, LipoP, HMMSEARCH, TMHMM. Finalizada a execução do SurfG, apenas as proteínas classificadas como exportadas passam para a próxima fase do *pipeline*, a análise de epitopos da porção predita como madura de uma proteína. Antes que o programa de predição de epitopos seja acionado cada proteína exportada é editada para ficar sem as porções não maduras, a saber, o peptídeo sinal para proteínas preditas como secretadas e as porções preditas como citoplasmáticas e integrais à membrana celular para as proteínas preditas como potencialmente expostas na superfície. O programa NetMHC 3.0 (Lundegaard e cols., 2008) analisará todas as janelas possíveis de 9-mer frente à possibilidade de serem epitopos com capacidade de ligação a qualquer um dos 55 alelos possíveis, em sua maioria alelos humanos, por intermédio desse programa de predição de epitopos. Ao termino da execução, as proteínas exportadas encontradas em um genoma sob análise são exibidos em ordem crescente da estatística MED. Também são listados o numerador, o denominador e afinidade média de ligação de um epitopo à molécula de MHC, dados utilizados no cálculo da MED. Também é listada a estatística denominada FOLD, com o mesmo significado do termo em inglês *fold,* representando em quantas vezes uma quantidade é maior do que outra, ou seja, quantas vezes a possibilidade de se encontrar epitopos em uma proteína é maior que a quantidade de epitopos preditos. Essa estatística FOLD fornece uma pista quanto a confiabilidade da estatística MED, assim como demonstrado no artigo científico da seção 3.2.5.1. Em resumo, se FOLD for menor do que 3 a estatística MED tem maiores chances de estar evidenciado uma possível proteína antigênica. Também é exibida a predição do local subcelular de proteínas exportadas, sob o rótulo SEC para secretado ou PSE para potencialmente exposto na superfície bacteriana.

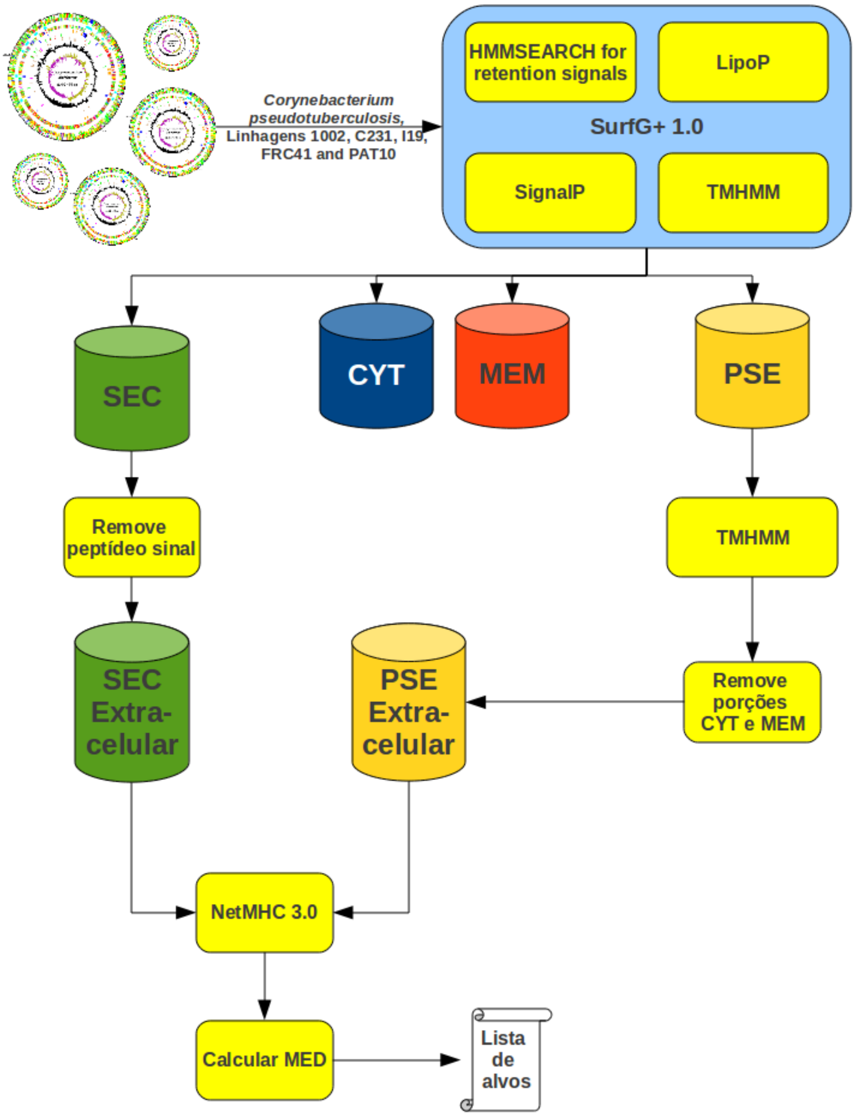


Figura 14: Esquema de predição de proteínas exportadas e cálculo de impacto na imunogenicidade.

6.1.3 Outros códigos fonte

Este conjunto de programas está disponível na internet, por meio de uma licença GNU *General Public License* (Licença Pública Geral) através do sítio http://sourceforge.net/projects/cpdb/:

6.1.3.1 splitfasta.c

Este programa recebe como parâmetro um arquivo em formato texto contendo várias sequências *fasta* e separa cada uma das sequências em arquivos texto individuais com o mesmo nome da primeira palavra do cabeçalho *fasta* de cada sequência.

Para compilar e executar o programa “splitfasta.c”, digite:

```
gcc splitfasta.c -o splitfasta
./splitfasta <arquivo multifasta>
```

A execução do programa *splitfasta* gera uma arquivo para cada cabeçalho do *multifasta*, contendo apenas uma sequência *fasta*.

6.1.3.2 bacparser

É um analisador léxico e semântico para criação de um mapa de BAC com dados oriundos de cromatogramas. Foram criados dois programas para gerar os mapas de *BAC ends* desse trabalho, exibidos na Figura 15 e na Figura 16. O primeiro programa é o *parser* composto por dois códigos fonte, analisadores léxico, semântico.

Os arquivos “*bacparser.l*” , “*bacparser.y*” e “*make*” devem ser estar localizados no mesmo diretório para que seja possível realizar a compilação com sucesso do código fonte e a geração do código executável do programa, que aqui é denominado *bacparser*. No programa “*make*”, quando o programa “*bacparser*” é compilado com sucesso ele também é executado em dois arquivos resultantes de alinhamentos de sequências de nucleotídeos. São os resultados tabulares (m8), entre os dois *contigs* da versão 4 do genoma e as sequências de BAC ends extraídas de cromatogramas. O resultado dessa execução é redirecionada para os arquivos “*out.12*” e “*out.21*”.

Ao executarmos o *bacparser* sobre um arquivo tabular resultante do alinhamento das sequências de nucleotídeos de BAC ends contra sequências de *contigs* de um genoma, é criada uma lista com possíveis BAC's, de acordo com as estatísticas de tamanho máximo e mínimo que foram parametrizados diretamente no código fonte do programa. No exemplo de execução do *bacparser*, no corpo do programa “*make*”, esses arquivos são “*out.12*” e “*out.21*”, arquivos que servem de entrada de dados para o programa que determina quais

BAC ends percorrem um caminho com interseções que permitam cobrir todo o genoma com o qual houve alinhamento. Essa lista de possíveis *contigs* possibilita criar um mapa assim como exibido na Figura 15. Porém, nessa figura ainda existem muitos BAC's e é desejada uma quantidade mínima de BAC's que permita cobrir todo o genoma. Para esse fim, criou-se o programa "*tilingpath.c*", que percorre a lista de todos os BAC's selecionando apenas um conjunto reduzido que nos permita cobrir a maior parte do genoma, assim como exibido na Figura 16. Este é o código do programa do caminho mínimo de BAC ends:

Para compilar e executar o programa "tilingpath.c", digite:

```
gcc tilingpath.c -o tilingpath
./tilingpath <arquivo de saida do bacparser>
```

A execução do programa *tilingpath* gera uma lista menor de BAC *ends*, que podem ser utilizados para produzir o gráfico exibido na Figura 16.

6.2 Montagem e anotação dos genomas de *C. pseudotuberculosis*

A montagem e anotação do primeiro genoma da *C. pseudotuberculosis*, referente à linhagem 1002, foi uma das etapas desafiadoras do projeto pangenoma desse organismo. Foi fundamental garantir a o máximo de acurácia no primeiro genoma a ser montado ou anotado, para evitar o risco de propagação de erros em diversos genomas que viriam a utilizar o genoma da linhagem 1002 como modelo.

Como consequência do esforço para garantir a finalização do primeiro genoma de *C. pseudotuberculosis,* bem como a sua acurácia, utilizaram-se as abordagens: análise da utilização de códons e o mapeamento de *BAC*'s (*Bacterial Chromossomes*) na sequência parcial do genoma que estava sendo montado. A primeira abordagem propiciou evidência a respeito de *contigs* para os quais se suspeitava serem oriundos de outro organismo, afastando a possibilidade de contaminação na amostra de DNA sequenciada; a segunda abordagem propiciou uma lista mínima de *BAC end*'s, cobrindo todo o genoma da linhagem 1002, para um eventual processo de fechamento de *gaps*. Para gerar esses resultados, foi necessária a criação de programas de computador específicos. As subseções a seguir documentam esses programas e o local de armazenamento na *web* dos mesmos.

6.2.1 Mapas com sequências de BAC's

A geração de um mapa de BAC end's foi um suporte no processo de verificação da qualidade da montagem do primeiro genoma de *C. pseudotuberculosis* realizado por nosso grupo de pesquisa. Através deste resultado foi possível estimar o tamanho médio dos intervalos da sequência de DNA do genoma para os quais não se conseguiu, durante a montagem, uma sequência de consenso entre os vários fragmentos oriundos da etapa de sequenciamento do genoma. Esses intervalos sem consenso, denominados *gaps*, foram preditos por esse trabalho como sendo de proporções ordinárias que seriam facilmente resolvidos por técnicas laboratoriais corriqueiras. Essa conclusão nos forneceu evidências de que os resultados obtidos sobre a montagem do genoma estavam dentro do esperado e que a montagem final, sem os *gaps*, não estava distante de se tornar realidade.

Um total de 1104 sequências de BAC *ends* foram alinhadas sobre uma fita de DNA representada pela junção dos dois *supercontigs* que existiam como consequência da montagem da versão três do genoma de *C. pseudotuberculosis.* Essas sequências por serem relativamente curtas, variando de 300 a 800 nucleotídeos, podiam se alinhar em mais de uma região dos *supercontigs*. Assim sendo uma análise estatística foi realizada para concluir qual poderia ser o tamanho médio dos BAC's dado combinações que eram aceitáveis para os pares de sequências. O tamanho médio foi calculado em 51 mil nucleotídeos. Nesse trabalho, o conceito de tamanho aceitável foi definido por meio de observação dos resultados das combinações possíveis. Por meio de observações estatísticas dos resultados obtidos na montagem *in silico* do mapa de BAC's de *C. pseudotuberculosis* determinou-se como 80 mil nucleotídeos o tamanho máximo aceito para um BAC. Esse limite foi evidenciado ao ser observado que acima desse patamar, os BAC's montados *in silico* possuíam tamanhos superiores aos conhecidos para essa biblioteca que era entre 25 a 120 Kb (Dorella e cols. 2006).

A Figura 15 mostra o alinhamento de todos os BAC's considerados viáveis sobre os dois *supercontigs* da versão três do genoma de *C. pseudotuberculosis* linhagem 1002. Nessa figura os tamanhos estimados dos *gaps* não foram representados, visto que ainda não eram conhecidos. Também não foram representados os BAC's que estariam transpondo ou ancorados na extremidade dos *gaps*.

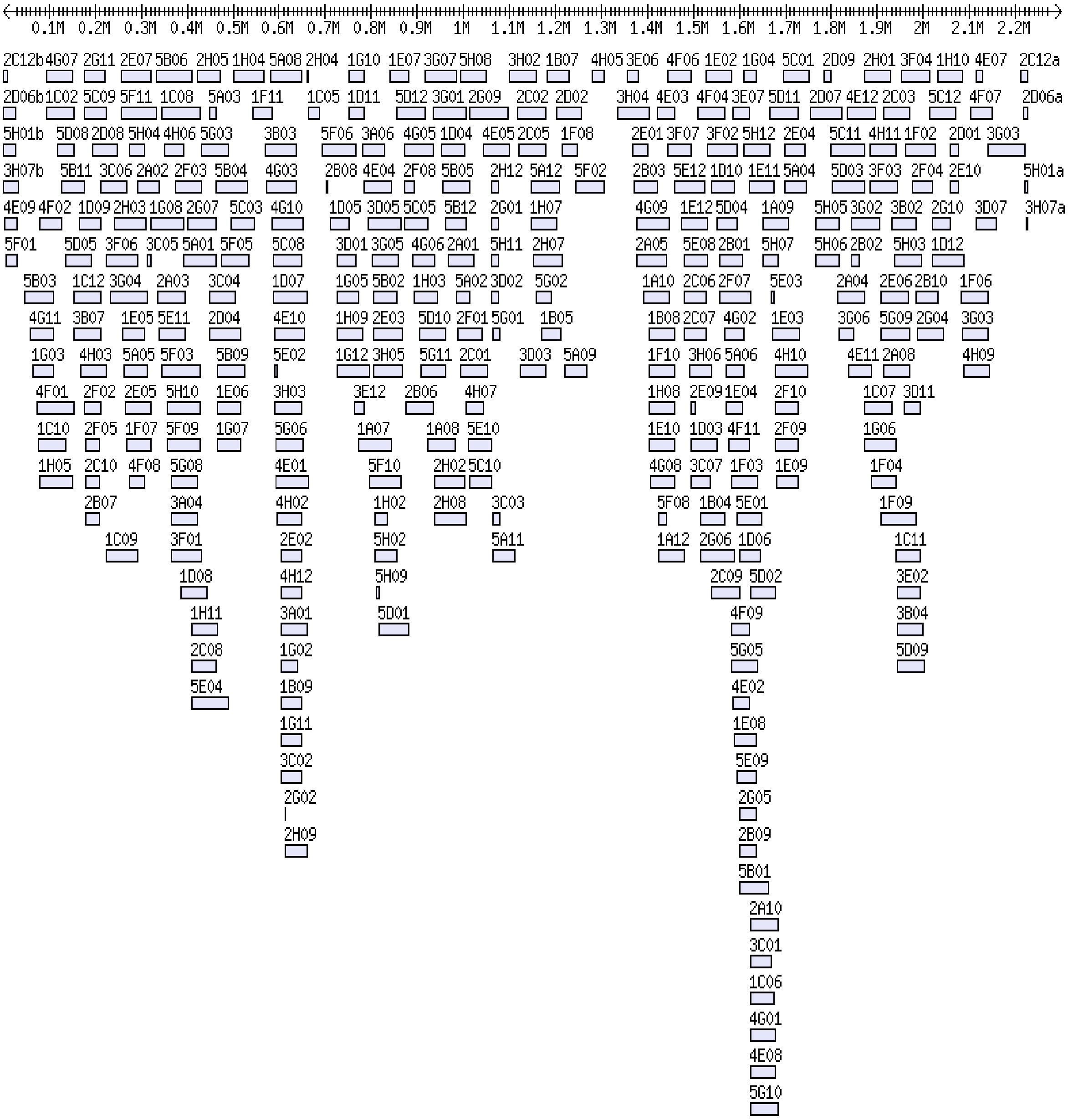

0.1M 0.2M 0.3M 0.4M 0.5M 0.6M 0.7M 0.8M 0.9M 1M 1.1M 1.2M 1.3M 1.4M 1.5M 1.6M 1.7M 1.8M 1.9M 2M 2.1M 2.2M
2C12b 4G07 2G11 2E07 5B06 2H05 1H04 5A08 2H04 1G10 1E07 3G07 5H08 3H02 1B07 4H05 3E06 4F06 1E02 1G04 5C01 2D09 2H01 3F04 1H10 4E07 2C12a
2D06b 1C02 5C09 5F11 1C08 5A03 1F11 1C05 1D11 5D12 3G01 2G09 2C02 2D02 3H04 4E03 4F04 3E07 5D11 2D07 4E12 2C03 5C12 4F07 2D06a
5H01b 5D08 2D08 5H04 4H06 5G03 3B03 5F06 3A06 4G05 1D04 4E05 2C05 1F08 2E01 3F07 3F02 5H12 2E04 5C11 4H11 1F02 2D01 3G03
3H07b 5B11 3C06 2A02 2F03 5B04 4G03 2B08 4E04 2F08 5B05 2H12 5A12 5F02 2B03 5E12 1D10 1E11 5A04 5D03 3F03 2F04 2E10 5H01a
4E09 4F02 1D09 2H03 1G08 2G07 5C03 4G10 1D05 3D05 5C05 5B12 2G01 1H07 4G09 1E12 5D04 1A09 5H05 3G02 3B02 2G10 3D07 3H07a
5F01 5D05 3F06 3C05 5A01 5F05 5C08 3D01 3G05 4G06 2A01 5H11 2H07 2A05 5E08 2B01 5H07 5H06 2B02 5H03 1D12
5B03 1C12 3G04 2A03 3C04 1D07 1G05 5B02 1H03 5A02 3D02 5G02 1A10 2C06 2F07 5E03 2A04 2E06 2B10 1F06
4G11 3B07 1E05 5E11 2D04 4E10 1H09 2E03 5D10 2F01 5G01 1B05 1B08 2C07 4G02 1E03 3G06 5G09 2G04 3G03
1G03 4H03 5A05 5F03 5B09 5E02 1G12 3H05 5G11 2C01 3D03 5A09 1F10 3H06 5A06 4H10 4E11 2A08 4H09
4F01 2F02 2E05 5H10 1E06 3H03 3E12 2B06 4H07 1H08 2E09 1E04 2F10 1C07 3D11
1C10 2F05 1F07 5F09 1G07 5G06 1A07 1A08 5E10 1E10 1D03 4F11 2F09 1G06
1H05 2C10 4F08 5G08 4E01 5F10 2H02 5C10 4G08 3C07 1F03 1E09 1F04
2B07 3A04 4H02 1H02 2H08 3C03 5F08 1B04 5E01 1F09
1C09 3F01 2E02 5H02 5A11 1A12 2G06 1D06 1C11
1D08 4H12 5H09 2C09 5D02 3E02
1H11 3A01 5D01 4F09 3B04
2C08 1G02 5G05 5D09
5E04 1B09 4E02
1G11 1E08
3C02 5E09
2G02 2G05
2H09 2B09
5B01
2A10
3C01
1C06
4G01
4E08
5G10

Figura 15: BAC's criados por meio de alinhamentos de extremidades de BAC's e os dois *supercontigs* da versão três do genoma de *C. pseudotuberculosis*.

Em seguida, foi realizada uma seleção da menor quantidade de BAC's possíveis para transpor o máximo possível do genoma, assim como exibido na Figura 16. O ancoramento de BAC's nas extremidades dos gaps conhecidos, além do ancoramento de BAC's entre duas extremidades transpondo os *gaps*, bem como as estimativas para tamanhos máximo e médio de BAC's foram decisivos elaborar uma estimativa sobre qual seriam os tamanhos dos prováveis gaps. Terminadas essas análises havia evidências estatísticas de que esses gaps não passariam de tamanhos que variavam de menos de 10 Kb até no máximo 20 Kb, tamanhos esses considerados normais para serem resolvidos por técnicas clássicas de fechamento de gaps como *primer walking* ou *longrange PCR*. Motivados por esses resultados e pela existência de intervalos supostamente redundantes no genoma, antes de nosso grupo partir de fato para técnicas de fechamento de *gaps*, foi feita uma nova tentativa de montagem do genoma por meio de uma calibração mais fina de parâmetros dos programas de montagem. Essa remontagem deu origem à versão quatro do genoma de *C. pseudotuberculosis*, linhagem 1002 na qual tanto os *gaps* quanto as regiões de repetição não mais existiam.

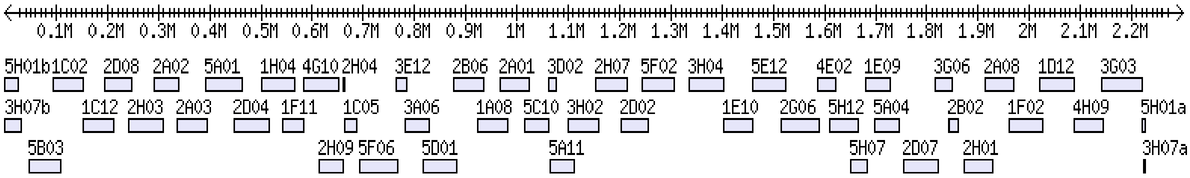


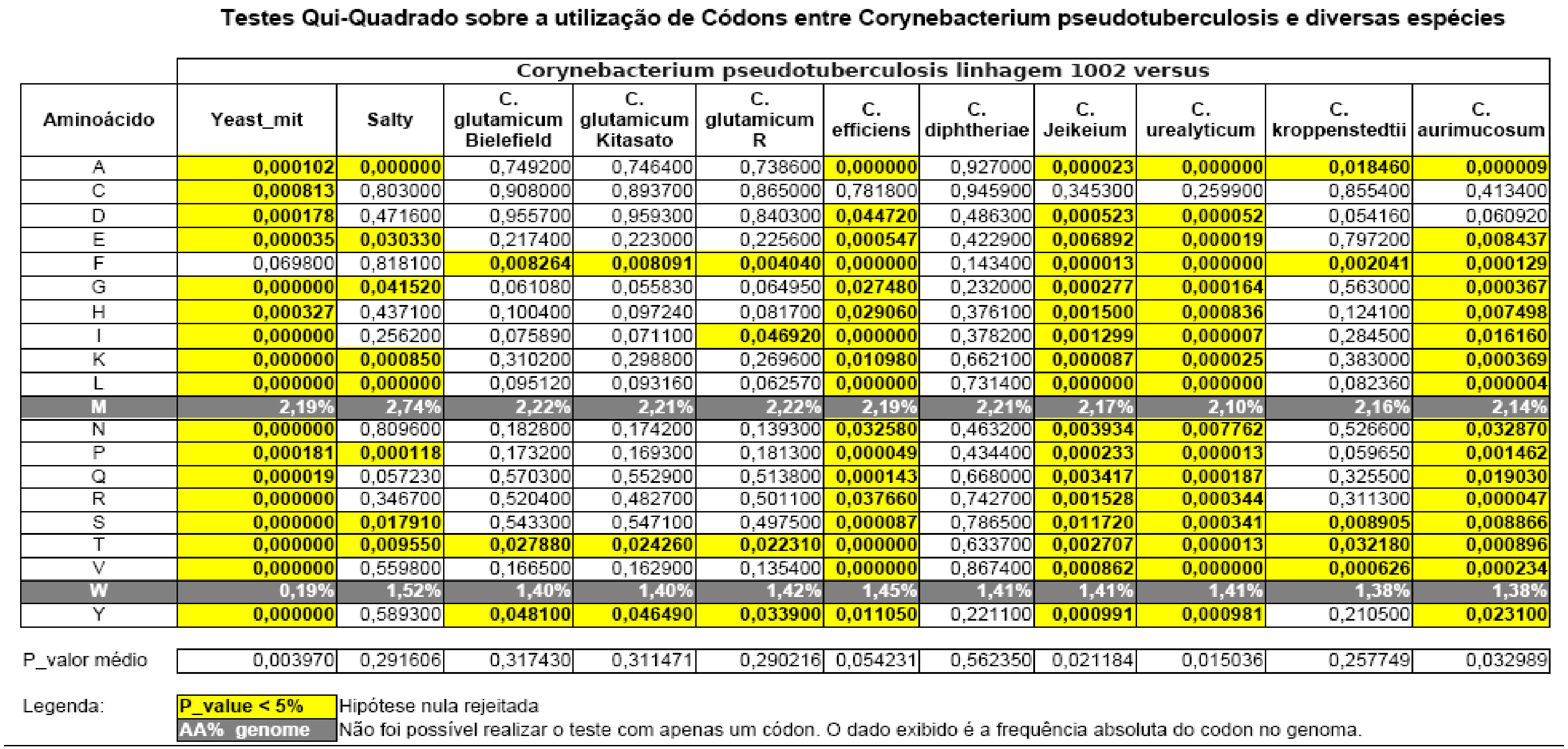
Figura 16: Biblioteca por ordenamento clonal de BAC's gerada com a menor quantidade de BAC's cobrindo a maior extensão possível do genoma de *C. pseudotuberculosis*.

Pela análise da Figura 16 é possível perceber que quase a totalidade do genoma de *C. pseudotuberculosis* foi coberta por apenas 49 BAC's.

6.2.2 Análise da utilização de códon da *C. pseudotuberculosis*

Na seção anterior foi relatado que durante a montagem do genoma de *C. pseudotuberculosis* houve diferentes versões do genoma de *C. pseudotuberculosis*. Na versão 2 houve um *supercontig* que não estava se alinhando nas extremidades de nenhum outro. Posteriormente, foi descoberto que o impedimento para esse alinhamento ocorreu porque a similaridade entre os *supercontigs* não ocorria quando comparadas bases nucleotídicas, mas ocorria com sucesso quando comparadas proteínas codificadas pelas extremidades dos *supercontigs*. Porém, esse obstáculo revelou-se uma excelente oportunidade para que fosse extraída mais informação a respeito desse genoma. Durante o processo de investigação do porque os *supercontigs* não se alinhavam foi feita uma análise da utilização de códon do genoma da *C. pseudotuberculosis* e isoladamente do *supercontig* que não se alinhava com os demais por meio de similaridade de nucleotídeos. Para se certificar que a análise da utilização de códon feita entre o genoma e o *supercontig* seria relevante, introduziu-se também controles positivos e negativos na comparação, nesse caso o genoma de outras bactérias e de uma Levedura. O resultado foi tão revelador para o grupo de trabalho que essa análise foi cogitada para o manuscrito principal do genoma de *C. pseudotuberculosis* (Ruiz e cols., 2011). É relevante a exibição dessa análise de utilização de códon como resultado desse trabalho de doutorado visto que a vacinologia reversa depende da obtenção da sequência correta do genoma para produzir análises com resultados igualmente corretos (Bambini e Rappuoli, 2009; Rinaudo e cols., 2009), bem como acarretar evidências que possam justificar escolhas de variáveis em análises de genômica comparativa e confecção de vacinas de DNA com a utilização de códons otimizada (Uchijima e cols., 1998).

O uso de códon de tradução de organismos pode fornecer evidências acerca de ancestralidade. Existe a hipótese sobre a ancestralidade de *C. diphtheriae* e *C. pseudotuberculosis*, bem como entre essas duas espécies e os três genomas da *C. glutamicum* sequenciadas, a saber, Bielefield, Kitasato e R (D'Afonseca e cols., 2009). Havia também suspeitas levantadas por experimentos de clonagem de realizados pelo nosso grupo de pesquisa (comunicação pessoal) sobre as diferenças na utilização de códon entre organismos do gênero *Corynebacterium*. Utilizando-se do genoma de *C. pseudotuberculosis* 1002 realizaram-se testes estatísticos para investigar essas hipóteses. Os resultados destes testes são apresentados na Tabela 6. Os testes foram realizados considerando a fração de probabilidade de ocorrência de cada códon, de cada aminoácido, de *C. pseudotuberculosis* linhagem 1002 contra a frequência de ocorrência do respectivo códon em organismos de comparação. Na Tabela 6, as células em amarelo e negrito representam hipótese nula rejeitada no tocante à utilização de um códon contra o mesmo códon no genoma representado na coluna. Células sem negrito representam hipótese nula aceita, ou seja, existe evidência estatística em nível de 5% de que a utilização de códons entre a espécie representada na coluna é igual à de *C. pseudotuberculosis*.



Testes Qui-Quadrado sobre a utilização de Códons entre Corynebacterium pseudotuberculosis e diversas espécies

	Corynebacterium pseudotuberculosis linhagem 1002 versus										
Aminoácido	**Yeast_mit**	**Salty**	**C. glutamicum Bielefield**	**C. glutamicum Kitasato**	**C. glutamicum R**	**C. efficiens**	**C. diphtheriae**	**C. Jeikeium**	**C. urealyticum**	**C. kroppenstedtii**	**C. aurimucosum**
A	**0,000102**	**0,000000**	0,749200	0,746400	0,738600	**0,000000**	0,927000	**0,000023**	**0,000000**	**0,018460**	**0,000009**
C	**0,000813**	0,803000	0,908000	0,893700	0,865000	0,781800	0,945900	0,345300	0,259900	0,855400	0,413400
D	**0,000178**	0,471600	0,955700	0,959300	0,840300	**0,044720**	0,486300	**0,000523**	**0,000052**	0,054160	0,060920
E	**0,000035**	**0,030330**	0,217400	0,223000	0,225600	**0,000547**	0,422900	**0,006892**	**0,000019**	0,797200	**0,008437**
F	0,069800	0,818100	**0,008264**	**0,008091**	**0,004040**	**0,000000**	0,143400	**0,000013**	**0,000000**	**0,002041**	**0,000129**
G	**0,000000**	**0,041520**	0,061080	0,055830	0,064950	**0,027480**	0,232000	**0,000277**	**0,000164**	0,563000	**0,000367**
H	**0,000327**	0,437100	0,100400	0,097240	0,081700	**0,029060**	0,376100	**0,001500**	**0,000836**	0,124100	**0,007498**
I	**0,000000**	0,256200	0,075890	0,071100	**0,046920**	**0,000000**	0,378200	**0,001299**	**0,000007**	0,284500	**0,016160**
K	**0,000000**	**0,000850**	0,310200	0,298800	0,269600	**0,010980**	0,662100	**0,000087**	**0,000025**	0,383000	**0,000369**
L	**0,000000**	**0,000000**	0,095120	0,093160	0,062570	**0,000000**	0,731400	**0,000000**	**0,000000**	0,082360	**0,000004**
M	**2,19%**	**2,74%**	**2,22%**	**2,21%**	**2,22%**	**2,19%**	**2,21%**	**2,17%**	**2,10%**	**2,16%**	**2,14%**
N	**0,000000**	0,809600	0,182800	0,174200	0,139300	**0,032580**	0,463200	**0,003934**	**0,007762**	0,526600	**0,032870**
P	**0,000181**	**0,000118**	0,173200	0,169300	0,181300	**0,000049**	0,434400	**0,000233**	**0,000013**	0,059650	**0,001462**
Q	**0,000019**	0,057230	0,570300	0,552900	0,513800	**0,000143**	0,668000	**0,003417**	**0,000187**	0,325500	**0,019030**
R	**0,000000**	0,346700	0,520400	0,482700	0,501100	**0,037660**	0,742700	**0,001528**	**0,000344**	0,311300	**0,000047**
S	**0,000000**	**0,017910**	0,543300	0,547100	0,497500	**0,000087**	0,786500	**0,011720**	**0,000341**	**0,008905**	**0,008866**
T	**0,000000**	**0,009550**	**0,027880**	**0,024260**	**0,022310**	**0,000000**	0,633700	**0,002707**	**0,000013**	**0,032180**	**0,000896**
V	**0,000000**	0,559800	0,166500	0,162900	0,135400	**0,000000**	0,867400	**0,000862**	**0,000000**	**0,000626**	**0,000234**
W	**0,19%**	**1,52%**	**1,40%**	**1,40%**	**1,42%**	**1,45%**	**1,41%**	**1,41%**	**1,41%**	**1,38%**	**1,38%**
Y	**0,000000**	0,589300	**0,048100**	**0,046490**	**0,033900**	**0,011050**	0,221100	**0,000991**	**0,000981**	0,210500	**0,023100**
P_valor médio	0,003970	0,291606	0,317430	0,311471	0,290216	0,054231	0,562350	0,021184	0,015036	0,257749	0,032989

Legenda:		
	P_value < 5%	Hipótese nula rejeitada
	AA% genome	Não foi possível realizar o teste com apenas um códon. O dado exibido é a frequência absoluta do codon no genoma.

Tabela 6: Utilização de códon entre *C. pseudotuberculosis*, linhagem 1002 e genomas do mesmo gênero.

Quando essa tabela foi criada, visou-se uma comparação do uso de códon com um organismo eucarioto e com uma bactéria Gram negativa. Na Tabela 6, as duas primeiras colunas de comparação são do genoma de uma levedura (organismo eucarioto inferior) e *Salmonella* (bactéria Gram negativa). O resultado foi o esperado: todos os testes sobre códons de levedura apresentaram resultados significativamente diferentes da *C. pseudotuberculosis*, exibidos pela legenda em fundo na cor amarela e texto em negrito. Para *Salmonella*, a diferença foi um pouco menor. Sendo assim, *C. pseudotuberculosis* foi comparada com bactérias de três grupos: não patogênicas, patogênicas e oportunistas. Estes são alguns dos resultados:

1) As três espécies de *C. glutamicum* apresentam poucas diferenças na utilização de códons em relação a *C. pseudotuberculosis*;

2) O uso de códons de *C. diphtheriae* é semelhante a *C. pseudotuberculosis*, de modo que todos os testes apresentaram probabilidade maior do que 5%. Portanto, não se pode rejeitar a hipótese nula de igualdade de utilização de códons;

3) Como *C. diphtheriae* é patogênico, bem como *C. jeikeium* e *C. urealyticum*, pode-se concluir que não há relação entre patogenicidade e utilização de códons, ao passo que o uso de códons de *C. jeikeium* e *C. urealyticum* são totalmente diferentes de *C. pseudotuberculosis*, o oposto do *C. diphtheriae*, que é totalmente igual a *C. pseudotuberculosis* e todos são patogênicos.

4) Foi mencionado anteriormente que o uso de códons entre bactérias do gênero *Corynebacterium* eram diferentes. Estes dados confirmam que há realmente muita diferença entre alguns grupos de *Corynebacterium*. É possível separar estes organismos em dois grupos quando o critério é o uso de códons em comparação com *C. pseudotuberculosis*, a saber:

Primeiro Grupo) *C. glutamicum B., C. glutamicum K., C. glutamicum R, C. Kroppenstedtii, C. diphtheriae* e *C. pseudotuberculosis;*

Segundo Grupo) *C. aurimucosum, C. urealyticum, C. jeikeium* e *C. efficiens.*

Esses resultados reforçam as evidências de ancestralidade entre *C. pseudotuberculosis* e *C. diphtheriae*, bem como reforça a ancestralidade entre ambas em relação às linhagens de *C. glutamicum*. Também reforçam evidências de ancestralidade entre as três espécies de *C. glutamicum* e *C. Kroppenstedtii*.

6.2.3 Segunda Revolução Genômica: Utilização de Sequenciadores de Próxima Geração

O texto a seguir, publicado na revista Microbiologia *in foco* no ano de 2009, discorre sobre sequenciadores de próxima geração (Next-Generation Sequencing ou NGS), com ênfase na plataforma SOLiD (Supported Oligonucleotide Ligation and Detection), utilizada pelo nosso grupo de pesquisa, lançada pela Applied Biosystems no ano de 2007. São explicados os avanços do NGS em relação à tecnologia Sanger e analisadas as diferenças entre três plataformas NGS: 454, Illumina e SOLiD. Também são exploradas características de algoritmos desenvolvidos para lidar com a grande quantidade de dados gerada por NGS. Por fim, são relatados resultados preliminares do nosso grupo de pesquisa, ainda no ano de 2009, em relação ao sequenciamento de duas linhagens de *C. pseudotuberculosis*, linhagens 162 e 258.

SEGUNDA REVOLUÇÃO GENÔMICA: UTILIZAÇÃO DE SEQUENCIADORES DE NOVA GERAÇÃO

[redacted] 1 [redacted] 2 [redacted] 3 [redacted] 4 [redacted] 5 [redacted] 6

[redacted] 7 [redacted] 8 [redacted] 9 [redacted] 10 [redacted] 11 [redacted] 12

1. Jeronimo Conceição Ruiz
2. Anderson Rodrigues dos Santos
3. Anne Cybelle Pinto
4. Daniela de Melo Resende
5. Louise Teixeira Cerdeira
6. Rommel Thiago Jucá Ramos
7. Sara Orellana Cuadros
8. Sintia Silva de Almeida
9. Siomar de Castro Soares
10. Vivian D'Afonseca
11. Vasco Azevedo
12. Artur Silva

Rede Genoma de Minas Gerais; Rede Paraense de Genômica e Proteômica

SEQÜENCIAMENTO GENÔMICO

É evidente o impacto das tecnologias de seqüenciamento genômico no modo com que a pesquisa biológica moderna vem sendo desenvolvida. Pela utilização dessas abordagens de estudo, áreas como a de biologia molecular, genética e microbiologia tem produzido grandes descobertas e ajudado nosso entendimento sobre os princípios da vida na terra.

Desde a introdução do método didesoxi por Frederick Sanger em 1977 o genoma de mais de 914 bactérias e 118 eucariotos foram seqüenciados (www.genomesonline.org). A técnica, que teve um início com a modesta capacidade de gerar cerca de mil pares de bases em aproximadamente um ano passou por um processo de automatização nas décadas de 80 e 90, tornando-se semi-automático e posteriormente automático, respectivamente. Em decorrência, trilhões de pares de bases estão atualmente depositados em bancos de dados de domínio público e a nossa perspectiva de entendimento da complexidade dos seres vivos e de nos mesmos nunca mais foi a mesma.

Hoje temos sete genomas humanos completamente seqüenciados, são eles: o Dr. J. Craig Venter, o pioneiro na decodificação e montagem do DNA humano utilizando a estratégia de *shotgun*, o do Dr. James D. Watson, o co-descobridor da fita dupla de DNA, o de dois coreanos, um chinês, um yoruban e o de uma vítima de leucemia. As bases moleculares associadas aos estados de saúde e doença e a identificação de mais de 19000 patologias vinculadas a alterações genéticas só puderam ser desvendadas graças a essas tecnologias (http://www.ncbi.nlm.nih.gov/omim).

Atrelado aos avanços técnicos, computacionais e da bioinformática que catalisaram sobremaneira tais desenvolvimentos vemos uma rápida queda do custo associado à decodificação dos genomas. Estima-se que o custo do primeiro genoma humano seqüenciado em 2003 tenha sido de algo em torno de 500 milhões de dólares, enquanto os genomas recentemente seqüenciados custaram aproximadamente 250 mil dólares, o que trouxe o custo, em proporção ao genoma, a dois milésimo do valor gasto inicialmente e estima-se que em três anos tenhamos um custo estimado em mil dólares por genoma.

A SEGUNDA REVOLUÇÃO GENÔMICA

A segunda revolução genômica aconteceu nos últimos cinco anos, com a entrada no mercado das plataformas de seqüenciamento genômico que utilizam tecnologia de próxima geração (*Next-Generation Sequencing, ou NGS).*

Rapidez, melhor rendimento, maior cobertura e menor custo são alguns dos atributos dessas novas ferramentas. Baseadas em nanotecnologias inovadoras e criativas, a plataforma 454 (FLX/Roche), a SOLEXA (Illumina) e o SOLiD (Applied Biosystems) abriram uma nova fronteira para a biociência. Pesquisadores em seus laboratórios e pequenos grupos de pesquisa têm hoje em mãos a mesma capacidade de seqüenciamento que há algum tempo somente era possível para um grande centro de seqüenciamento genômico (Schendure e Ji, 2008).

O pirosequenciador 454 FLX da Roche foi o primeiro seqüenciador de nova geração a entrar no mercado, em 2004. Compartilha com o método didesoxi o fato de realizar o seqüenciamento pela síntese. Porém, para detecção de nucleotídeos e posterior caracterização da sequência de DNA-alvo, utiliza bases nitrogenadas marcadas com pirofosfato, que na presença de luciferase e outros substratos é convertido em luz visível detectado por uma câmera. Essa inovação permitiu a realização de seqüenciamento massivo de milhões de fragmentos de DNA simultaneamente (Mardis, 2008). O rendimento do 454 alcança aproximadamente 100 milhões de bases em uma única corrida de 7,5 horas de duração, e a extensão dos fragmentos gerados varia de 250 a 350 bases, em média, com um custo estimado de 10 mil dólares por corrida. O lançamento do kit de seqüenciamento Titanium, pela Roche, promete uma melhora no rendimento que agora pode chegar a 500 Mb (Megabases) e elevar a extensão média das leituras para 400 bases.

O Solexa 1G Genetic Analyzer, da Illumina, disponibilizado para o mercado no segundo semestre de 2006, também tem como princípio o seqüenciamento pela síntese (Mardis, 2008).

A plataforma SOLiD (*Supported Oligonucleotide Ligation and Detection*), da Applied Biosystems, lançada em 2007, difere das demais por realizar a leitura da seqüência de nucleotídeos de uma fita de DNA durante a reação de incorporação de dinucleotídeos marcados, catalisada pela DNA ligase. A incorporação de um dinucleotídeo é seguida da excitação do fluoróforo, da leitura do sinal e da remoção do fluoróforo antes da incorporação do dinucleotídeo seguinte. O resultado da leitura dos sinais fluorescentes gera um código de cores que é analisado na forma de uma matriz de cores para então ser transformado no tradicional código de letras (Mardis, 2008). A versão 2 do SOLiD é capaz de gerar de 3 a 5 Gb (GigaBase) e leituras de cerca de 35 bases de extensão, a um custo aproximado de 2 mil dólares por corrida.

O lançamento da versão 3 do SOLiD permitiu elevar ainda mais o desempenho do seqüenciamento, tanto no sentido de reduzir o tempo de corrida, de 8 para 6 dias, mas também no sentido de gerar leituras maiores (~70 bases) e com um rendimento que chega a 20 Gb. A versão 3 Plus, que será lançada pela Applied Biosystems até o final deste ano, elevará esse rendimento para 100 Gb. Uma aplicação revolucionária do SOLiD é a análise de transcriptoma, com uma eficiência comparável à dos métodos tradicionalmente utilizados, como PCR em tempo real (RT-PCR), atrelada a facilidade de análise de resultados.

Vários laboratórios em diferentes estados do Brasil já estão implementando e fazendo bom uso das tecnologias NGS. A comunidade científica nacional e internacional, ao mesmo tempo em que acompanha atentamente a publicação dos resultados produzidos por esses equipamentos, já aguarda o lançamento das nanotecnologias de seqüenciamento baseadas em análise *single-molecule*, que estão neste momento em desenvolvimento, e já têm sido chamadas pela comunidade científica de "*next-next-generation sequencing"* ou de seqüenciamento de terceira geração.

OS DESAFIOS ASSOCIADOS À MONTAGEM DE GENOMAS UTILIZANDO AS ESTRATÉGIAS DE SEQUENCIAMENTO DE NOVA GERAÇÃO

Por mais de 30 anos o seqüenciamento genômico foi realizado utilizando a tecnologia de Sanger com equipamentos que geram leituras relativamente grandes (de 600 a 1500 pares de bases rotineiramente) de bibliotecas do tipo *mate-pair* construídas com uma grande variedade de tamanhos de insertos (plasmidiais de 2 a 10kb, de cosmídeos e fosmídeos de 35 a 40kb, de BACs de 50 a 150Kb e de YACs variando de 150 a 3000Kb).

Como citado anteriormente, a série de estratégias de seqüenciamento genômico que surgiram nos últimos anos tem em comum o elevado grau de paralelismo do processo de seqüenciamento, capaz de produzir dados em uma escala de magnitude de seqüenciamento *highthroughput* nunca antes alcançada. Apesar disso, esse desempenho tem um custo: o tamanho da leitura gerada pelo seqüenciador. Enquanto que os aparelhos 454 geram em média 400pb o SOLiD e Helicos geram de 25 a 50pb por leitura.

Além do viés associado ao tamanho das leituras geradas, algumas dessas novas plataformas de seqüenciamento genômico produzem erros ou tem limitações inerentes à tecnologia empregada. Os aparelhos 454, por exemplo, tem dificuldade de seqüenciar longos trechos homopoliméricos e "escondem" esses trechos no genoma seqüenciado, enquanto que os aparelhos Helicos da atualidade seqüenciam cada fragmento de DNA mais de uma vez, produzindo leituras duplicadas com o objetivo de reduzir os índices de erro.

Quando se discute a montagem de genomas, é importante que se faça uma distinção entre algumas abordagens amplamente empregadas. As abordagens denominadas de novo têm como objetivo a montagem de genomas sem a utilização de qualquer outra informação além das leituras geradas pelo processo de seqüenciamento genômico enquanto que as abordagens denominadas *re-sequencing* ou *reference assembly* utilizam um genoma altamente relacionado, filogeneticamente próximo e anotado, durante o processo de montagem. Essa última estratégia já foi utilizada para a montagem do genoma de diversas linhagens de organismos complexos, como *Drosophila melanogaste*r e *Caenorhabditis elegans*, e está sendo usada em larga escala para a montagem de genomas de cânceres humanos (Pop & Salsberg, 2008).

Apesar das vantagens e facilidade em se utilizar um genoma de referência, essa abordagem é obviamente limitada não permitindo a montagem de genomas que não têm um organismo próximo já seqüenciado. Por outro lado, a montagem de novo continua sendo uma etapa fundamental, complexa, hardware-dependente e carente de novas metodologias analíticas, sendo esta na atualidade uma frente efervescente de estudo e desenvolvimento em bioinformática.

As limitações das abordagens de montagem de novo também se associam diretamente às limitações tecnológicas relacionadas às características dos dados gerados pelos sequenciadores de nova geração. Estudos desenvolvidos por Chaisson et al. (2004) e Whiteford et al. (2005) mostraram uma rápida deterioração na qualidade das montagens quando o tamanho das leituras diminui. Chaisson *et al.* (2004) mostraram que, para leituras de 750pb, obtidas com seqüenciamento utilizando tecnologia Sanger, a montagem do genoma de *Neisseria meningitidis* resultou em 59 contigs, dos quais 48 eram maiores que 1kpb. Por outro lado, com leituras de aproximadamente 70pb, a montagem gerou mais de 1.800 contigs, dos quais apenas um sexto eram maiores que 1kpb. Até para leituras relativamente longas (200pb), a montagem resultante foi muito fragmentada (296 contigs). Resultados semelhantes foram obtidos por Whiteford et al.(2005), que observaram uma rápida diminuição no tamanho dos contigs com leituras menores que 50pb.

Em linhas gerais, os algoritmos originalmente desenvolvidos para montagem utilizando leituras geradas pela tecnologia Sanger não podem ser diretamente aplicados sem alguma alteração para os dados gerados pelas tecnologias NGS. O tamanho das leituras e o volume de dados gerados têm um efeito logarítmico no tempo de processamento, às vezes inviabilizando a montagem e caindo na classe de problemas *NP hard* de difícil resolução. Como exemplo, oito vezes a cobertura de um genoma de mamífero de 3Gb em tamanho requerem 30 milhões de leituras do Sanger enquanto que seriam necessárias 750 milhões de leituras Illumina (Pop & Salsberg, 2008).

MODELOS COMPUTACIONAIS UTILIZADOS NA NOVA ERA GENÔMICA

Os softwares existentes hoje para o processo de alinhamento, análise e montagem de genomas utilizam diferentes abordagens computacionais, o que resulta em uma grande variedade de soluções, e dentre as principais estratégias temos: a) Overlap-layout-consensus (OLC); b) Greedy e c) Eurelian path. Independente da estratégia empregada pelo algoritmo, um objetivo comum a todos é a redução de tempo do processo de alinhamento sem prejuízo da acurácia na análise e montagem final, e a obtenção de uma seqüência consenso com boa qualidade.

Dentre os principais formatos de entrada desses programas temos o CSFASTA (color space fasta) que apresenta o código de cores produzido pelo processo de seqüenciamento SOLiD, e o tradicional FASTA.

Seja na montagem *de novo* ou *reference assembly* uma das metodologias computacionais mais utilizadas é a transformada BWT - *Burrows-Wheeler Transform* (Burrows and Wheeler, 1994). Este software foi utilizado inicialmente para compressão de dados e hoje temos seu emprego em vários algoritmos como exemplo o software MAQ (Li et al., 2008a) que possui uma ferramenta baseada no BWT chamada BWA - *Burrows-Wheeler Alignment tool* (Li and Durbin, 2009) e também o software SOAP2 (http://soap.genomics.org.cn/), que devido o uso do BWT teve seu processo de alinhamento das leituras mais rápido, reduzindo drasticamente o uso de memória RAM. Além disso, o SOAP2 permite *a* identificação de regiões de dissimilaridade e detecção de alterações nucleotídicas utilizando o teorema de bayes para realização de tais inferências.

O algoritmo de Rabin-karp também tem sido utilizado com o objetivo de aumentar a velocidade do processo de alinhamento das leituras com o genoma de referência, e tem sido usado no software SOCS **-** *Short Oligonucleotide Color Space (*http://socs.biology.gatech.edu/*)* que permite inclusive que o usuário informe a quantidade de regiões de dissimilaridade através de um parâmetro específico. Além destes, utilizando outra estratégia o software SHRiMP (http://compbio.cs.toronto.edu/shrimp) realiza alinhamento das leituras com a referência usando o algoritmo de Smith-Waterman que realiza um alinhamento rigoroso e veloz, contudo, não gera a seqüência consenso. O MOM - *Maximum Oligonucleotide Mapping* (Eaves and Gao, 2009) possui uma estratégia similar, mas possui maior sensibilidade na identificação de regiões de alta similaridade e um melhor percentual de mapeamento de uma leitura em uma única região quando comparado aos programas SOAP, MAQ e SHRiMP.

Além das iniciativas acadêmicas voltadas ao desenvolvimento de algoritmos específicos para montagem genômica temos aqueles desenvolvidos por empresas privadas como é o caso do Corona Lite (http://solidsoftwaretools.com/gf/project/corona/) desenvolvido pela Applied Byosistem e que hoje representa uma das principais ferramentas aplicadas nos pipelines de montagem aplicados pela empresa.

Por outro lado, nas montagens sem genoma de referência um dos algoritmos mais empregados tem sido o Velvet *(*http://www.ebi.ac.uk/~zerbino/velvet/*)*. Esse algoritmo utiliza o caminho Eureliano (teoria

TABELA 1. PANORAMA GERAL DOS DADOS BRUTOS DO SEQUENCIAMENTO DOS GENOMAS DAS LINHAGENS CpCamelo **E**CpCavalo **GERADOS NO** SOLiD

DADOS	CpCamelo	CpCavalo
Número de leituras	21.102.241	31.294.379
Tamanho das leituras (pb)	35	35
Número de pares de bases	738.578.435	10.953.032.650
Cobertura média (X)	150-200	200
Detecção de SNP	18.000	19.000

TABELA 2. VISÃO GERAL DOS GENOMAS DE DIFERENTES LINHAGENS DE *C. pseudotuberculosis* **SEQUENCIADAS NO** SOLiD

DADOS	CpCamelo	CpCavalo
Tamanho do genoma (pb)	2.273.983	2.273.983
Número de genes	2.230	2.224
Tamanho médio dos genes (pb)	876	880
Densidade gênica	0,98	0,97
Conteúdo GC (%)	52	52
% do genoma codificante	85,90	86,10

dos grafos) para encontrar um único caminho que percorra, todas as leituras, ou pelo menos a grande maioria e que seja capaz de gerar a seqüência consenso e isso é feito através da identificação das sobreposições que ocorrem no gráfico de diBruijn gerado o que demanda relativamente pouco tempo de processamento computacional.

UTILIZAÇÃO DE BANCOS DE DADOS RELACIONAIS PARA O ARMAZENAMENTO DE DADOS ORIUNDOS DE SNG

Um aspecto importante que deve ser comentado no processo de montagem genômica diz respeito à utilização de um Sistema Gerenciador de Banco de Dados (SGBD) relacional. Dentre as principais vantagens do uso de um banco de dados relacional podemos citar a centralização de dados em um servidor com controle de acessos, a garantia da manutenção da integridade de dados e a facilidade da extração de dados.

Outro aspecto positivo está relacionado à possibilidade de elaboração fácil de rotinas de exportação de dados nos formatos EMBL e GFF, amplamente utilizados e adotados por ferramentas de anotação genômica como o Artemis (http://www.sanger.ac.uk/Software/Artemis/) e o Apollo (http://www.dhgp.org/), além do GBROWSE que utiliza ambiente web para apresentação do genoma anotado.

UTILIZANDO A PLATAFORMA SOLID DE SEQUENCIMENTO

Pioneiros no Brasil e na América Latina na utilização das tecnologias NGS, o esforço integrado de vários pesquisadores da Universidade Federal do Pará (Dr. Artur Silva), da Universidade Federal de Minas Gerais (Dr. Vasco Azevedo) e da Fundação Oswaldo Cruz de Minas Gerais (Dr. Jeronimo Ruiz e Dr. Guilherme Oliveira) foi capaz de gerar o seqüenciamento, montagem e anotação da primeira bactéria não *E. coli*.

A Rede Paraense de Genômica e Proteômica juntamente com as instituições supracitadas deram início aos seus trabalhos em parceria, seqüenciando diferentes linhagens do patógeno *Corynebacterium pseudotuberculosis*, agente etiológico da doença denominada Linfadenite Caseosa (LC).

A montagem de genomas tendo como fonte exclusiva de dados fragmentos pequenos representa uma área efervescente de pesquisa na área de bioinformática. Nesse contexto e não existindo algoritmos eficientes para montagem de novo de genomas, o principio de uma parceria com a Applied Biosystems foi estabelecido visando, entre outros objetivos, a integração de várias rotinas de pré-filtragem, e de montagem integrando estratégias de montagem *de novo* e *reference assembly* desenvolvidas pelo grupo Paraense-Mineiro e empregadas com sucesso na montagem de genomas bacterianos, no pipeline atualmente disponibilizado pela empresa.

Para tal, a parceria vem acompanhando o desenvolvimento de algoritmos específicos principalmente no que diz respeito à montagem *ab initio* dos dados oriundos do SOLiD, ou seja, sem um genoma de referência, para dar continuidade aos trabalhos vinculados ao uso do SOLiD. Um problema inicialmente enfrentado foi que a principal abordagem utilizada pela vasta maioria dos programas para montagem de genomas com a utilização de pequenas leituras é baseado em um genoma de referência. A dificuldade que esses programas de alinhamento e montagem mostram é que na maioria das vezes regiões de não-similaridade entre o genoma de referência e o genoma a ser montado são descartadas.

GENOMA DE *Corynebacterium Pseudotuberculosis*

As duas diferentes linhagens de *C. pseudotuberculosis* seqüenciadas pela estratégia SOLiD foram isoladas de diferentes hospedeiros, sendo um camelídeo e um equino, respectivamente. A biblioteca utilizada na geração dos dados genômicos foi de fragmentos e os dados brutos gerados no SOLiD estão ilustrados na tabela 1.

A massiva geração de dados pelas técnicas apresentadas tem viabilizado o crescimento dos mais vastos campos da ciência tais como biologia geral, medicina, veterinária e biotecnologia. Aliado a esse notório desenvolvimento, atualmente as mais diversas áreas tem atrelado seus projetos, tornando os diversos campos de pesquisas cada vez mais interdisciplinares, e com isso, permitido o avanço da pesquisa como um todo no país.

REFERÊNCIAS BIBLIOGRÁFICAS

Burrows M, Wheeler DJ. A block-sorting lossless data compression algorithm. Technical report. 2004 124, Palo Alto, CA, Digital Equipment Corporation.

Chaisson M et al. Fragment assembly with short reads. Bioinformatics. 2004 20:2067-2074.

Eaves HL, Gao Y.MOM: maximum oligonucleotide mapping. Bioinformatics. 2009 25:969–970.

Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009 15;25(14):1754-60

Li H et al. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008 18:1851–1858.

Mardis, ER. The impact of next generation sequencing technology on genetics. Trends Genet. 2008 24:133-141.

Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trend Genet. 2008 24:142-149.

Schendure J, Ji H. Nat Biotechnol. 2008 26:1135-1145.

Whiteford N et al. An analysis of the feasibility of short read sequencing. Nucleic Acids Res. 2005 33, e171.

6.2.4 Plasticidade Genômica e Evolução Bacteriana

O conhecimento sobre fatores que levam à modificação do DNA bacteriano, propriedade conhecida como plasticidade genômica, é útil tanto para análises sobre a correção de um genoma que está sendo montado, quanto para a pesquisa de elementos genéticos que podem causar, por exemplo, o aumento da virulência de uma bactéria. Como exemplos de plasticidade em genomas bacterianos podem ser citados a ausência ou presença de grandes regiões codificadoras, quando comparadas com espécies evolutivamente próximas, e a presença significativa de proteínas hipotéticas, possivelmente oriundas de transferências horizontais de bactérias ambientais, as quais no máximo 1% são cultiváveis e, portanto, passíveis de terem seus genomas sequenciados pelas tecnologias atualmente conhecidas.

O texto a seguir, publicado na revista Microbiologia *in foco* no ano de 2011, lista e conceitua alguns dos principais fatores responsáveis ou característicos da plasticidade genômica bacteriana, bem como analisa o conjunto de programas disponível para identificação desses fatores, para então propor um novo programa de computador para predição de ilhas genômicas relacionadas com a virulência bacteriana denominado PIPS (*Pathogenecity Island Prediction Software*), o qual, posteriormente, foi publicado na revista *PlosOne*.

PLASTICIDADE GENÔMICA E EVOLUÇÃO BACTERIANA

Siomar de Castro Soares, Artur Luiz da Costa Silva, Rommel Thiago Jucá Ramos, Louise Teixeira Cerdeira, Amjad Ali, Anderson Rodrigues dos Santos, Anne Cybelle Pinto, Aryane Aparecida Magalhães Cassiano Rocha, Eudes Guilherme Vieira Barbosa, Laiane Alves Leão, Sintia Silva de Almeida, Vinícius Augusto Carvalho de Abreu, Anderson Miyoshi, Vasco Azevedo

A reemergência de doenças causadas por bactérias patogênicas e o aparecimento de linhagens resistentes a antibióticos têm chamado a atenção da comunidade científica a patógenos que, até então, estavam sob controle. Este fenômeno se deve, principalmente, às pressões geradas dentro do hospedeiro que contribuem para o direcionamento evolutivo de bactérias que passaram por modificações genômicas. Esta propriedade dinâmica do DNA de se modificar é chamada de plasticidade genômica e advem de diversos mecanismos, tais como mutações pontuais, conversões gênicas, rearranjos (inversão ou translocação), deleções e inserções de DNA de outros organismos (plasmídeos, bacteriófagos, transposons, elementos de inserção e ilhas genômicas) (Schmidt & Hensel, 2004).

O presente artigo explora os conceitos dos mecanismos citados acima, que podem levar à plasticidade genômica, e como estes podem influenciar no estilo de vida bacteriano. Além disso, serão abordados os padrões gerados nos genomas bacterianos em decorrência de eventos de transferência horizontal de genes, sejam eles: desvio na assinatura genômica; presença de transposase, tRNAs flanqueadores e seqüências de inserção; e grande concentração de proteínas hipotéticas. Por fim, serão descritas ferramentas de bioinformática voltadas para a análise *in silico* de características padrões para identificação de elementos móveis dando maior enfoque nas ilhas de patogenicidade, que é a classe de ilhas genômicas mais estudada nos dias atuais.

PLASMÍDEOS

Os plasmídeos contribuem para a plasticidade genômica através de sua capacidade de auto-transferência, de mobilizar outros plasmídeos co-residentes, além da possibilidade de integração cromossômica. No DNA plasmidial podem ser encontrados genes de resistência a antibióticos e determinantes associados à patogenicidade (Dobrindt & Hacker, 2001). Existem vários exemplos de plasmídeos que carregam genes de virulência e que provêm vantagens adaptativas às respectivas bactérias aceptoras, como os plasmídeos: ColV, que confere à *Shigella kentucky* maior facilidade de colonização e adaptabilidade a diversos hospedeiros (Johnson *et al.*, 2010); pCD1, pFra, e pPCP1, responsáveis pela evolução e, em grande parte, pela virulência das diferentes linhagens de *Yersinia pestis*(Rajanna *et al.*, 2010); e pO26-Vir e pO26-L, que contêm genes de virulência e de resistência a antibióticos importantes para a patogênese de *Escherichia coli* produtora de toxina Shiga (STEC) (Fratamico *et al.*, 2011).

Em casos extremos, um único plasmídeo pode ser responsável pela emergência de uma linhagem patogênica de determinada espécie. Um exemplo que ilustra esse fato é a bactéria *Rhodococcus equi*, pertencente à família das Actinobactérias, que possui um plasmídeo de virulência que se encontra ausente em linhagens avirulentas. Este possui uma ilha de patogenicidade (Pathogenicity Island, PAI) onde estão localizados genes que codificam proteínas de superfície associadas à virulência (genes *vap*) (Takai *et al.*, 2000).

BACTERIÓFAGOS

Os bacteriófagos podem afetar a plasticidade genômica de um organismo através do mecanismo de transdução. Através deste método, fagos funcionais injetam o DNA de uma bactéria em outra sem causar dano à aceptora. Esse DNA exógeno pode incorporar-se no genoma do microrganismo aceptor e lhe proporcionar vantagens adaptativas. Os profagos conferem proteção contra infecção lítica e carregam genes de toxicidade que podem ser adquiridos pelo organismo aceptor. Os genes de toxicidade, juntamente com outros fatores de virulência, possuem funções que envolvem a neutralização dos sistemas de defesa e subversão ou destruição das células do organismo hospedeiro, contribuindo, assim, para o espalhamento e sobrevivência da bactéria no organismo. *Clostridium botulinum* (botulismo), *Streptococcus pyogenes* (febre escarlate, artrite reumatóide e glomerulonefrite pós-estreptocócica), *Staphylococcus aureus* (endocardite e impetigo), *Escherichia coli* (enteropatias) e *Corynebacterium diphtheriae* são exemplos de bactérias que adquiriram genes de toxicidade por esse mecanismo (Brüssow *et al.*, 2004).

ILHAS GENÔMICAS

As ilhas genômicas (Genomic Islands, GEIs), por sua vez, afetam a plasticidade através de sua capacidade de transferência e incorporação de grande número de genes em bloco (operons e grupos de genes codificantes de funções correlatas) que podem causar mudanças drásticas, levando a saltos evolutivos em relação à sua linhagem parental. GEIs são caracterizadas por serem regiões de DNA adquiridas de outros organismos que apresentam tamanhos variáveis entre 10 e 200 kilobases; possuírem seqüências derivadas de fago e/ou plasmídeo incluindo genes de transferência ou integrase, além de sequências de inserção (Insertion Sequences, IS); e aparecerem inseridas entre genes de tRNA ou flanqueadas por repetições diretas que parecem estar envolvidas na sua instabilidade (Hacker & Carniel, 2001). Além disso, as GEIs sofrem eventos de deleção com freqüências distintas, são passíveis de transferência e possuem estruturas do tipo mosaico (Schmidt & Hensel, 2004).

Em análise das GEIs de *Burkholderia pseudomallei,* bactéria saprofítica do solo, causadora da melioidose em humanos, observou-se rápida aquisição e/ou perda de genes e mudança na organização das mesmas em diferentes linhagens. A GEI 11 da linhagem modelo foi substituída por ilhas alternativas de tamanhos diferentes nas linhagens 1106a e 1106b e está ausente nas demais linhagens estudadas no trabalho (Tumapa *et al.*, 2008).

Vale ressaltar que as GEIs podem ser classificadas em diversas classes de acordo com o conteúdo gênico das mesmas. As classes existentes são: Ilhas Simbióticas, que podem estar envolvidas na associação de bactérias a plantas hospedeiras da família *Leguminosae*(Barcellos *et al.*, 2007); Ilhas de Resistência, que possuem genes relacionados à resistência a antibióticos (Krizova & Nemec, 2010); Ilhas Metabólicas, onde estão localizados diversos genes associados com a biossíntese de metabólitos secundários (Tumapa *et al.*, 2008); e as Ilhas de Patogenicidade (PAIs), que apresentam uma alta concentração de genes de virulência, aparecem associadas à bactérias patogênicas e estão envolvidas na reemergência de vários patógenos (Dobrindt *et al.*, 2000).

ILHAS DE PATOGENICIDADE

O termo PAI foi utilizado pela primeira vez em 1990 para descrever grandes regiões instáveis dentro do cromossomo de algumas linhagens patogênicas de *E. coli*, *in vitro* (Hacker *et al.*, 1990). A identificação baseou-se na observação da íntima relação entre a deleção de regiões codificadoras de hemolisina e adesinas fimbriais e a geração de linhagens não patogênicas de *E. coli*. A estratégia utilizada envolveu técnicas de clonagem gênica, eletroforese em gel de campo pulsado (PFGE) e hibridização através da técnica de Southern Blot. Com isso, Hacker *et al.* (1990) demonstraram que os genes codificadores de hemolisina e adesinas fimbriais estão posicionados na mesma região cromossômica em linhagens selvagens de *E. coli* e que passam por eventos de deleção tanto *in vivo* quanto *in vitro*(Hacker *et al.*, 1990).

Atualmente, o termo PAI é utilizado para descrever regiões cromossômicas que foram adquiridas pelo organismo por transferência horizontal de genes e que: contenham conteúdo G+C anômalo; estejam ausentes em organismos não patogênicos do mesmo gênero ou espécie correlata; e codifiquem fatores de virulência de bactérias patogênicas (Gal-Mor & Finlay, 2006; Schmidt & Hensel, 2004). Os genes de virulência adquiridos via PAI estão envolvidos nos processos de adesão, invasão, colonização, multiplicação dentro do hospedeiro e evasão do sistema imune, além de permitirem o contato, penetração e sobrevivência de bactérias patogênicas dentro do hospedeiro (Schumann, 2007).

Uma característica geral das PAIs é a presença de genes codificadores de proteínas relacionadas com o transporte de ferro (Brown *et al.*, 2002). A High Pathogenicity Island (HPI) é um exemplo de PAI que possui genes relacionados à aquisição de ferro do meio. Essa PAI é compartilhada por diversas enterobactérias, sejam elas: *Escherichia*, *Klebsiella*, *Enterobacter*, *Citrobacter*, *Salmonella* e *Serratia*. A deleção desses genes de aquisição de ferro está relacionada à perda de virulência de *E. coli* extraintestinal e de espécies de *Yersinia* patogênicas a humanos (Benedek & Schubert, 2007).

Em *C.diphtheriae*, o principal gene presente nas PAIs e responsável pelos efeitos causados pela bactéria, é o que codifica a proteína DT (gene *tox*), uma toxina encontrada em corynefagos lisogênicos β *tox*+, γ *tox*+ e ω *tox*+. Foi relatado que somente as linhagens lisogênicas para este fago produzem a toxina, no entanto, as linhagens atoxigênicas também têm surgido como causa freqüente de processos infecciosos que podem variar de lesões cutâneas e faringites a doenças invasivas severas. Tendo em vista que a vacina contra essa bactéria é baseada no toxóide produzido a partir da inativação da toxina diftérica, a instabilidade das PAIs é um fator determinante na resposta imune de pacientes imunizados, visto que os mesmos apresentam bacteremia e endocardite na ausência de lesões mediadas pela toxina (Mattos-Guaraldi *et al.*, 2000).

Existem diversos outros exemplos relatados da influência das Ilhas de Patogenicidade na virulência de diferentes espécies de bactérias. Evidências sugerem que a transferência lateral de genes em *Vibrio cholerae* atoxigênica foi responsável pela emergência de novas linhagens patogênicas. Foi demonstrado que a PAI de *V. cholerae* (Vibrio Pathogenicity Island, VPI) está associada com linhagens epidêmicas e pandêmicas. Todos os genes da VPI supostamente são importantes para causar a doença e tem um papel direto na patogênese de *V. cholerae* ou um papel indireto na mobilidade ou transferência da VPI, levando potencialmente ao surgimento de novas linhagens virulentas, ou a reemergência da doença (Karaolis *et al.*, 1998).

Por fim, como um último exemplo de influência de PAIs na patogenicidade de bactérias, podemos citar *Helicobacter pylori, cujas linhagens* patogênicas podem ter evoluído de sua contraparte benigna pela incorporação de grande quantidade de informação genética na forma de PAI que teria conferido uma capacidade aumentada de colonizar novos hospedeiros. A ilha de patogenicidade cag (cag-PAI), além das várias seqüências de inserção e regiões de plasticidade presentes no organismo, é a principal responsável pela diversidade das linhagens de *H. pylori*. O rearranjo dentro de cag-PAI é um fenômeno comum e os genes que estão presentes nessa ilha estão sobre pressão seletiva maior que os demais genes do genoma. A presença de cag-PAI intacta está associada a um aumento na severidade da doença causada por *H. pilory* indicando seu papel de virulência (Kauser *et al.*, 2004).

"BLACK HOLES"

Assim como as inserções de DNA, a deleção de genes também apresenta grande importância para a adaptabilidade de muitos organismos a novos ambientes e hospedeiros. O conceito de "Black Hole" foi criado para caracterizar eventos de deleção de fatores de antivirulência, ou seja, genes cuja expressão em organismo patogênico é incompatível com a virulência deste patógeno. A evolução a partir da deleção de genes de antivirulência parte da premissa de que os genes requeridos para a adaptação do organismo em determinado nicho podem inibir a adaptabilidade do mesmo em outro nicho (Maurelli, 2007).

Em *E. coli*, por exemplo, a perda de *cad*A, gene codificador de lisina descarboxilase (LDC), e *omp*T, que sintetiza uma protease de membrana externalizada, tornam a bactéria virulenta (Maurelli *et al.,* 1998; Suzuki & Sasakawa, 2001). O mecanismo de ação da cadaverina, produto da descarboxilação de lisina pela LDC, ainda é desconhecido, mas existem duas hipóteses: cadaverina inativa a enterotoxina sintetizada por *E. coli*; ou, cadaverina age diretamente na célula alvo para protegê-la. Maurelli *et al.* (1998) demonstraram que células de mucosa de coelho pré-tratadas com cadaverina e lavadas foram protegidas dos efeitos da enterotoxina. Já a ausência da proteína *Omp*-T em espécies do gênero *Shigella* e em linhagens de *E. coli* enteroinvasiva é crucial para a manutenção da proteína *Vir*G na superfície da célula que, por conseguinte, é um pré-*requi*sito para que essas bactérias se movimentem em células de mamíferos, incluindo disseminação bacteriana pelas células epiteliais (Suzuki & Sasakawa, 2001).

IDENTIFICAÇÃO DE REGIÕES ADQUIRIDAS POR TRANSFERÊNCIA HORIZONTAL DE GENES (HGT)

Transferência horizontal de genes é um processo comum que ocorre quando o DNA de um organismo é transferido para outro, podendo, ou não, se tornar estável e incorporado no aceptor. A aquisição e perda de genes por todos os mecanismos referenciados anteriormente (plasmídeos, bacteriófagos, etc) refletem no estilo de vida e na versatilidade fisiológica do microrganismo (Dobrindt & Hacker, 2001) e, portanto, revelam um grande potencial para novas descobertas. O interesse pelo potencial latente dessa área, associado ao número crescente de seqüências genômicas completas disponíveis para análises, tem dirigido esforços de diversos pesquisadores para a implementação de ferramentas *in silico* que têm como intuito principal, a identificação de eventos de transferência horizontal. Para trabalhar com tais ferramentas, primeiramente, é importante entender como esses eventos de transferência horizontal modificam o genoma do organismo aceptor e quais padrões gerados por esses eventos podem ser utilizados em análises *in silico* tais como: desvios na assinatura genômica, i.e., conteúdo G+C anômalo e desvio de uso de códon; presença de transposase, tRNAs flanqueadores e sequências de inserção; e grande concentração de proteínas hipotéticas.

ASSINATURA GENÔMICA

A identificação de regiões adquiridas por transferência horizontal baseia-se na observação do padrão de conteúdo G+C e uso de códon, visto que organismos diferentes apresentam padrões específicos que constituem sua assinatura genômica. Grupos de genes adquiridos por transferência horizontal apresentam um desvio em relação a esse padrão, pois refletem a assinatura do genoma de origem (Langille *et al.*, 2008). Contudo, devido a diversos fatores como a força de ligação códon/anticódon e maior disponibilidade de determinado gene de tRNA, a pressão evolutiva direciona os genes a adaptarem seu uso de códon ao do genoma onde se encontram inseridos, de modo a aumentar a expressão dos mesmos (Karlin *et al.*, 1998).

Além disso, a preferência de códon em bactérias encontra-se intimamente relacionada à composição de bases inclusive de regiões intergênicas. A adoção de códons preferenciais ricos em GC ou AT leva a um padrão de conteúdo G+C similar entre os genes distribuídos pelo genoma (Hershberg & Petrov, 2009). Levando-se em conta a grande densidade de regiões codificantes no genoma de procariotos, a adaptação do uso de códon direciona a uma distribuição homogênea da composição de bases nesses organismos.

Atualmente, existem diversas ferramentas criadas com o intuito de identificar regiões adquiridas por HGT através de desvio na assinatura genômica, isto é, desvio do conteúdo G+C (Wavelet analysis of G+C content, Cumulative GC Profile, δ_P-web, IVOM, IslandPath and PAI-IDA) e do Uso de Códon (SIGI-HMM and PAI-IDA) (Gao & Chen, 2010). Contudo, devido às adaptações no uso de códon e conteúdo G+C referenciadas acima, a identificação de regiões móveis baseada na assinatura genômica só é possível para regiões que foram adquiridas recentemente de organismos distantes filogeneticamente, isto é, que possuam assinatura genômica discrepante em relação ao genoma aceptor.

INFLUÊNCIA DAS TRANSPOSASES, TRNAS E SEQÜÊNCIAS DE INSERÇÃO NA MOBILIDADE GENÔMICA

Regiões adquiridas por inserção mediada por transposase apresentam regiões flanqueadoras de repetição invertida (Inverted Repeat, IR), que freqüentemente estão contidas em seqüências codificadoras de tRNA (Hou, 1999). Essas seqüências de repetição invertida (palíndromos) são alvos de transposases que medeiam a integração e deleção de elementos de inserção no genoma do hospedeiro e são conhecidas como seqüências de inserção (Insertion Sequences, IS) (Tobes & Pareja, 2006). Genes de tRNA são conhecidos como "hot spots" para elementos de inserção por possuírem uma seqüência 3' terminal reconhecida por diversas integrases e que aparecem com grande freqüência em genes de tRNA de selC e leuX (selenocisteína e leucina, respectivamente) (Hou, 1999; Ou *et al*., 2006).

Blum et al (1994) demonstraram que as ilhas de patogenicidade PAI I e PAI II de *E. coli* linhagem 536 estão inseridas em regiões flanqueadas por genes de tRNA de selC e leuX, respectivamente. Essas PAIs passam por eventos de deleção, tanto *in vivo* quanto *in vitro*, levando à inibição do fenótipo hemolítico e a uma maior susceptibilidade de infecção em pacientes diabéticos por esse organismo (Hacker *et al*., 1990). Além disso, a quebra do gene de tRNA de selC durante o evento de deleção da PAI I leva a uma regulação negativa de genes que possuem códon para selenocisteína, como os genes das formato desidrogenases (fdh), envolvidas no metabolismo de energia, fixação de carbono e homeostase do pH (Blum *et al*., 1994). As FDHs são as selenoproteínas mais distribuídas na natureza, e foi sugerido que os genes codificantes de FDH e da maquinaria de incorporação e síntese de selenocisteína sofreram vários processos de transferência horizontal (Stock & Rother, 2009).

Lesic *et al.* (2004) analisaram a transferência horizontal da High Pathogenicity Island (HPI) entre diferentes linhagens de *Yersinia pseudotuberculosis* e desta para *Y. pestis*, e observaram que HPI quase sempre se inseriu em um gene de tRNA de *asn*3 e que a transferência mostrou-se Rec-A dependente nessas linhagens.

PROTEÍNAS HIPOTÉTICAS

A grande concentração de proteínas hipotéticas em regiões adquiridas por transferência horizontal é uma nova característica que revela um grande potencial na análise de novas proteínas de uso industrial. Hsiao *et al.* (2005) demonstraram que a concentração de genes de proteínas hipotéticas, isto é, que não apresentam função conhecida, é maior dentro de GEIs em comparação com a média genômica. Nesse mesmo trabalho, foi demonstrado que a alta concentração de genes de proteínas hipotéticas não está relacionada a uma baixa acurácia na predição gênica, mas possivelmente, a eventos de aquisição gênica de organismos não seqüenciados e estudados, incluindo nesse grupo bactérias não cultiváveis.

Neste contexto, é importante ressaltar que vários produtos naturais de grande valor econômico, tais como antibióticos e outros fármacos, são derivados de micro-organismos cultivados de amostras ambientais e a grande densidade de microorganismos, que pode atingir 500.000 espécies bacterianas em 30g de solo, gera um grande potencial para descoberta de novos genes via análises metagenômicas. Contudo, apenas 0,1% a 1% das bactérias presentes em amostras ambientais são cultiváveis (Daniel, 2004). Dessa forma, a análise e caracterização de genes presentes em GEIs associada à metagenômica podem ajudar não somente na descoberta de novas proteínas, outrora tratadas como hipotéticas, como na elucidação da fonte e função desses genes.

IDENTIFICAÇÃO DE ILHAS DE PATOGENICIDADE *IN SILICO*

Apesar de eficientes na identificação de eventos de transferência horizontal, os métodos baseados nas características descritas acima, i.e., assinatura genômica, tRNAs flanqueadores, presença de transposase e proteínas hipotéticas, não são voltados para a classificação das GEIs, uma vez que os mesmos não consideram o conteúdo gênico da região como um todo. Além disso, as regiões adquiridas podem exibir desvios somente no conteúdo G+C ou no uso de códon separadamente, o que pode criar um obstáculo durante o processo de identificação quando se utiliza apenas uma das características citadas anteriormente. Contudo, existem ferramentas voltadas para a identificação das PAIs (uma das classes de GEIs) que usam uma estratégia combinada para superar esses obstáculos, isto é, consideram várias características relacionadas às mesmas durante o processo de identificação.

O primeiro deles, PredictBias, realiza uma análise da assinatura genômica e identificação de proteínas com papel na virulência, classificando-as em: PAIs (Composição enviesada), quando apresentam características de provável transferência horizontal; e, PAIs (Composição não enviesada), quando não apresentam sinais indicativos de transferência, mas estão ausentes em organismo correlato (Pundhir *et al*., 2008).

O segundo, IslandViewer, realiza uma análise combinada utilizando 3 programas já descritos na literatura, sejam eles: Colombo/SIGI-HMM, baseado na análise de uso de códon de cada CDS do genoma; IslandPick, que caracteriza as PAIs pela ausência em organismo filogeneticamente próximo; e, IslandPath-DIMOB, que realiza a classificação através da identificação de regiões que apresentam desvio no conteúdo de dinucleotídeos e que possuem genes relacionados à mobilidade (Langille & Brinkman, 2009; Langille *et al.*, 2008; Waack *et al.*, 2006).

Apesar de o PredictBias e o IslandViewer serem programas robustos que usam uma estratégia combinada, eles apresentam algumas retrições. O PredictBias, por exemplo, só pode ser utilizado através de uma interface web disponível "online" e, para tanto, a seqüência genômica deve ser submetida através dessa interface para então ser analisada no servidor. Essa interface web torna-se uma limitação em casos onde a seqüência genômica ainda não foi devidamente publicada e, portanto, os dados não podem ser enviados a terceiros. O programa IslandViewer, por outro lado, apresenta um código fonte para instalação em uma máquina local, contudo, um dos programas requeridos por ele, o programa IslandPick, apresenta uma grande dependência de um banco de dados MySQL composto de todos os genomas bacterianos publicados e, por essa razão, sua análise demanda muito tempo. Além disso, este programa depende de um servidor com alto desempenho de processamento e com uma configuração fora do convencional.

Ao enfrentar tais problemas técnicos, nosso grupo (Laboratório de Genética Celular e Molecular, LGCM), juntamente com o grupo do Laboratório de Polimorfismo de DNA (LPDNA), optou por desenvolver uma ferramenta, o PIPS (Pathogenicity Island Prediction Software – Artigo submetido), tendo como principais objetivos: (1) disponibilizar um código fonte aberto para instalação que exigisse um conhecimento mínimo da plataforma Linux; (2) realizar análises de forma robusta; e (3) possuir uma maior eficiência na identificação das PAIs quando comparado com os programas previamente descritos. O programa PIPS já encontra-se submetido e possui um Sítio com o código fonte disponibilizado e com a possibilidade de realização de análises *online* de forma fácil e intuitiva (http://www.genoma.ufpa.br/lgcm/pips).

O PIPS realiza análises das PAIs empregando várias estratégias baseadas em características, como: desvio de uso de códon (Colombo/SIGI-HMM) e de conteúdo G+C (EMBOSS/geecee; gccontent.pl); presença de fatores de virulência (mVIRdb), tRNAs flanqueadores (tRNAscan-SE) e transposase (ENTREZ database); e ausência em bactéria não-patogênica do mesmo gênero ou espécie correlata (ACT: the Artemis Comparison tool; plasticity.pl; plasticity2.pl). Na figura 01 pode ser visualizado um organograma onde estão representados os passos realizados pelo PIPS na identificação das PAIs.

Como pode ser visto, o programa realiza uma identificação automáticas das PAIs, assim como provê os arquivos necessários para a realização de uma posterior curadoria manual.

Comparação entre o PIPS, IslandViewer e PredictBias

Para comparar a eficiência do PIPS na identificação das PAIs em relação a outros programas disponíveis, foram realizadas análises de sensibilidade e especificidade usando dados sobre as 13 PAIs de *C. diphtheriae* retirados da literatura como controle positivo (Tabela 1). As seqüências codificantes (Coding Sequences, CDSs) de *C. diphtheriae* foram dadas como positivas quando as mesmas estavam localizadas em uma PAI e como negativas quando estavam localizadas em outra região ao longo da seqüência genômica.

Como visto na tabela 1, o programa PredictBias apresentou uma boa sensibilidade, mas com especificidade baixa, ao serem utilizadas apenas regiões identificadas como Ilhas de Patogenicidade como positivas para o teste (PredictBias_PAI). Ao contrário, ao considerar as regiões definidas como Ilhas Genômicas como positivas para o teste (PredictBias_GEI), o mesmo apresentou baixa sensibilidade e alta especificidade. Este resultado é um reflexo da classificação errônea de algumas Ilhas de Patogenicidade como Ilhas Genômicas. Este erro pode ter ocorrido, principalmente, devido ao banco de dados de fatores de virulência utilizado pelo programa. O banco de dados utilizado pelo PredictBias foi criado utilizando-se uma pesquisa no NCBI pelas palavras-chave: 'Virulence', 'Adhesin', 'Siderophore', 'Invasin', 'Endotoxin' e 'Exotoxin' (Zhou *et al.*, 2007). O seu tamanho é um fator determinante para o discernimento entre PAIs e GEIs, pois quanto maior a quantidade de informações, maior a probabilidade de uma classificação correta dos genes de virulência e, conseqüentemente, das PAIs.

O programa IslandViewer não identificou 3 PAIs de *C. diphtheriae* e, mesmo dentre as identificadas, houve grandes variações em comparação com as Ilhas descritas na literatura. Dentre os 3 programas utilizados pelo IslandViewer, o que obteve melhor desempenho foi o Is-

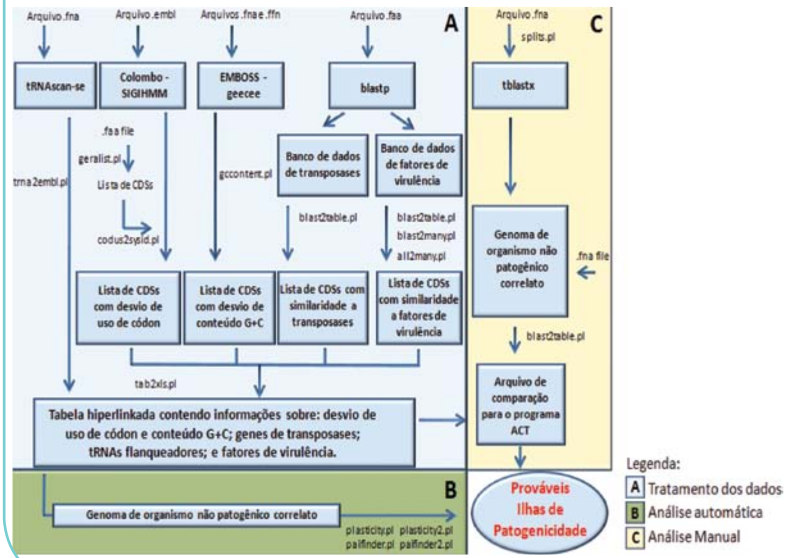

Arquivo .fna
Arquivo .embl
Arquivos .fna e .ffn
Arquivo .faa
A
tRNAscan-se
Colombo - SIGIHMM
EMBOSS - geecee
blastp
.faa file
geralist.pl
Lista de CDSs
codus2sysid.pl
trna2embl.pl
gccontent.pl
Banco de dados de transposases
Banco de dados de fatores de virulência
blast2table.pl
blast2table.pl
blast2many.pl
all2many.pl
Lista de CDSs com desvio de uso de códon
Lista de CDSs com desvio de conteúdo G+C
Lista de CDSs com similaridade a transposases
Lista de CDSs com similaridade a fatores de virulência
tab2xls.pl
Tabela hiperlinkada contendo informações sobre: desvio de uso de códon e conteúdo G+C; genes de transposases; tRNAs flanqueadores; e fatores de virulência.
B
Genoma de organismo não patogênico correlato
plasticity.pl plasticity2.pl
paifinder.pl paifinder2.pl
Arquivo .fna
C
splits.pl
tblastx
Genoma de organismo não patogênico correlato
.fna file
blast2table.pl
Arquivo de comparação para o programa ACT
Prováveis Ilhas de Patogenicidade
Legenda:
A Tratamento dos dados
B Análise automática
C Análise Manual

Figura 1. *Organograma demonstrando a execução coordenada dos programas do método combinado de identificação de Ilhas de Patogenicidade. (A) tratamento dos dados; (B) Análise automática; e (C) Análise Manual.*

TABELA 1 - COMPARAÇÃO ENTRE OS PROGRAMAS NA IDENTIFICAÇÃO DAS 13 PAIS DE C. DIPHTHERIAE.

Programas	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
IslandPath_DIMOB	13.6	98.3	89.2
IslandPick	65.2	81.9	80.1
SIGI_HMM	14.0	94.9	86.2
IslandViewer	74.4	76.4	76.2
PredictBias_GEI	30.8	84.4	78.6
PredictBias_PAI	2.4	88.7	79.4
PIPS_Auto	86.4	85.0	85.1
PIPS_Manual	96.8	87.1	88.1

landPick que realiza análise baseado na ausência em organismo correlato, nesse caso, *C. glutamicum*. Este fato corrobora para a importância da comparação genômica entre a bactéria analisada e organismo não patogênico do mesmo gênero ou espécie correlata.

O IslandPath-DIMOB apresentou menor precisão na predição entre os programas testados, identificando corretamente apenas as PAIs 7 e 9 de *C. diphtheriae*. O Colombo/SIGI-HMM apresentou um desempenho médio, contudo, isso se deve principalmente à alta estringência na configuração do programa quando utilizado pelo IslandViewer. No PIPS, utilizando uma sensibilidade de 95%, o programa Colombo/SIGI-HMM apresentou um desempenho melhor e mostrou-se uma abordagem eficaz na identificação de regiões com desvio de uso de codon para análise manual.

O programa PIPS identificou corretamente 12 das 13 PAIs de *C. diphtheriae*. De acordo com a anotação genômica de *C. diphtheriae*, a única PAI não identificada pelo PIPS, PAI 5, apresentou um conteúdo G+C anômalo correspondente a 52,2%. Contudo, quando utilizado um valor limítrofe de 1.5 desvios padrão para a identificação de conteúdo G+C anômalo, foram encontrados valores de referência variando de 49,95% a 60,4%. Além disso, o programa Artemis: the Annotation Tool (Rutherford *et al.*, 2000)K;Parkhill, J;Crook, J;Horsnell, T;Rice, P;Rajandream, M A;Barrell, B</Author><Journal>Bioinformatics</Journal><Month>Oct</Month><Number>10</Number><Pages>944-5</Pages><Title>Artemis: sequence visualization and annotation</Title><Volume>16</Volume><Year>2000</Year><URL>http://view.ncbi.nlm.nih.gov/pubmed/11120685</URL><ISBN>1367-4803</ISBN><CitationRanges>;0*17*-1;17*19*-1;19*23*-1;10*17*1</CitationRanges><DuplicateInfo></DuplicateInfo> não identificou nenhuma variação no conteúdo G+C na PAI 5. Junto a isso, exceto pela ausência em *C. glutamicum* (bactéria não patogênica do mesmo gênero), a PAI 5 não apresenta nenhuma outra característica de ilha de patogenicidade. Por fim, os programas IslandViewer e PredictBias corroboram com o PIPS na possível classificação errônea, na literatura, da PAI 5 como uma Ilha de Patogenicidade.

A análise automática usando PIPS apresentou um melhor desempenho em relação a técnicas disponíveis atualmente, contudo, os resultados da análise manual permitiram uma melhor identificação das PAIs, mostrando a importância de uma curadoria manual dos dados tendo como base o conhecimento biológico do organismo estudado.

IDENTIFICAÇÃO DAS ILHAS DE PATOGENICIDADE DE *E. COLI*UROPATOGÊNICA LINHAGEM *CFT073*

Além da validação do PIPS em *C. diphtheriae*, foi realizada uma comparação entre PIPS e os demais programas na identificação das PAIs de *E. coli* uropatogênica linhagem *CFT073* para analisar a performance do PIPS com bactérias Gram negativas. *E. coli* uropatogênica foi escolhida por possuir várias PAIs descritas na literatura. Foram utilizadas 13 PAIs descritas por Lloyd *et al.* (2007) como padrão ouro e a acurácia do PIPS foi comparada com IslandViewer e PredictBias como descrito previamente para *C. diphtheriae*. *E. coli* linhagem *K-12* foi utilizada como organismo não patogênico proximamente relacionado para essas análises e a sensibilidade e especificidade dos métodos estão representados na Tabela 2.

Os resultados demonstraram que, apesar das especificidades alcançadas pelos outros programas serem maiores, o PIPS exibe uma sensibilidade muito alta em comparação aos mesmos. Essa perda na especificidade pode ser proveniente de novas PAIs não identificadas previamente na literatura em vez de resultados falso positivos e a maior acurácia do PIPS corrobora para essa informação, isto é, PIPS apresentou o melhor desempenho na identificação de CDSs verdadeiro positivas e verdadeiro negativas dentro das PAIs de *E. coli* uropatogênica linhagem *CFT 073*.

O PIPS apresentou maior acurácia que os demais programas na análise das PAIs de *E. coli* além de identificar 11 PAIs adicionais em *C. diphtheriae*. Contudo, apesar da acurácia do PIPS e existência de outros programas, como os descritos

TABELA 2 - COMPARAÇÃO ENTRE OS PROGRAMAS NA IDENTIFICAÇÃO DAS 13 PAIS DE *E. COLI* UROPATOGÊNICA LINHAGEM *CFT073.*

Programas	Sensibilidade (%)	Especificidade (%)	Acurácia (%)
IslandPath_DIMOB	44.5	99.3	90.2
IslandPick	7.5	99.7	84.5
SIGI_HMM	21.9	96.9	84.5
IslandViewer	55.8	96.2	89.5
PredictBias_IG	60.0	93.7	88.1
PredictBias_PAI	39.2	96.2	86.8
PIPS_Auto	94.8	93.7	93.9

acima, ainda existe uma carência de programas voltados para a identificação das demais classes de Ilhas Genômicas que, quando somado ao número crescente de genomas bacterianos disponibilizados nos bancos de dados de domínio público, revela um grande potencial para novas abordagens bioinformáticas na área.

REFERÊNCIAS BIBLIOGRÁFICAS

Barcellos FG, Menna P, da Silva Batista JS & Hungria M. **Evidence of horizontal transfer of symbiotic genes from a *Bradyrhizobium japonicum* inoculant strain to indigenous diazotrophs Sinorhizobium (Ensifer) fredii and Bradyrhizobium elkanii in a Brazilian Savannah soil**. *Appl Environ Microbiol* (2007) 73: pp. 2635-2643.

Benedek O & Schubert S. **Mobility of the *Yersinia* High-Pathogenicity Island (HPI): transfer mechanisms of pathogenicity islands (PAIS) revisited (a review)**. *Acta Microbiol Immunol Hung* (2007) 54: pp. 89-105.

Blum G, Ott M, Lischewski A, Ritter A, Imrich H, Tschäpe H & Hacker J. **Excision of large DNA regions termed pathogenicity islands from tRNA-specific loci in the chromosome of an Escherichia coli wild-type pathogen**. *Infect Immun* (1994) 62: pp. 606-614.

Brown JS, Gilliland SM, Ruiz-Albert J & Holden DW.**Characterization of pit, a Streptococcus pneumoniae iron uptake ABC transporter**. *Infect Immun* (2002) 70: pp. 4389-4398.

Brüssow H, Canchaya C & Hardt W. **Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion**. *Microbiol Mol Biol Rev* (2004) 68: pp. 560-602.

Daniel R. **The soil metagenome--a rich resource for the discovery of novel natural products**.*Curr Opin Biotechnol* (2004) 15: pp. 199-204.

Dobrindt U & Hacker J. **Whole genome plasticity in pathogenic bacteria**.*Curr Opin Microbiol* (2001) 4: pp. 550-557.

Dobrindt U, Janke B, Piechaczek K, Nagy G, Ziebuhr W, Fischer G, Schierhorn A, Hecker M, Blum-Oehler G & Hacker J. **Toxin genes on pathogenicity islands: impact for microbial evolution**. *Int J Med Microbiol* (2000) 290: pp. 307-311.

Fratamico PM, Yan X, Caprioli A, Esposito G, Needleman DS, Pepe T, Tozzoli R, Cortesi ML & Morabito S. **The complete DNA sequence and analysis of the virulence plasmid and of five additional plasmids carried by Shiga toxin-producing Escherichia coli O26:H11 strain H30**. *Int J Med Microbiol* (2011) 301: pp. 192-203.

Gal-Mor O & Finlay BB.**Pathogenicity islands: a molecular toolbox for bacterial virulence**. *Cell Microbiol* (2006) 8: pp. 1707-1719.

Gao J & Chen L. **Theoretical methods for identifying important functional genes in bacterial genomes**.*Res Microbiol* (2010) 161: pp. 1-8.

Hacker J & Carniel E. **Ecological fitness, genomic islands and bacterial pathogenicity.A Darwinian view of the evolution of microbes**.*EMBO Rep* (2001) 2: pp. 376-381.

Hacker J, Bender L, Ott M, Wingender J, Lund B, Marre R & Goebel W. **Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal Escherichia coli isolates**. *Microb Pathog* (1990) 8: pp. 213-225.

Hershberg R & Petrov DA. **General rules for optimal codon choice**. *PLoS Genet* (2009) 5: p. e1000556.

Hou YM. **Transfer RNAs and pathogenicity islands**. *Trends Biochem Sci* (1999) 24: pp. 295-298.

Hsiao WWL, Ung K, Aeschliman D, Bryan J, Finlay BB & Brinkman FSL. **Evidence of a large novel gene pool associated with prokaryotic genomic islands**. *PLoS Genet* (2005) 1: p. e62.

Johnson TJ, Thorsness JL, Anderson CP, Lynne AM, Foley SL, Han J, Fricke WF, McDermott PF, White DG, Khatri M, Stell AL, Flores C & Singer RS. **Horizontal gene transfer of a ColV plasmid has resulted in a dominant avian clonal type of Salmonella enterica serovar Kentucky**. *PLoS One* (2010) 5: p. e15524.

Karaolis DK, Johnson JA, Bailey CC, Boedeker EC, Kaper JB & Reeves PR. **A Vibrio cholerae pathogenicity island associated with epidemic and pandemic strains**. *Proc Natl Acad Sci U S A* (1998) 95: pp. 3134-3139.

Karlin S, Mrázek J & Campbell AM. **Codon usages in different gene classes of the Escherichia coli genome**.*Mol Microbiol* (1998) 29: pp. 1341-1355.

Kauser F, Khan AA, Hussain MA, Carroll IM, Ahmad N, Tiwari S, Shouche Y, Das B, Alam M, Ali SM, Habibullah CM, Sierra R, Megraud F, Sechi LA & Ahmed N. **The cag pathogenicity island of *Helicobacter pylori* is disrupted in the majority of patient isolates from different human populations**. *J Clin Microbiol* (2004) 42: pp. 5302-5308.

Krizova L & Nemec A.**A 63 kb genomic resistance island found in a multidrug-resistant Acinetobacter baumannii isolate of European clone I from 1977**. *J Antimicrob Chemother* (2010) 65: pp. 1915-1918.

Langille MGI & Brinkman FSL.**IslandViewer: an integrated interface for computational identification and visualization of genomic islands**. *Bioinformatics* (2009) 25: pp. 664-665.

Langille MGI, Hsiao WWL & Brinkman FSL. **Evaluation of genomic island predictors using a comparative genomics approach**. *BMC Bioinformatics* (2008) 9: p. 329.

Lesic B, Bach S, Ghigo J, Dobrindt U, Hacker J & Carniel E. **Excision of the high-pathogenicity island of *Yersinia pseudotuberculosis* requires the combined actions of its cognate integrase and Hef, a new recombination directionality factor**. *Mol Microbiol* (2004) 52: pp. 1337-1348.

Lloyd AL, Rasko DA & Mobley HLT.**Defining genomic islands and uropathogen-specific genes in uropathogenic Escherichia coli**. *J Bacteriol* (2007) 189: pp. 3532-3546.

Mattos-Guaraldi AL, Duarte Formiga LC & Pereira GA. **Cell surface components and adhesion in Corynebacterium diphtheriae**. *Microbes Infect* (2000) 2: pp. 1507-1512.

Maurelli AT. **Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens**.*FEMS Microbiol Lett* (2007) 267: pp. 1-8.

Maurelli AT, FernÃ¡ndez RE, Bloch CA, Rode CK & Fasano A. **“Black holes” and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli**. *Proc Natl Acad Sci U S A* (1998) 95: pp. 3943-3948.

Ou H, Chen L, Lonnen J, Chaudhuri RR, Thani AB, Smith R, Garton NJ, Hinton J, Pallen M, Barer MR & Rajakumar K. **A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria**. *Nucleic Acids Res* (2006) 34: p. e3.

Pundhir S, Vijayvargiya H & Kumar A. **PredictBias: a server for the identification of genomic and pathogenicity islands in prokaryotes**. *In Silico Biol* (2008) 8: pp. 223-234.

Rajanna C, Revazishvili T, Rashid MH, Chubinidze S, Bakanidze L, Tsanava S, Imnadze P, Bishop-Lilly KA, Sozhamannan S, Gibbons HS, Morris JG & Sulakvelidze A. **Characterization of pPCP1 Plasmids in Yersinia pestis Strains Isolated from the Former Soviet Union**. *Int J Microbiol* (2010) 2010: p. 760819.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B. **Artemis: sequence visualization and annotation**. *Bioinformatics*(2000) 16: pp. 944-5.

Schmidt H & Hensel M. **Pathogenicity islands in bacterial pathogenesis**. *Clin Microbiol Rev* (2004) 17: pp. 14-56.

Schumann W. **Thermosensors in eubacteria: role and evolution**. *J Biosci* (2007) 32: pp. 549-557.

Stock T & Rother M. **Selenoproteins in Archaea and Gram-positive bacteria**. *Biochim Biophys Acta* (2009) : .

Suzuki T & Sasakawa C. **Molecular basis of the intracellular spreading of Shigella**. *Infect Immun* (2001) 69: pp. 5959-5966.

Takai S, Hines SA, Sekizaki T, Nicholson VM, Alperin DA, Osaki M, Takamatsu D, Nakamura M, Suzuki K, Ogino N, Kakuda T, Dan H & Prescott JF. **DNA sequence and comparison of virulence plasmids from Rhodococcus equi ATCC 33701 and 103**.*Infect Immun* (2000) 68: pp. 6840-6847.

Tobes R & Pareja E. **Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements**. *BMC Genomics* (2006) 7: p. 62.

Tumapa S, Holden MTG, Vesaratchavest M, Wuthiekanun V, Limmathurotsakul D, Chierakul W, Feil EJ, Currie BJ, Day NPJ, Nierman WC & Peacock SJ. **Burkholderia pseudomallei genome plasticity associated with genomic island variation**. *BMC Genomics* (2008) 9: p. 190.

Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P & Merkl R. **Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models**. *BMC Bioinformatics* (2006) 7: p. 142.

Zhou CE, Smith J, Lam M, Zemla A, Dyer MD & Slezak T. **MvirDB--a microbial database of protein toxins, virulence factors and antibiotic resistance genes for bio-defence applications**. *Nucleic Acids Res* (2007) 35: p. D391-4.

6.2.5 Montagem de genomas bacterianos

As tecnologias de sequenciamento de próxima geração (NGS) produzem uma grande quantidade de dados, com milhões de leituras curtas por rodada, alta precisão, a um custo reduzido, uma realidade inimaginável em comparação com o sequenciamento de Sanger. No entanto, leituras curtas, regiões repetitivas do genoma e regiões de baixa complexidade das leituras, ou seja comuns em outros organismos, trazem novos desafios para genomas obtidos a partir dessas tecnologias. Esses desafios impulsionaram o desenvolvimento de novas abordagens computacionais para armazenamento e análise de dados oriundos dessa tecnologia. No texto a seguir, são apresentadas algumas ferramentas computacionais utilizadas no processo de montagem de genoma por meio dessas tecnologias, considerando desde o tratamento dos dados brutos, a geração de sequências *ab initio*, até a etapa de fechamento, produzindo a primeira versão de um genoma. Este capítulo de livro foi aceito para publicação pela editora científica *CRC Press* (*Taylor & Francis Group*), no livro intitulado "*OMICS: Applications in Biomedical, Agricultural and Environmental Sciences*".

ASSEMBLY OF BACTERIAL GENOMES

ARTUR SILVA[b], VINICIUS DE ABREU[a], ADRIANA CARNEIRO[b], SINTIA ALMEIDA[a], ROMMEL RAMOS[b], ANDERSON SANTOS[a], LOUISE CERDEIRA[b], SIOMAR SOARES[a], ANNE PINTO[a], LUIS GUIMARÃES[a], EUDES BARBOSA[a], PAULA SCHNEIDER[b], ANDERSON MIYOSHI[a], VASCO AZEVEDO[a]

[a]Universidade Federal de Minas Gerais – UFMG, Belo Horizonte – Minas Gerais, Brasil
[b]Universidade Federal do Pará – UFPA, Belém – Pará, Brasil

ABSTRACT: The next-generation sequencing technologies (NGS) generate a large amount of data capable of generating millions of short reads per run with high accuracy, at a reduced cost, a reality unimaginable with Sanger sequencing. However, short readings, repetitive regions in the genome and regions of low complexity of the readings bring new challenges for genome assemblies obtained from these technologies, which drove the development of new computational approaches to storage and analysis of these data. In this chapter, we present some computational tools used in the genome assembly process since the treatment of the raw data and generation of *ab initio* sequences until the finish step that produced the scaffold of the genome and close remains gaps.

KEY WORDS: Short reads; *ab initio* assembly; Reference assembly;

CONTENT

*Corresponding author: E-mail: vasco@icb.ufmg.br; ☎: +55-31-34092610

1. INTRODUCTION

One of the most important advances in biology has been our capacity to sequence the DNA of organisms. However, long after the conclusion of the human genome sequencing, there are still regions of the genome that are unworkable; that is they are difficult to mount and remain incomplete. Answers may come from second generation sequencing, which generates large volumes of data, generating millions of short reads per run, a reality that was unimaginable with Sanger sequencing.

Though we can now generate a high degree of sequencing coverage (*Figure 1*), the mounting of short reads *ab initio* is more complex than is re-sequencing. Various algorithms and bioinformatics tools have been developed to take care of these new problems and computational challenges, such as identification of repeat regions, sequencing errors and simultaneous manipulation of short reads [1,2].

After reads are generated by the sequencer, it is necessary to join them in a logical fashion to mount the final sequence. Over the years, various tools have been developed to resolve this issue, as for example, the assemblers PHRAP (www.phrap.org), ARACHNE [3] and Celera [4]. They have a paradigm in common, often referred to as overlap layout consensus [5]. This approach is quite similar to that used to resolve a jigsaw puzzle, as described below.

The first step consists of aligning the reads, two by two, exhaustively; the pairs of reads should present a consistent overlap from one read to another, similar to the search for pieces in a jigsaw puzzle that fit each other and have colors that match. Especially in eukaryotic genomes, the main difficulty is in distinguishing inexact overlaps due to sequencing error and similarities within the genome, such as highly conserved repeat regions [6]. Sequence alignment is a much studied area of bioinformatics, which consists in supplying the ideal alignment between two sequences as a function of an evaluation ’score’. Most of these methods are based on the Needleman-Wunsch algorithm [7], which uses the spatial dynamics of the possible alignments between the sequences. Many extensions have been conceived, as for example, for multiple alignments [8], local alignments [9] or rapid research in large data banks [10].

Currently, techniques that can be executed rapidly in parallel processors. In order to process short reads generated by second generation sequencing platforms, one of the solutions found for simultaneously manipulating thousands of sequences has been the use of computing clouds [11]. The assemblers detect a group of reads with consistent alignments with each other, forming contiguous sequences (contigs). This would be equivalent to partially forming an image by putting together pieces of a jigsaw puzzle. In both montages, genome and jigsaw puzzle, the process can be interrupted in ambiguous regions, where various continuations or holes are possible, and where no connecting piece was found [12].

Finally, the assembler tries to order and orient the contigs with each other in an *ab initio* manner; that is without the help of a reference sequence. Returning to the metaphor of a jigsaw puzzle, this would correspond to identifying corners and different parts of the image that relate to each other. In the final mounting phase, a scaffolded group of contigs will become available. Nevertheless, it is desirable to remove most possible holes, eventually converging to a group of integral chromosomes, that is, those that do not include breaks. This phase, called finishing, can be expensive and take considerable time, depending on the strategy used to close the occasional holes in the genome scaffold [13].

In the pre-mendelian era, inheritance was believed to be a merging of characteristics of both parents. Mendel’s experiments indicated that the reciprocal crosses between parents with contrasting characters produce same results. But there was existing evidence showing that reciprocal crosses may not always be alike. A classical example is a cross between a female horse and a male donkey which produces a mule, the reciprocal cross produces a hinney. No suitable explanation was available till lately when the significance of epigenetics started to be realized. In Greek language the prefix “Epi”means features that are “above” or in addition to something. Therefore epigenetic traits exist in addition to the traditional molecular basis for inheritance.

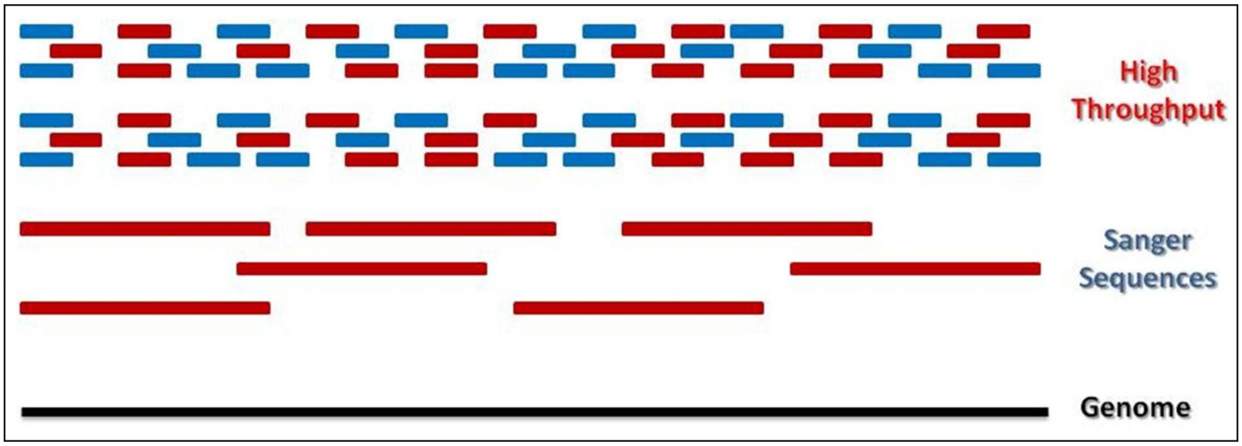


Figure 1: Coverage comparison - This figure is a qualitative comparison between the sequences generated by Sanger and those from the NGS platforms. There is higher abundance and depth of coverage with the short reads, but they are also significantly shorter, with little overlap available for assurance. Adapted from [14].

2 TREATMENT OF THE DATA

Pre-processing of the data, involving a quality filter and correction of the sequencing errors, is essential to increase the accuracy of the assemblies, as it prevents incorrect or low quality reads from becoming part of the genome assembly process.

2.1 Base quality

In 1998, Ewing and collaborators developed the PHRED algorithm, with the objective of determining the probability of occurrence of the one of the four nucleotides (A, C, G or T) for each base of a DNA sequence during the base-calling process; the intensity of the wave length that is obtained is used to calculate the PHRED quality value (Q), which is logarithmically related to observed probability of error for each base (P), according to the formula presented in *Figure 2*.

$$Q = -10 \log_{10} P$$

Figure 2: Formula for calculating the PHRED quality (Q) associated with the probability of error (P).

In *Table 1*, we can observe examples of PHRED quality associated with the probability of being incorrect and the precision of the identification of the base.

Table 1. The relation between PHRED quality-value and the probability of error and accuracy in the determination of a base.

PHRED Quality Score	***Error Probability***	***Accuracy***
10	1/10	90%
20	1/100	99%
30	1/1000	99.9%
40	1/10000	99.99%
50	1/100000	99.999%

The sequences obtained from automatic sequencers are not considered to be reliable do to their low quality at the extremities (*Figure 3*) and because of contaminants. Consequently, working on the quality of the data was fundamental so that the following phases of processing biological information would not be compromised [15]. In the case of the sequences obtained from next generation sequencing (NGS), despite the high degree of coverage, base quality should be evaluated. In this way the reads can be trimmed and quality filters applied. As examples of tools that do this quality treatment, we can cite Quality Assessment [16], Galaxy [17], ShortRead [18] and PIQA [19], the latter being used exclusively for Illumina data.

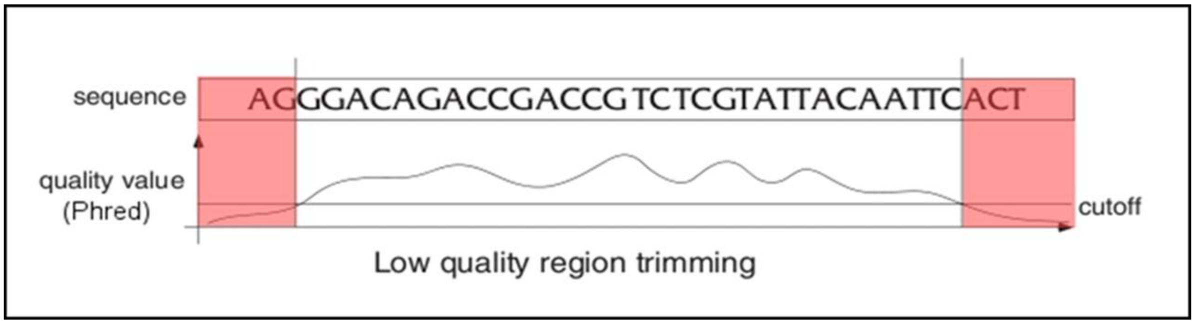

sequence
AGGGACAGACCGACCGTCTCGTATTACAATTCACT
quality value
(Phred)
cutoff
Low quality region trimming

Figure 3: Low quality of the ends of the reads obtained from automatic sequencers, which use the dideoxynucleotide method. Adapted from [15].

Analysis of sequence quality, followed by data treatment, makes it possible to reduce alignment errors because it provides precise alignment parameters, according to the data that are the objects of study [20]. In the assembly of genomes, software such as QRSA [21] propose mounting genomes with extension of sequences through analysis of base quality, giving better results than the assembler on which it was based: VCAKE [22], which does not take base quality into account to extend sequences.

In transcriptome studies with NGS platforms using RNA-Seq, evaluation of the quality of the data is extremely important, since the coverage represents the level of expression; consequently, quality filters using stringent parameters can provoke variations in the expression levels that are found [23].

2.2 Error Correction (Tools)

Despite the high degree of accuracy provided by NGS platforms, due to the extensive coverage generated by this equipment, sequencing errors can cause problems in the mounting of genomes when using an *ab initio* approach, because generation of contigs is very sensitive to these errors [24]. Consequently, to obtain better results, it is necessary to correct the errors before mounting the genomes, which will make the data more reliable [25].

In re-sequencing projects, in which the reads obtained from the sequencers are aligned against a reference genome, error correction can avoid elimination (trimming) of the 3' extremities of the read, due to the low quality observed when one tries to improve the alignments [26].

Some genome assemblers already include error correction procedures: SHARCGS only considers reads that have been produced by the sequencer N times, a parameterized value, and those that present overlap with other reads [25]. In 2001, Pevzner and collaborators used the spectral alignment method to correct errors, which consists of: given a string S in a spectrum T, formed by all of the continuous strings of fixed size (T-String), a search is made for the smallest number of modifications that need to be made in S to transform it into a T-string. This method of correction is implemented by the assembler EULER-SR [27] before the process of mounting the genome.

As examples of independent tools that can correct errors, we can cite SHREC [25] for SOLEXA/Illumina data, which uses a generalized tree of suffixes to process the data, the SOLiD Accuracy Enhancement Tool (SAET, http://solidsoftwaretools.com/gf/project/saet/) for SOLiD data, using an approach similar to that of EULER-SR, and Hybrid SHREC [24], which is based on the SHREC algorithm, but can process files from various different sequencing platforms.

3 Strategies for mounting genomes

Reads from the sequencer should be submitted to pre-processing, where base quality and sequencing errors are evaluated with software, commonly specific to corresponding sequencing platforms; they are then submitted to *ab initio* mounting and then oriented and ordered to produce the scaffold (*Figure 4*). If the mounting is done with a reference genome, after pre- processing, the reads are mapped against this reference, and after alignment is finished, a consensus sequences is produced.

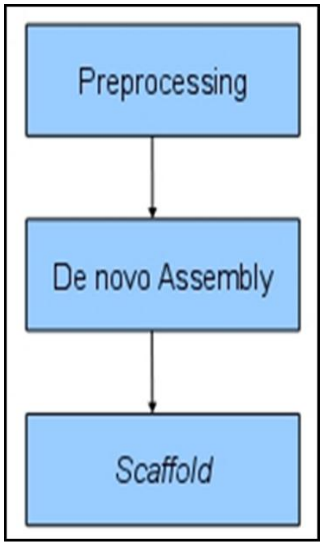


Figure 4: Steps used for ab initio mounting of genomes. After data treatment in the preprocessing stage, ab initio assembly is run, generating the contigs, which then are oriented and ordered to generate the scaffold.

3.1 Reference mounting

Basically, reference mounting consists in mapping the reads obtained from the sequencing against a reference genome (*Figure 5*), preferentially, of a phylogenetically closely related organism, making it possible to align a large part of the reads. However, the alignment configurations will also influence the quantity of leads that is utilized; consequently, the parameters such as depth of coverage and the number of mismatches that is permitted should be defined based on the sequencing information: estimated coverage and PHRED quality of the bases [20]. Mapping using a reference sequence provides identification of the nucleotide substitutions as well as indels, principally with the use of NGS platforms, due to the high degree of sequence coverage [28].

After mounting, regions of the reference genome that are covered are observed, representing gaps, which can occur as a function of the presence of a nucleotide sequence in the reference that does not occur in the sequenced organism, or because this region was not sequenced.

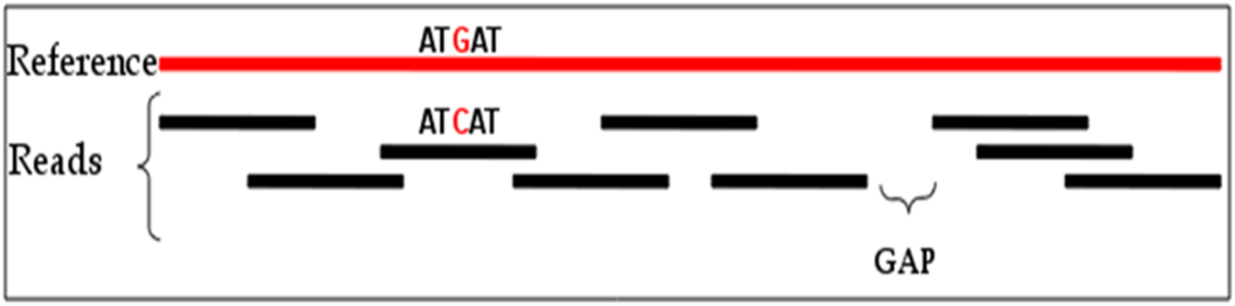


Figure 5: Alignment of reads against a reference genome, showing mismatches and a gap.

Among the problems with sequence mounting with the use of a reference, we can cite the representation of repeated regions, such as for example: the case of a reference genome that has two such regions and the sequenced organism has only one; during mapping against a reference the two reference regions will be covered, which can result in mounting errors (*Figure 6*).

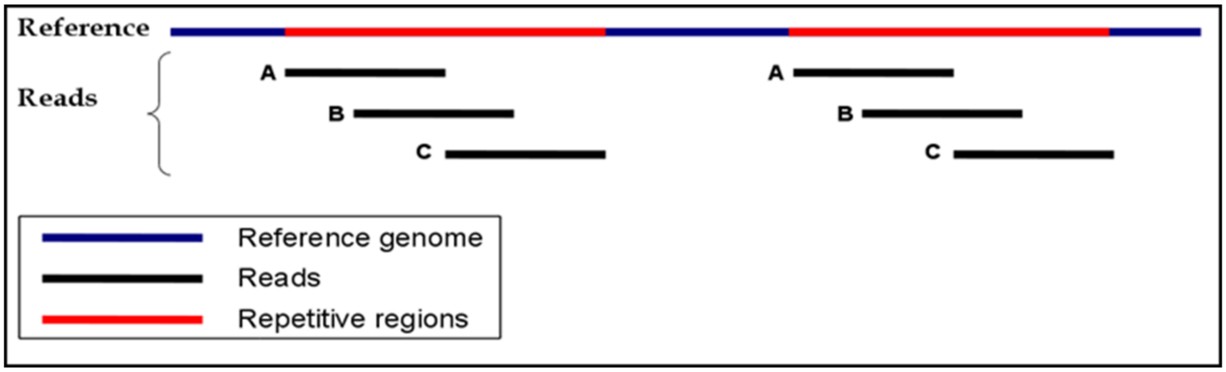


Figure 6: Double mapping of reads A, B and C in the reference genome, because it involves a repetitive region. However, in the sequenced genome the number of repeats can be different from that observed in the reference genome.

For mapping reads using a reference genome, one can use the softwares SHRiMP [29] and SOLiD BioScope (Applied Biosystems); both align in color-space SOLiD, SOAP2 [30], MAQ [31], RMAP [26] and ZOOM [32].

The program SOLiDTM BioScopeTM is an application based on Java that has various integrated tools in a web interface for re-sequencing and transcriptome analyses. The re-sequencing pipeline permits mapping of reads based on reference genomes, identifying single nucleotide polymorphisms (SNPs), INDELS (INsertions and DELetions) and inversions; as well as generating a consensus genome (Applied Biosystems).

3.2 Assemblers *ab initio*

This consists of reconstructing genome sequences without the aid of any other information besides the reads produced by the sequencing process. With this strategy, similarity alignments can be made among the reads themselves, or through overlap of k-mers. This allows, at the end of the alignment process, the formation of contiguous sequences (contigs) as seen in *Figure 7*. In most NGS assemblers, because they are based on graph theory, in which vertices and edges can represent overlap, a k-mer or a read varies according to the strategy that is used, in which the contigs are the paths formed in the graph. In this manner, the assemblers can be divided into: Greedy, Overlap layout consensus (OLC) or de Bruijn graph (DBG) algorithms; the latter uses a Eulerian pathway [28].

DBG is the approach that is mostly widely used by assemblers of short reads, as it works better with large numbers of reads, typical of NGS sequencers. The main programs that adopt this approach are: AllPaths [33], Euler-SR [34], SOAPdenovo [35] and Velvet [36]. Among these programs, Velvet is the only one that mounts short sequences in the color space format.

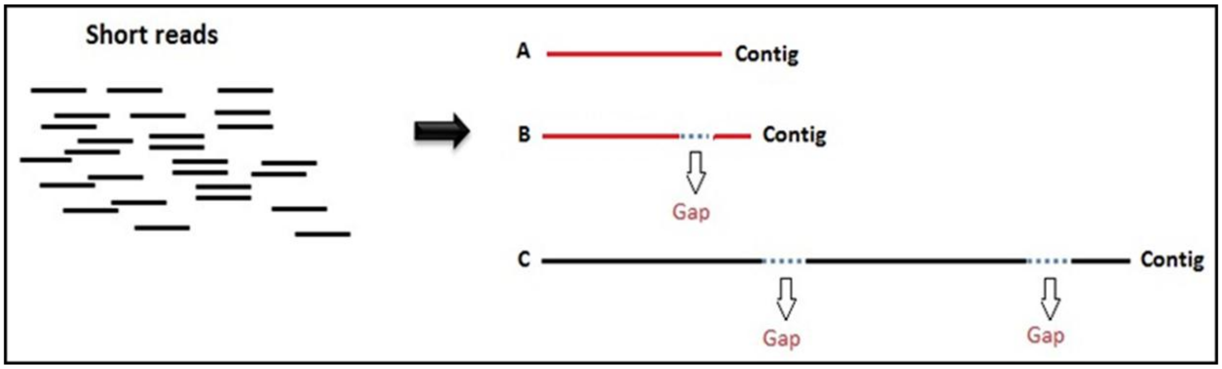


Figure 7: Alignment between the reads generated by the sequencing, finally obtaining scaffolded contigs, with and without gaps.

3.3 Challenges and difficulties for *ab initio* assembly

The limitations of *ab initio* mounting approaches are directly associated with the technological limitations because of the characteristics of the data generated by second-generation sequences, as well as the sizes of the reads and the volume of data that is generated, which exponentially increases the processing time and sometime makes mounting unviable. Within this context, various problems can occur, such as grouping of repeat regions, as can be seen in *Figure 8*; there are also regions in which sequences are of low quality, base compression in the sequencing, and even regions with a low degree of coverage due to the random character of the sequencing [37].

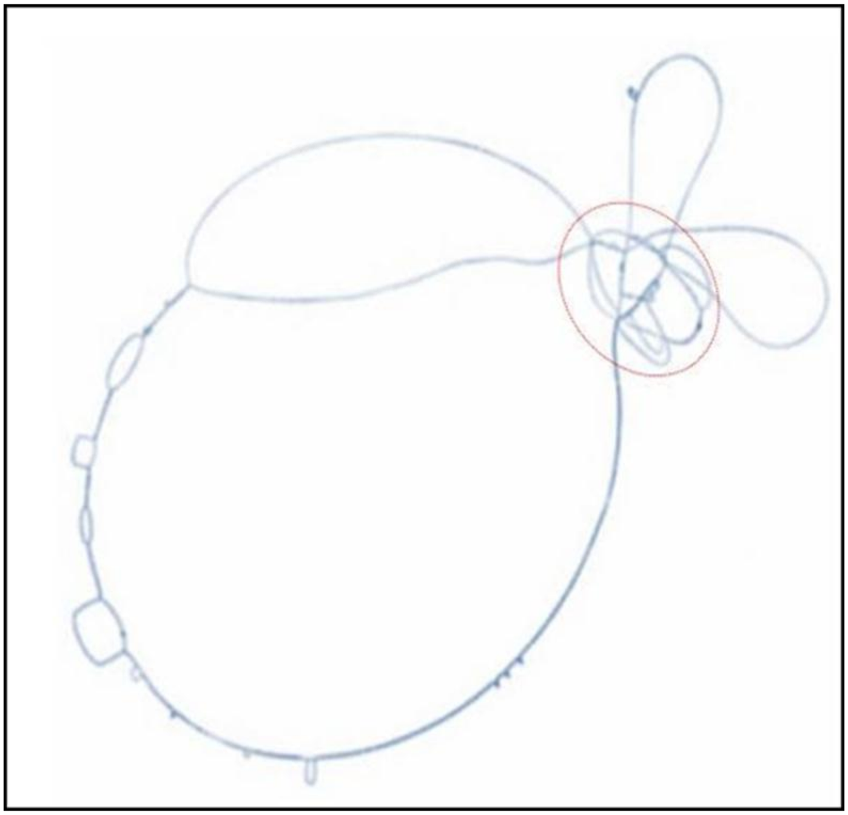


Figure 8: Figure showing two plasmids with a common locus. Adapted from [38].

One of the classic examples of problems with *ab initio* mounting is finding a path in the overlap graph that passes through each of the vertices only once (Hamiltonian pathway) or each edge only once (Eulerian pathway); this often results in the loss of connectivity between very distant sequences, showing that strategies based on graphs, especially the de Bruijn strategy, are extremely sensitive to sequencing errors [39].

These problems are more complex and common in assemblies made with short reads, as the number of reads is larger than with Sanger sequences, since the lengths are much shorter, which exponentially increases the size of the problem.

The large sizes of the conserved repeated regions also make the process of mounting the genome difficult and as it involves Eukaryotic genomes, which have very large repeat regions, sometimes this task is a problem that is hard to resolve [5].

In spite of the problems cited above, studies show that up to 96.29% of a gene can be reconstructed using short sequences, with sizes starting at 25 nucleotides [40].

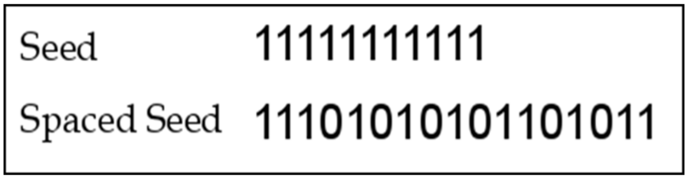
4 TOOLS FOR ASSEMBLING GENOMES

Second generation sequencers are capable of generating thousands of reads, providing a high degree of coverage and accuracy. As examples of these platforms, we can cite: SOLiD, Illumina and 454 FLX Titanium. Despite the reduction in sequencing costs, among other advantages, the reduction in the sizes of the reads, along with the increase in the number of reads, results in computational challenges for the processing of this data, principally for mounting genomes [28].

Mounting a genome consists of overlapping based on similarity of reads generated by a sequencer, in order to produce contiguous sequences (contigs), which in turn are aligned and oriented with each other to construct the scaffold. Mounting is called referenced mounting when it involves mapping reads in comparison with a reference sequence, while mounting reads without such a reference, is called *ab initio* mounting [31].

4.1 Tools for reference assembly

For alignment against a reference genome, there are two approaches: using hash tables and using prefix/suffix trees [20]. Many softwares that use hash tables define as the key, the subsequences obtained from the search sequence. The program tries to map identical sequences, known as seeds from the reference, so that the sequences can be subsequently extended. However, the use of templates with spaced seeds gives better results, because it considers internal mismatches (*Figure 9*). Even so, independent of whether seeds or spaced seeds, are used, the alignments do not accept gaps; identification of such gaps is made in a step after extension of the alignment.



Seed	11111111111
Spaced Seed	11101010101101011

Figure 9: Templates using seed in which an exact match of 11 bases is necessary to initiate the extension, and spaced seeds in which a match of 11 bases in required, but permitting the existence of internal mismatches

The mapping algorithms that use prefix/suffix trees search for exact alignments, represented by suffix trees, enhanced suffix array and FM-index [41]; then they extend the alignments considering mismatches. Among the tools that use suffix trees, we have MUMmer [42] and OASIS [43]. Among the alignment softwares based on enhanced suffix arrays, we can cite Vmatch [44] and Segemehl [45]. The FM-index method uses a small amount of memory (from 0.5 – 2 bytes per nucleotide), which can vary as a function of implementation and parameters that are used [20]; examples of such programs include: Bowtie [46], BWA [47], BWA-SW [48], SOAP2 [30], and BWT-SW [49].

4.2 Tools for *ab initio* assembly

Ording to Miller et al. [28], the mounting of genomes *ab initio* consists in aligning the reads with each other in order to produce contiguous sequences (contigs). The principal methodologies for NGS data are based on graphs, these being:

- Overlap Layout Consensus (OLC);
- de Bruijn Graph (DBG);
- Greedy

4.2.1 Tools that use Overlap Layout Consensus (OLC)

This is the most widely used approach for large sequences, such as those produce by Sanger; nevertheless, there are also applications based on this method for short reads, such as Edena [50]. The OLC method can be divided into three phases: overlap, layout and consensus. In the overlap phase, each read is compared to all of the others to identify overlaps, considering the minimum size of overlap and k-mer, which will affect the accuracy of the contigs.

Among the types of overlaps that are recorded, four categories are possible: containment, normal fitting, prefix and suffix fitting, as shown in *Figure 10* [51].

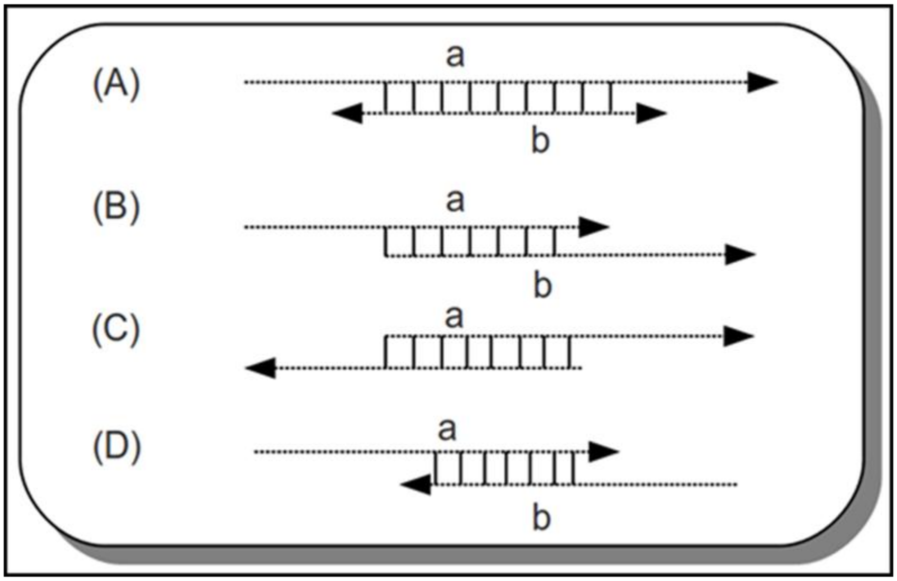


Figure 10: (A) Containment, (B) Partial overlap, (C) Prefix overlap (D) Suffix overlap. Adapted from [51].

In the layout phase, the information obtained from the previous phase is used to construct a graph *(Figure 11)*, which is reconstructed at each actualization. At the end of this phase, the first draft representing the genome will be generated, taking into account that in this phase various methods are used to simplify the pathways and remove errors detected in the graph, such as bubbles and linear extensions, known as dead paths [28].

In the consensus phase, multiple alignments are made of the fragments, progressively, to develop a consensus sequence [28], as shown in *Figure 12*.

As examples of the assemblers that use the OLC strategy we have: Celera Assembler [4], Arachne [3], CAP and PCAP [52]. Edena is the only program for the platforms Solexa and SOLiD that uses OLC [50].

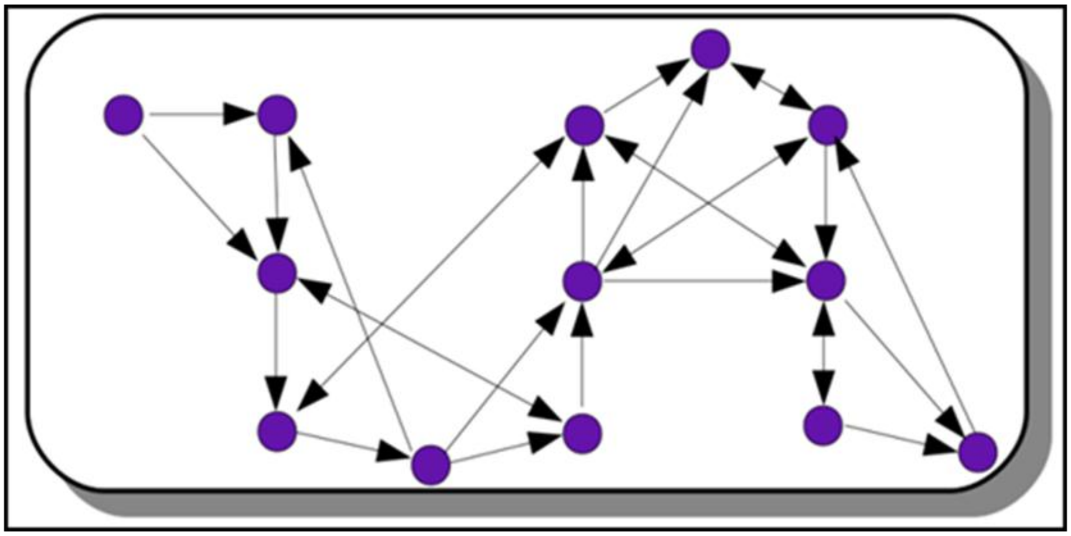


Figure 11: Layout Graph: G graph valid based on graph construction theory. First draft of what could be the genome. Adapted from [51].

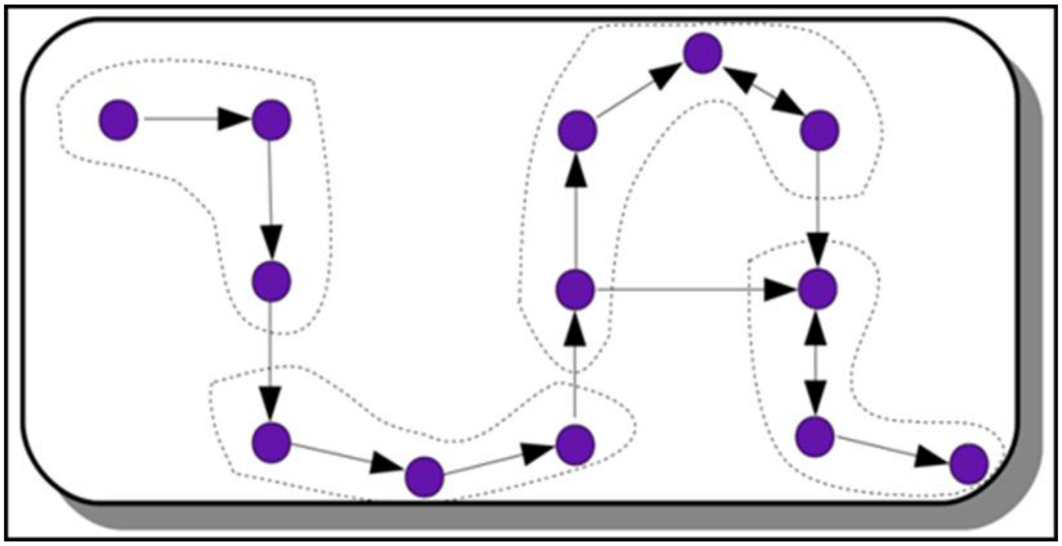


Figure 12: Consensus graph: Using progressive alignment guided by pairs. Adapted from [51]

4.2.2 Tools that use the de Bruijn graph

In 1995, Idury and Waterman introduced the use of a graph to represent a sequence assembly. Their method consisted of creating a vertex for each word. Then the vertices that correspond to the overlap of k-mers are connected; k can be represented by a sequence with a specific number of bases. The original vertex corresponds to prefix k-1 of the corresponding overlap region (k-mer), and the vertex destination of suffix k-1 of the same region, providing reconstruction of the sequence through a path that traverses each edge exactly once. [39] proposed a representation slightly different from the graph of the sequence, the so-called de Bruijn graph, which uses a Eulerian pathway; that is a pathway that visits each edge exactly once, through which the k-mers are represented as arcs or edges, and overlapping of the k-mers join their ends.

The classic method of assembling fragments is based on the notion of a graph of overlaps. The DNA sequence in *Figure 13a*, consists of four unique segments, A, B, C and D, and a repetition R. Each read corresponds a vertex in the overlap graph and two vertices are connected by an arc, if the corresponding reads overlap, as observed in *Figure 13b* [53]. It is possible to visualize the construction of this graph representing a DNA sequence as a ’line’ with repeat regions covered by a ’glue’ that ’links’ these regions *Figure 13c*. The de Bruijn in this case *Figure 13d*, consists of five arcs, in which each repetition corresponds to an arc instead of a collection of vertices in the overlap graph (*Figure 13b*). It can be seen that the de Bruijn graph (*Figure 13d*) represents the repetitions in a much simpler form than the overlap graph (*Figure 13b*). In *Figure 13*, there are two Eulerian pathways: one corresponding to reconstruction of the ARBRCRD sequence, while the other corresponds to the reconstruction of the ARCRBRD sequence. Different from the problem of the Hamiltonian pathway, the Eulerian pathway is less complex and is resolved even with graphs with millions of vertices, as there are linear-time algorithms that can provide a solution for them [54]. It is important to emphasize that a de Bruijn graph is centered in the k-mer, which means that its topology is not affected by fragmentation of the reads [12]. And compared with the overlap phase in OLC, the computational cost is much smaller, because the overlaps are not done against all [28].

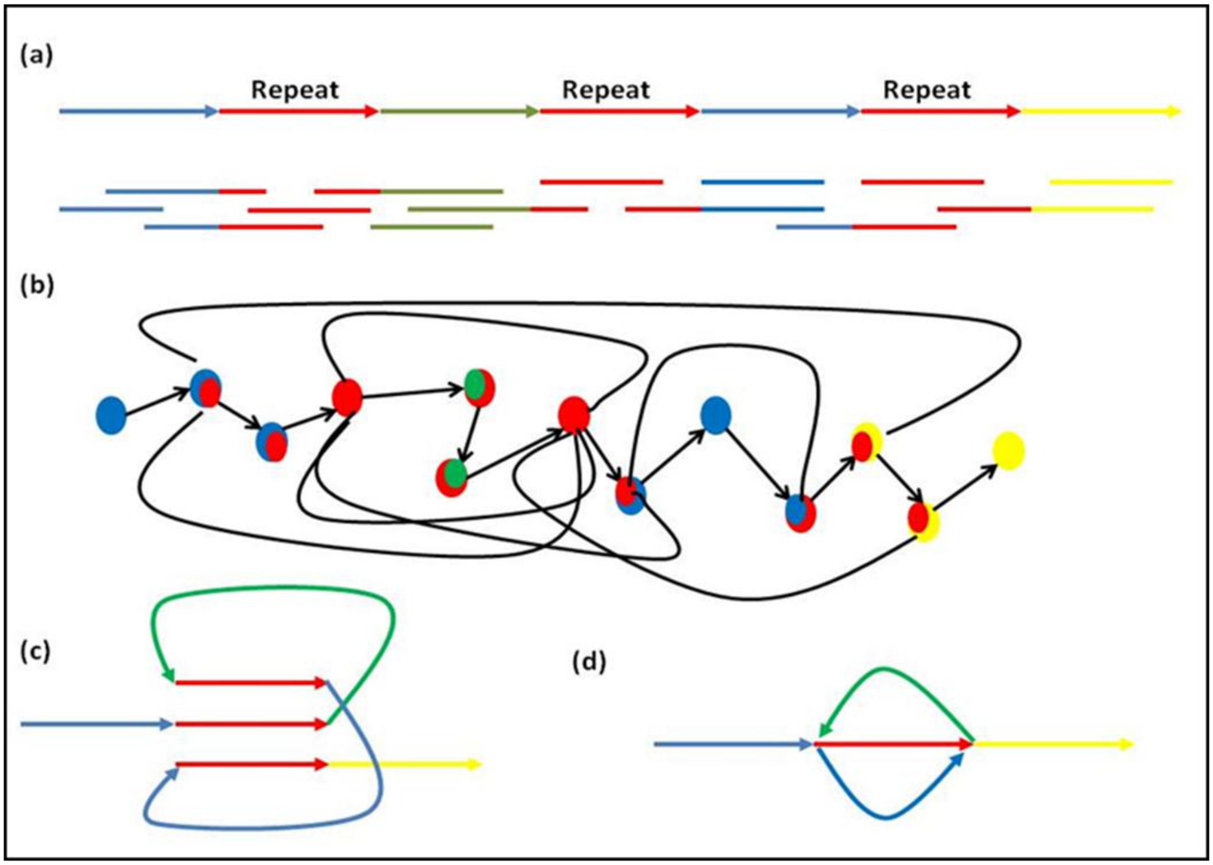


Figure 13: (a) DNA sequence with three repetitions R; (b) a diagram of the layout; (c) construction of the de Bruijn graph on the repetitions; (d) de Bruijn graph. Adapted from [53].

Operationally, the de Bruijn graph containing the vertices with length K is constructed with the result of the division of the K-mer, being linked in exact, identical overlapping for the previously-defined k-mer values. After construction of the graph, generally it is possible to simplify it without any loss of information. The reads of the vertices are interrupted and initiated again at each simplification. Simplification of two vertices is similar to the concatenation of two strands of characters [51].

As in other approaches, the assemblers add to their main algorithm, accessory algorithms to help remove assembly errors, such as reduction of redundant pathways, removal of bubbles or pathway loops and linear extensions; that is those that do not possess defined pathways. The main programs that adopt the use of the de Bruijn graph are: AllPaths [33], Euler-SR for short sequences [34], SOAPdenovo [54] and Velvet [36] specialized in the localization of the use of pared reads. Velvet is the only one that assembles short sequences in a color-space. Other assembly programs, such as for example, ABySS [56] were successful in constructing de Bruijn graphs, eliminating the limitations in the use of memory that are common during assemblies; *Table 2* shows the principal characteristics of each software [28].

In 2009, inspired by the ideas of [05], Zerbino implemented the program Velvet, the structure of which differs in various aspects. Among these, maps of k-mers are generated for the vertices and not for the arcs and there can be reverse complementary associated sequences, in order to obtain a bidirectional graph, as can be observed in *Figure 14* [57]. In this way the vertices can be connected by a directed edge or an arc. Due to the symmetry of the blocks, an arc goes from vertex A to B, a symmetrical arc goes from B to A. With any alteration of an arc, it is implicit that the same change will be made symmetrically in the paired arc.

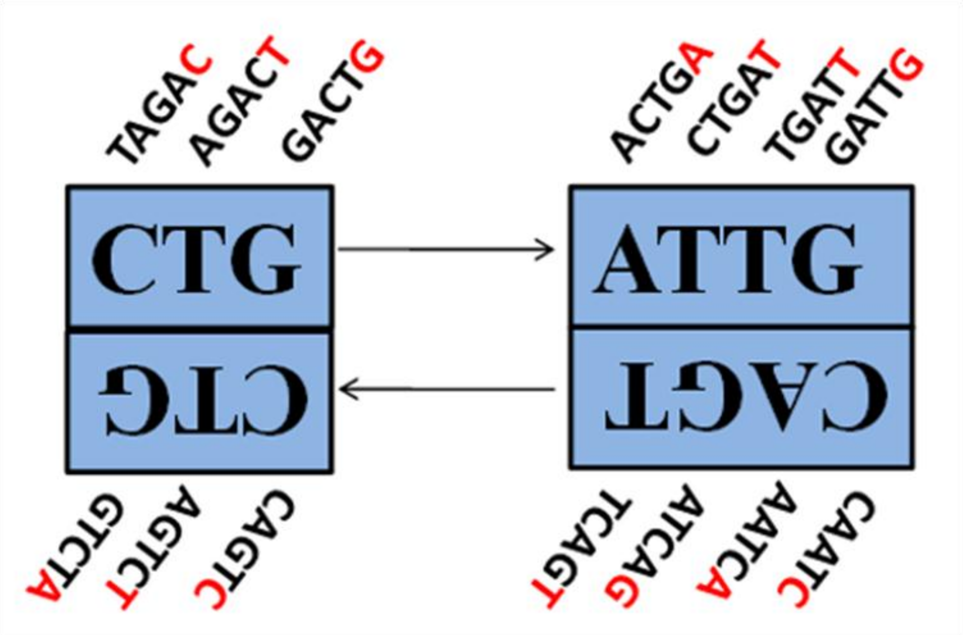


Figure 14: De Bruijn graph scheme implemented in Velvet. Adapted from [12].

Each vertex can be represented by a single rectangle, which represents a series of k-mer over- laps (in this case, k = 5) listed directly above or below. The only nucleotide of each k-mer is colored red. The arcs are represented as arrows between knots. The last k-mer has overlaps of an arc of origin with the first of its destination arcs. Each arc has a symmetrical arc [12].

Table 2. Feature comparison between de novo assemblers for whole-genome shotgun data from next-generation sequencing platforms. OLC refers to the overlap/layout/consensus architecture. DBG refers to the de Bruijn graph architecture. The table is based on the literature cited in the text. It may not reflect the current state of each software package. Adapted from [28].

	Algorithms Feature	**Greedy Assemblers**	**OLC Assemblers**	**DBG Assemblers**
Modeled features of reads	Base substitutions	-	-	Euler, AllPaths. SOAP
	Homopolymer miscount	-	CABOG	-
	Concentrated error in 3' end	-	-	Euler
	Flow space	-	Newbler	-
	Color space	-	Shorty	Velvet
Removal of erroneous reads	Based on K-mer frequencies	-	-	Euler, Velvet, AllPaths
	Based on K-mer freq. and QV	-	-	AllPaths
	For multiple values of K	-	-	AllPaths
	By alignment to other reads	-	CABOG	-
	By alignment and QV	SHARCGS	-	-
Correction of erroneous base calls	Based on K-mer frequencies	-	-	Euler, SOAP
	Based on K-mer freq. and QV	-	-	AllPaths
	Based on alignments	-	COBOG	-
Approaches to graph construction	Implicit	SSAKE, SHARCGS, VCAKE	-	-
	Reads as graph nodes	-	Edena, CABOG, Newbler	-
	K-mer as graph nodes	-	-	Euler, Velvet, AbySS, SOAP
	Simple path as graph nodes	-	-	AllPaths
	Multiple values of K	-	-	Euler
	Multiple overlap stringencies	HARCGS	-	-

Approaches to graph reduction	Filter overlaps	-	CABOG	-
	Greedy contig extension	SSAKE, SHARCGS, VCAKE	-	-
	Collapse simple paths	-	CABOG, Newbler	Euler, SOAP, Velvet
	Erosion of spurs	-	CABOG, Edena	Euler, AllPaths, SOAP, Velvet
	Transitive overlap reduction	-	Edena	-
	Bubble smoothing	-	Edena	Euler, SOAP, Velvet
	Bubble detection	-	-	AllPaths
	Reads separate tangled paths	-	-	Euler, SOAP
	Break at low coverage	-	-	SOAP, Velvet
	Break at high coverage	-	CABOG	Euler
	High coverage indicates repeat	-	CABOG	Velvet
	Special use of long reads	-	Shorty	Velvet

4.2.3 Tools that use Greedy Graph

The greedy algorithms were widely implemented in assembly programs for Sanger data, such as for example PHRAP, TIGR Assembler and CAP3 [1]. When new sequencing technology became available, other software was developed for assembling the NGS data (short reads) using different greedy strategies, such as SSAKE [58], SHARGCS [59], and VCAKE [22].

These algorithms can use an overlap-layout-consensus (OLC) approach or a de Bruijn graph, applying a basic function (*Figure 15*); starting with any read of a group of data, add another, and in this way numerous interactions are run until all possible operations are tested and the overlaps are identified, where a suffix of a read overlaps a prefix of another (*Figure 15a*). Each operation uses the overlap of a major score, measured by the size of the overlap between reads to make the next junction [28, 1, 2].

The quality of the overlaps is measured by the size and degree of identity (percentage bases shared between two reads in the overlap region. Also, simplification of the graph is based only on the size of the overlaps between reads, it being necessary to implement mechanisms to prevent misassemblies [28, 1]. The term "greedy" refers to the fact that the decisions taken by the algorithm occur as a function of a local quality (in the case of assembly, the quality of the overlaps between the reads), which may not be an optimal global solution; in this way, assemblers based on greedy can generate numerous misassemblies (*Figure 15b* and *15c*) [01].

During the assembly process, where the reads are added by iteration, the fragments are considered in descending order according to their quality, as explained previously. Consequently, in order to avoid misassemblies, the extension process is finalized when conflicting information is identified, as for example: two or more reads that extend a contig, but with no overlap between them (*Figure 15d*) [1].

The approaches based on greedy algorithms need a mechanism to avoid incorporating false-positive overlaps into contigs. The overlaps induced by repeat sequences can have high scores more than the overlaps of regions without repetitions; also, an assembly that generates a false-positive overlap will unite unrelated sequences to the ends of a repetition and produce a chimera [28]. Some assemblers use other algorithms to avoid including errors; SHARCGS, for example, includes a pre-processing step that filters out erroneous reads. The parameters of this filter can be modified by the user of the program [59].

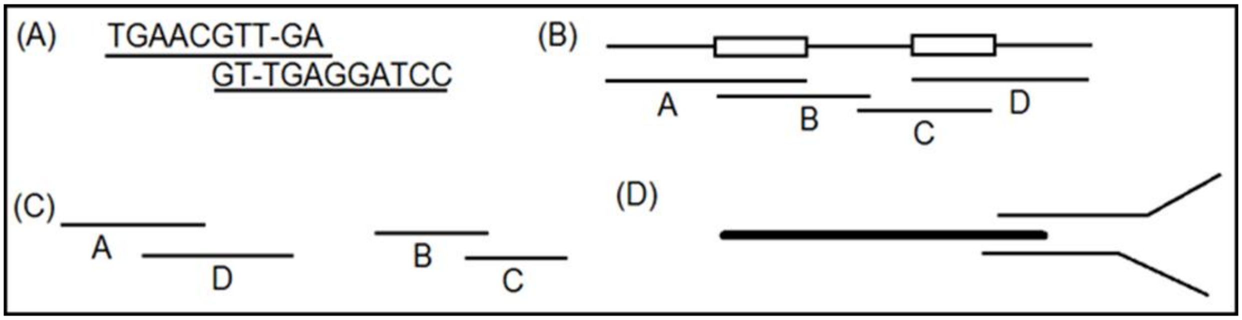


Figure 15: (A) Overlap between two reads, in which the overlapping region does not need to be a perfect match; (B) Example of correct assembly of a region of the genome that has two repetitive regions (box) using four reads (A to D). (C) Assembly generated by the greedy approach. Reads A and D are mounted first, incorrectly, due to identification of better overlap, and (D) discordance between two reads (fine lines) that could extend a contig (thick line). Extension of the contig could be finalized to avoid misassemblies. Adapted from [1].

5 TOOLS FOR VISUALIZING NGS DATA AND PRODUCING A SCAFFOLD

The development of Next-Generation Sequencing – NGS platforms, also named high-throughput sequencing (HTS) machines has opened new opportunities for biological applications, including re-sequencing of genomes, sequencing of the transcriptome, ChIP-seq, and discovery of miRNA [60]. These NGS technologies created a necessity to develop new tools to visualize the results of the assemblies and alignments of short reads [61]. Consequently, new challenges arose: a need to rapidly and efficiently process an enormous quantity of reads, a need for high quality interpretation of data, a user friendly interface, and the capacity to accept various formats of files produced by different sequencers and assemblers [62].

Visualization of the sequences generated from the process of mounting genomes can be done, for example, by the software Consed [63], which allows the data to be edited, and Hawkeye [64]. Various programs have been developed for NGS reads, including EagleView [65], Tablet [62], MapView [66], MaqView [31], SAMtools [55, 48] and IGV (Integrative Genomics Viewer) (www.broadinstitute.org).

The main differences between the visualizers are in the interfaces for presentation to the user, data processing velocity, as well as the different formats of the data entry files and develop- ment of the scaffold. Loading NGS data in programs such as Consed and Hawkeye, for exam- ple, requires a large amount of memory, which is normally not available to users of desktop computers. Eagleview is a visualization tool developed only for NGS, but it does not permit visualization of paired reads and it has memory limitations. MapView permits analysis of genetic variation, supports paired-end data and single-end reads and various different entry and output file formats [66].

The scaffold is made of DNA sequences that are reconstructed after sequencing; it can be composed of contigs, which should be ordered and oriented with each other with the help of a reference genome, and by gaps: regions where the DNA sequence is not recognized because it does not exist in the genome or because it is not covered in the sequencing or assembly [67]. Options for software for generating genome scaffolds using a reference genome include: Bambus [68] from the software package AMOS, which can be used as en- try and exit by the software Mummer [42]. Genscaff [69] uses contiguous sequences generated by an assembly program without the help of a reference genome, through implementation of graph theory. CLCBio Workbench (www.clcbio.com) and the software package Lasergene (www.dnastar.com), besides other functionalities, produce and edit the scaffold; however, they are commercial software packages.

6 CLOSING GAPS

An artifact related to the mounting of a genome is the formation of gaps (holes or spaces). Commonly, the strategy used to resolve these spaces would be to design specific primers for this region, and posterior alignment of the amplified sequeces by the primers, thereby closing the gap. However, for large gaps (2 Kb or longer), new primers are needed. Therefore, this process requires considerable time and becomes expensive [70].

Given this situation, we describe here an in silico strategy to resolve this type of artifact, a solution consisting of use of short reads generated by SOLID (Next Generation Sequencing) that were not mapped during the assembly process.

6.1 Description of the gap closing strategy

Align the short sequences in the flanking regions of the mapped genes. Then, the nucleotides that have a PHRED quality of 20 or more and a minimum of 10X coverage should be added manually (*Figure 16*; Items 1 and 2). This extension will close the small gaps (1 – 100 bp).

If there still are gaps, the short reads should be realigned in relation to the reference genome (*Figure 16*; Item 3), because with the production of the new contigs, new short reads align in the flanking regions of the gaps, forming what we call a merged contig. In this way, the genome could possibly be closed completely in silico, without using PCR. We emphasize that this strategy was used during the mounting of the genome of a strain of *Corynebacterium pseudoturbeculosis*, which was sequenced with the SOLiD plataform, which generated 19,091,361 reads (140X coverage). This system mapped 590 gap regions, closing 100% of the gaps [71].

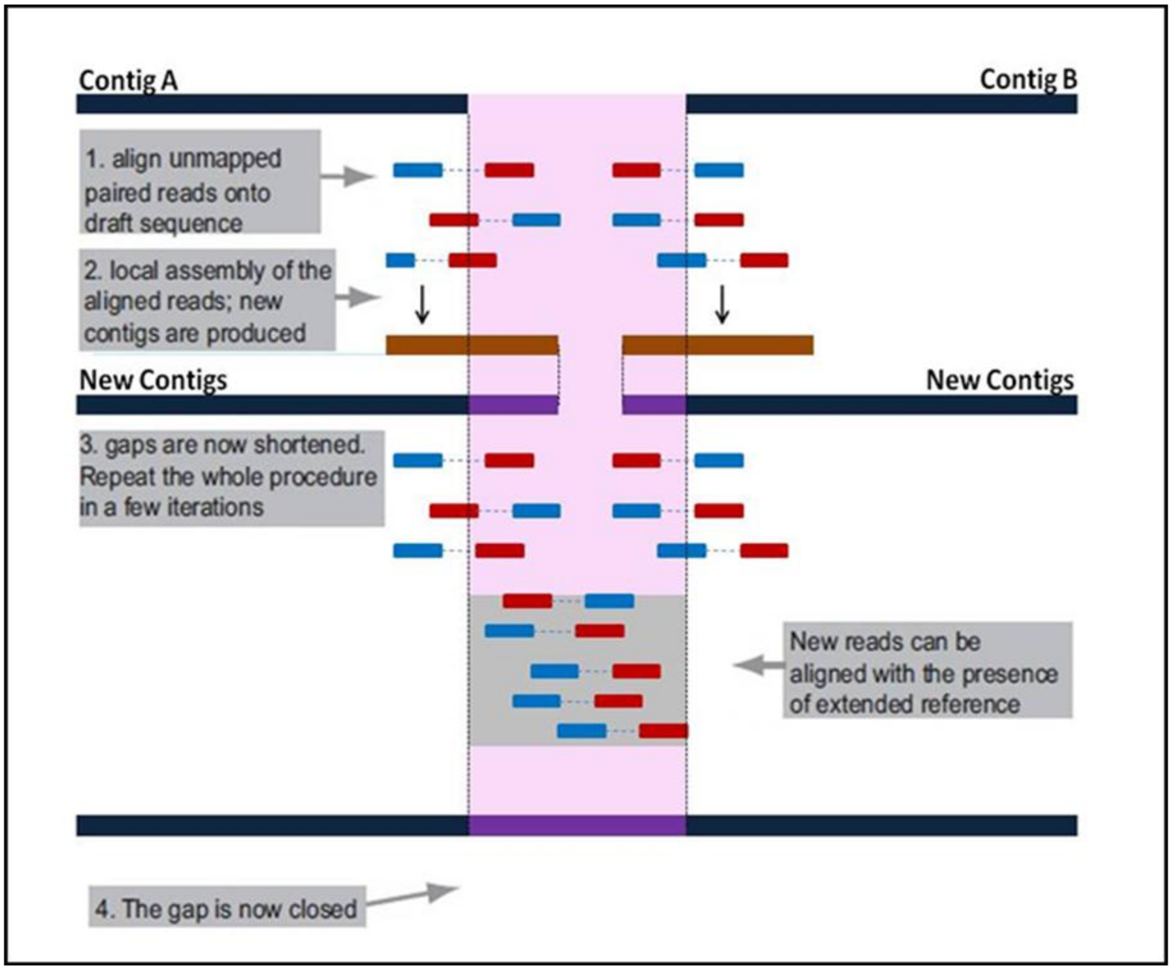

Contig A
Contig B
1. align unmapped paired reads onto draft sequence
2. local assembly of the aligned reads; new contigs are produced
New Contigs
New Contigs
3. gaps are now shortened. Repeat the whole procedure in a few iterations
New reads can be aligned with the presence of extended reference
4. The gap is now closed

Figure 16: Description of the strategy for closing the gaps: Step 1 – Short reads are aligned in the initial assembly; Step 2 – Short reads that align in terminal contigs are mounted in new contigs; Step 3 – The short reads are aligned against the updated sequence and the process is repeated until the gap is closed. Adapted from [70].

REFERENCES

[1] Pop M. (2009). Genome assembly reborn: recent computational challenges. Brief Bioinform, Vol 10(4), 354–366.

[2] Pop M, Salzberg SL (2008). Bioinformatics challenges of new sequencing technology. Trends in Genetics, Vol 24(3), 142–149..

[3] Batzoglou, S., Jaffe, D.B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J.P., and Lander, E.S. (2002). ARACHNE: A whole genome shotgun assembler. Genome Research, Vol 12, 177 – 189.

[4] Myers, E W, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, Kravitz SA, Mo- barry CM, Reinert KHJ, Remington KA, et al. (2000). A whole-genome assembly of drosophila. Science Vol. 287, 2196–2204.

[5] Pevzner, P. A., Tang, H., and Waterman, M. S. (2001). An eulerian path approach to DNA fragment assembly. Proc. Nati. Acad. Sci. USA 98.

[6] Phillippy, A. M., Schatz, M. C., and Pop, M. (2008). Genome assembly forensics: finding the elusive misassembly. Genome Biology Vol 9, R55.

[7] Needleman, S. and Wunsch, C. (1970). A general method applicable to the search for similar- ities in the amino acid sequence of two proteins. Journal of Molecular Biology Vol 48, 443–453.

[8] Higgins, D. and Sharp, P. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene, Vol 73, 237–244.

[9] Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. Journal of Molecular Biology, Vol 147, 195–197.

[10] Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local aligment search tool. Journal of Molecular Biology, Vol 215, 403 – 410.

[11] Bateman, A and Wood, M. (2009). Cloud computing. Bioinformatics, Vol 25(12), 1474.

[12] Zerbino D. (2009). Genome assembly and comparison using de Bruijn graphs. Ph.D. thesis, University of Cambridge.

[13] Cole, C. G., McCann, O. T., Oliver, J. E. C. K., Willey, D., Gribble, S. M., Yang, F., McLaren, K., Rogers, J., Ning, Z., Beare, D. M., et al. (2008). Finishing the finished human chromo- some 22 sequence. Genome Biology, Vol 9, R78.

[14] Sasson, SA (2010). From millions to one: Theoretical and concrete approaches to de novo assembly using short read DNA sequences. Ph.D. thesis, Graduate School-New Brunswick Rutgers, The State University of New Jersey.

[15] Chou HH, Holmes MH. (2001). DNA sequence quality trimming and vector removal. Bioinformatics, Vol 17(12), 1093–1104.

[16] Ramos RT, Carneiro AR, Baumbach J, Azevedo V, Schneider MP, Silva A. (2011). Analysis of quality raw data of second generation sequencers with Quality Assessment Software. BMC research notes, Vol 4, 130.

[17] Blankenberg, D, Gordon, A, Von Kuster, G, Coraor, N, Taylor, J, and Nekrutenko, A. (2010).

[18] Morgan M, Anders S, Lawrence M, Aboyoun P, Pagès H, Gentleman R. (2009). ShortRead: a bioconductor package for input, quality assessment and exploration of high- throughput sequence data. Bioinformatics, Vol 25, 2607–2608.

[19] Martínez-Alcántara, a, Ballesteros, E, Feng, C, Rojas, M, Koshinsky, H, Fofanov, VY, Havlak, P, Fofanov, Y (2009) PIQA: pipeline for Illumina G1 genome analyzer data quality assessment. Bioinformatics (Oxford, England), Vol 25, 2438–2439.

[20] Li H, Homer N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Briefings Bioinformatics, Vol 11, 181–197.

[21] Bryant DW, Wong W-K, Mockler TC. (2009). QSRA – a quality-value guided de novo short read assembler.BMC Bioinformatics 10–69.

[22] Jeck W, Reinhardt J, Baltrus D, Hickenbotham M, Magrini V, Mardis E, Dangl J, Jones C. (2007). Extending assembly of short DNA sequences to handle error. BMC Bioinformatics, Vol 23, 2942–2944.

[23] Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. (2008). RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Re- search, Vol 18, 1509–1517.

[24] Salmela L. (2010). Correction of sequencing errors in a mixed set of reads. Bioinformatics (Ox- ford, England, Vol 26, 1284–1290.

[25] Schroder J, Schroder H, Puglisi SJ, Sinha R, Schmidt B (2009) SHREC: a short-read error cor- rection method. Bioinformatics (Oxford, England), Vol 25, 2157–2163.

[26] Smith AD, Xuan Z, Zhang MQ. (2008). Using quality scores and longer reads improves accu- racy of Solexa read mapping. BMC bioinformatics, Vol 9, 128.

[27] Chaisson, MJ, and Pevzner, PA. (2008). Short read fragment assembly of bacterial genomes. Genome Research, Vol 18, 324–30.

[28] Miller, JR, Koren, S, Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. Genomics, Vol 95, 315-327.

[29] Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M. (2009). SHRiMP: Accurate Mapping of Short Color-space Reads. PLoS Computational Biology, Vol 5, 5.

[30] Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics, Vol 25, 1966–1967.

[31] Li H., Ruan J., Durbin R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Research, Vol 18, 1851–1858.

[32] Lin H, Zhang Z, Zhang MQ, Ma B, Li M. (2008) ZOOM! Zillions of oligos mapped. Bioinformatics, Vol 24, 2431–2437.

[33] Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I.A., Belmonte, M.K., Lander, E.S., Nusbaum, C. and Jaffe, D.B. (2008). ALLPATHS: de novo assembly of whole-genome shotgun microreads. Genome Research, Vol 18, 810 – 820.

[34] Chaisson M, Pevzner P, Tang H. (2004): Fragment assembly with short reads. Bioinformatics, Vol 20, 2067–2074

[35] Li Y, Hu Y, Bolund L, Wang J. (2010). State of the art de novo assembly of human genomes from massively parallel sequencing data. Hum Genomics, Vol 4(4), 271–277.

[36] Zerbino DR, Birney E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research. 18:821?829.

[37] Ewing B, Green P. (1998). Base-calling of automated sequencer traces using PHRED. II. Error probabilities. Genome Research, Vol 8(3), 186–194.

[38] Flicek P, and Birney E. (2009). Sense from sequence reads: methods for alignment and assem- bly. Nature Methods, Vol 6 (11 Suppl), pp. S6–S12.

[39] Pevzner, P. A. and Tang, H. (2001). Fragment assembly with doublebarreled data. Bioinformat- ics Vol 17, 225–233.

[40] Kingsford C, Schatz MC, Pop M. (2010). Assembly complexity of prokaryotic genomes using short reads. BMC Bioinformatics Vol 11(1), 21.

[41] Ferragina P, Manzini G. (2000). Opportunistic data structures with applications. Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000),, Redondo Beach,CA, USA, 390–8.

[42] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. (2004). Versatile and open software for comparing large genomes. Genome Biol, Vol 5, R12.

[43] Meek C, Patel JM, Kasetty S. (2003). OASIS: an online and accurate technique for local-alignment searches on biological sequences. In: Proceedings of 29th International Con- ference on Very Large Data Bases (VLDB 2003), Berlin. 2003, 910–921.

[44] Abouelhoda MI, Kurtz S, Ohlebusch E. (2004). Replacing suffix trees with enhanced suffix arrays. J DiscreteAlgorithms, Vol 2, 53 – 86.

[45] Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol, Vol Sep;5(9):e1000502.

[46] Langmead B, Trapnell C, Pop M, Salzberg SL. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol, Vol 10, R25.

[47] Li H, Durbin R. (2010). Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, Vol 26(5), 589–595.

[48] Li H, Durbin R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, Vol 25, 1754–1760.

[49] Lam TW, Sung WK, Tam SL, Wong CK, Yiu SM. (2008). Compressed indexing and local aligment of DNA. Bioinformatics, Vol 24, 791–7.

[50] Hernandez, D., François, P., Farinelli, L., Osterås, M. and Schrenzel, J. (2008). De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer. Genome Research, Vol 18, 802–809.

[51] Myers EW. (1995). Towards simplifying and accurately formulating fragment assembly. Jour- nal of Computational Biology, Vol 2.

[52] Huang, X. and Yang, S. (2005). Generating a genome assembly with PCAP. Curr Protoc Bioinformatics, Chapter 11, Unit11.3.

[53] Lemos M, Basílio A, Casanova A. (2003). Um Estudo dos Algoritmos de Montagem de Frag- mentos de DNA. PUC Rio, Rio de Janeiro.

[54] Fleishner, H. (1990). Eulerian Graphs and Related Topics. Elsevier Science, London.

[55] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Align- ment/Map format and SAMtools. Bioinformatics, Vol 16, 2078–2079.

[56] Simpson JT, Wong K, Jackman SD, Schein JE, Jones, SJM, Birol I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Research, Vol 19, 1117–1123.

[57] Medvedev, P., Georgiou, K., Myers, G., Brudno, M. (2007). Computability of models for sequence assembly. In Proceedings of Workshop on Algorithms in Bioinformatics WABI. 289?301.

[58] Warren RL, Sutton GG, Jones SJ, Holt RA. (2007). Assembling millions of short DNA sequences using SSAKE. Bioinformatics, Vol 15, 23(4).

[59] Dohm JC, Lottaz C, Borodina T, Himmelbauer H. (2007). SHARCGS, a fast and highly accu- rate short-read assembly algorithm for de novo genomic sequencing. Genome Research.

[60] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. Nature Biotechology, Vol 26, 1135-1145.

[61] Magi, A., Benelli, M., Gozzini, A., Girolami, F., Torricelli, F. and Brandi, M. L. (2010). Bioinfor- matics for Next Generation Sequencing Data. Genes, Vol 1, 294–307.

[62] Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, Marshall D. (2010). Tablet – next generation sequence assembly visualization. Bioinformatics, Vol 3, 401–402.

[63] Gordon D, Abajian C, Green P. (1998). Consed: a graphical tool for sequence finishing. Genome Research, Vol 8, 195–202.

[64] Schatz MC, Phillippy AM, Shneiderman B, Salzberg SL. (2007). Hawkeye: an interactive visual analytics tool for genome assemblies. Genome Biology, Vol 8, R34.

[65] Huang, W.; Marth, G. (2008). EagleView: a genome assembly viewer for next-generation sequencing technologies. Genome Research, Vol 9, 1538–1543.

[66] Bao, H.; Guo, H.; Wang, J.; Zhou, R.; Lu, X.; Shi, S. (2009). MapView: visualization of short reads alignment on a desktop computer. Bioinformatics, Vol 12, 1554 – 1555.

[67] Schuster SC. (2008). Next-generation sequencing transform today’s biology. Nature Methods, Vol5, 16-18.

[68] Pop, M., Kosack, DS., Salzberg, SL. (2004) Hierarchical scaffolding with Bambus. Genome research, Vol 14, 149–159.

[69] Setúbal JC and Werneck R. (2001). A program for building contig scaffolds in double-barrelled shotgun genome sequencing. Campinas Instituto de Computação, Unicamp.

[70] Tsai IJ, Otto DT, Berriman M. (2010). Improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biology, Vol 11, R41.

[71] Silva A, Schneider MP, Cerdeira L, Barbosa MS, Ramos RT, Carneiro AR, Santos R, Lima M, D’Afonseca V, Almeida SS, Santos AR, Soares SC, Pinto AC, Ali A, Dorella FA, Rocha F, de Abreu VA, Trost E, Tauch A, Shpigel N, Miyoshi A, Azevedo V. (2011). Complete Genome Sequence of *Corynebacterium pseudotuberculosis* I19, a Strain Iso- lated from a Cow in Israel with Bovine Mastitis. Journal Of Bacteriology, VOl 193(1), 323–324.

6.2.6 Anotação completa de genomas: análises *in silico*

A montagem completa de um genoma é um dos passos iniciais em estudos relacionados ao combate a microrganismos patogênicos. Não menos importante está a etapa complementar: a anotação desse genoma. Esta anotação fornecerá pistas a respeito, por exemplo, dos genes essências de um organismo sob estudo. Tanto a qualidade ou completude da montagem de um genoma, quanto de sua anotação, podem afetar as conclusões obtidas no organismo sob estudo.

Neste capítulo de livro discutimos programas de computadores com o intuito de mapear diferentes dados de um genoma tomando como modelo a *C. pseudotuberculosis.* São abordadas técnicas e ferramentas para predição gênica, anotação automática e curadoria manual, determinação do pangenoma, depósito de genomas em sítios públicos, análises de plasticidade genômica, regulação gênica e a busca por antígenos imunoprotetores.

Essa publicação encontra-se disponível na internet pelo endereço: http://www.intechopen.com/articles/show/title/whole-genome-annotation-in-silico-analysis.

Whole Genome Annotation: In Silico Analysis

Vasco Azevedo et al.*
Federal University of Minas Gerais (UFMG) and Federal University of Pará (UFPA), Brazil

1. Introduction

After a genome is assembled, the next step is genomic annotation, which can generate data that will allow various types of research of the model organism. Complete DNA sequences of the organism are then mapped in areas pertinent to the research objectives. In this chapter, we explore relevant ongoing research on genes and consider the gene as a basic mapping unit. Gene prediction is the first hurdle we come across to begin the extensive and intensive work demonstrated in first item, which deals with assembly of the genome. Gene prediction can be made with computational techniques for recognizing gene sequences, including stop codons and the initial portions of nucleotide sequences; it involves empirical rules concerning minimum coding sequences (CDS's) and is limited due to overlapping sequences coding forward and reverse.

Finishing gene prediction step by a computer initiates the functional annotation stage. Functional annotation, item 3, can be done initially by computer, using similarity in sequence alignment. However, no software is capable of generating a functional annotation without many false positive results, since conserved protein domains with varied functions make gene sequence alignment difficult. In this case, after automatic annotation, the predicted genes need to be revised manually. In manual curation, item 4, an expert can more accurately locate frameshifts in the DNA strand. Depending on the number of errors found, genomic annotation may be postponed, requiring a return to the previous stage of genome assembly. In manual curation, the principal contributions are usually correction of the start codon position, gene name, gene product and, finally, identification of frameshifts.

When functional annotation is completed, the genome should subsequently be submitted. It occurs after the assembly and annotation steps making the data generated available in public-access databanks. Submission is a pre-requisite for publication in scientific journals. Another advantage of genome publication in public-access sites is that it permits use of various genome analysis tools. For example, searches for genomic plasticity, pangenomic study, exported antigens and evaluation of innate and adaptive immune responses. The pangenome approach, item 5, concepts of species can be used as a filter for targeting candidates for vaccines, diagnostic kits and drug development. For drug development, the

* Vinicius Abreu, Sintia Almeida, Anderson Santos, Siomar Soares, Amjad Ali, Anne Pinto, Aryane Magalhães, Eudes Barbosa, Rommel Ramos, Louise Cerdeira, Adriana Carneiro, Paula Schneider, Artur Silva and Anderson Miyoshi
Federal University of Minas Gerais (UFMG) and Federal University of Pará (UFPA), Brazil

core set of proteins is a more likely source of useful information, for developing both vaccines and diagnostic materials for a unique pangenome set of a species of interest.

Genomic plasticity, item 6, is the dynamic property of genomes, involving DNA gains, losses, and rearrangement; it allows bacteria to adapt to new hosts and environments. There are several mechanisms that can drive these changes, including point mutations, gene conversions, rearrangements (inversion or translocation), deletions and DNA insertions from other organisms (through plasmids, bacteriophages, transposons, insertion elements and genomic islands). Gene acquisition and loss by all these mechanisms influences bacterial lifestyles and physiological versatility. Analyses of HGT regions in silico has become feasible due to the introduction of next–generation sequencing technologies, which allows sequencing of prokaryotic genomes at a faster rate than the earlier Sanger method and at a considerably lower operational cost. Consequently, the number of complete genome sequences available for analysis has grown and continues to grow rapidly.

In post-genomics, study of Reverse Vaccinology (RV), item 7, can provide predictions of the sub cellular locations of an entire predicted proteome. Additionally, these previous annotations, prediction of peptides with high affinity for class I and II MHC proteins is another in silico analysis that increases the probability of selecting antigens that can promote immune responses in organisms infected by a pathogen. The field of research referred to as immunoinformatics, item 8, is giving us the opportunity to analyze antigens with greater selectivity and increase the likelihood of developing a successful vaccine.

2. Gene prediction

The development of modern sequencing technology has resulted in an exponential increase in the number of available genome sequences. To illustrate, in 1997 there were 10 complete genome sequences of bacteria available in the NCBI (Lukashin & Borodovsky, 1998); by 2011, this number had sharply increased to 1,538 http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi. This enormous increase in the quantity of available information stimulated the development of tools for gene prediction. The development of these tools is a tremendous challenge, and it is a major contribution of Bioinformatics to the field of genomics.

2.1 Gene prediction strategies

Gene prediction programs can be divided into two categories: an empirical category, which relies on sequence similarity; and ab initio, which uses signal and content sensors. Empirical gene predictors search for similarity in the genome; they predict genes based on homologies with known databases, such as genomic DNA, cDNA, dbEST and proteins. This approach facilitates the identification of well–conserved exons. Ab initio gene finders use sequence information of signal and content sensors. Usually, these programs are based on Hidden Markov Models. Ab initio can be organized into categories based on the number of genome sequences used in gene analysis; it includes single, dual and multiple–genome predictors. Integrated approaches couple the extrinsic methodology of empirical gene–finders and intrinsic ab initio prediction. This technique significantly improves gene prediction protocols (Allen et al., 2004).

2.2 Eukaryotes

The complexity of the challenge faced by Bioinformatics is only completely understood when we look at the complexity of the eukaryotic genome. Within genomes, genes are not organized in a continuous cluster. Instead, the coding regions (exons) are often widely interspersed with non–coding intervening sequences (introns). Furthermore, in many cases the intronic region is much larger than the exonic region. These low–density coding sequences are evident in the human genome, in which only approximately 3% of the DNA generates proteins. The exon and intron issue can be compared to trying to read a non–continuous article in a journal. In an analogy, one must first identify in which part of the journal (genome) the article (gene) of interest is; then, as the DNA sequences are read, it is necessary to identify which part is informative (exon) and which part contains random information (intron). Also, genes can be altered by alternative splicing, which is a process that generates multiple protein sequences from the same gene sequence template (Schellenberg et al., 2008).

Gene prediction methodology for eukaryotes involves two distinct aspects; the first focuses on the information utilized for gene recognition, basically recognizing signal functions in the DNA strand; the second uses algorithms implemented by prediction programs for accurate prediction of gene structure and organization. The signal function search can be divided into two mechanisms utilized for locating genes. One classifies the content of the DNA strand and the other searches for functional signals in the genome:

(i) The content sensor classifies the DNA regions into coding and non-coding segments (introns, intergenic regions and untranslated regions). This mechanism involves two approaches, intrinsic and extrinsic. The extrinsic approach relies on the assumption that coding regions are evolutionarily more conserved than non–coding regions. Consequently, this methodology employs local alignment tools, like BLAST (Johnson et al., 2008) ; this makes it possible to make comparisons within the genome and between closely-related species. However, one important flaw in this approach involves the necessity of identifying homologies within the database in order to extract results. If none is found, this methodology is unable to determine if a region "codes" for a protein (Sleator, 2010). (ii) The functional sensor approach searches the genome for consensus sequences. Consensus sequences are extracted from multiple alignments of functionally-related documented sequences. The functional signals involve transcription, translation and splice sites. Transcriptional signals includes the CAP signal at the transcriptional start site and the polyadenylation signal located 20 to 30 bp downstream of the coding region. Another important signal to identify is the translation initiation site, although this feature has limitations due to a lack of knowledge concerning initiation sites in eukaryotes (Mathé et al., 2002).

2.3 Prokaryotes

Unlike eukaryotes, the archaeal, bacterial and virus genomes are highly gene-dense. The protein coding regions usually represent more than 90% of the genome. Therefore the accuracy of gene predictors depends primarily on determining which of the six frames contains the real gene. The simplest approach in gene prediction is to look for Open Reading Frames (ORFs). An ORF is a DNA sequence that initiates at a start codon and ends at a stop codon, with no other intervening stop codon. One way to locate genes is to look for ORFs with the mean size of proteins (roughly 900 base pairs) (Allen et al., 2004). Therefore, long ORFs indicate possible genes, although this methodology fails to predict small genes.

The major problem in simply applying this technique is the possibility of ORF overlap in the different DNA strains. This approach must be used along with guidelines to avoid overlapping, choosing the more likely candidates. Also, numerous false positives are found in non-coding regions. Due to the high gene density, it is difficult to confidently state that any gene predicted in a non–coding region is false. This problem can be minimized by searching for homologies in closely–related organisms. If we do not find a conserved sequence in related species, it is assumed that the prediction (of a gene) is false.

Another problem faced by prediction programs in prokaryotes is how to determine the start codon of a sequence. The first initiation site in a sequence is not necessarily the true one. To solve this problem, programs can employ ribosome binding sites (RBS), which provide a strong signal, indicating the position of the true start site. In conclusion, there is a drop in prediction accuracy in high–GC–content genomes. Rich GC genomes contain fewer stop codons and more spurious ORFs. These false ORFs are often chosen by prediction programs instead of the real ones in the same DNA region. Additionally, the longer ORFs in GC–rich genomes contain more potential start codons, leading to a drop in the accuracy of translation initiation site prediction (Hyatt et al., 2010).

2.4 Tools

2.4.1 Glimmer

The first version of Glimmer (Gene Locator and Interpolated Markov ModelER) was released in 1998 ; the 3.02 version was released in 2006. Glimmer is a system for finding genes in microbial DNA, especially the genomes of bacteria, archaea, and viruses. Glimmer uses interpolated Markov models (IMMs) to identify coding regions and distinguish them from noncoding DNA. Glimmer was the primary microbial gene finder used at The Institute for Genomic Research (TIGR), where it was first developed, and it has been used to annotate the complete genomes of over 100 bacterial species from TIGR and other labs. Like other gene prediction programs, Glimmer can be installed and run locally and has a web-based platform (Salzberg et al., 1998). All one needs for online gene prediction of a genome is the fasta version of the sequence and access to the site:
http://www.ncbi.nlm.nih.gov/genomes/MICROBES/glimmer_3.cgi.

2.4.2 FgenesB

FgenesB is a package developed by Softberry Inc. for automatic annotation of bacterial genomes. The gene prediction algorithm is based on Markov chain models of coding regions and translation and termination sites. The package includes options to work on sets of sequences, such as scaffolds of bacterial genomes or short sequencing reads extracted from bacterial communities. For community sequence annotation, it includes ABsplit program, which separates archebacterial and eubacterial sequences. FGENESB was used in the first published bacterial community annotation project (Tyson et al., 2004).

2.4.3 Prodigal

Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm) is a microbial (bacterial and archaeal) gene finding program developed at Oak Ridge National Laboratory and the University of Tennessee. Prodigal focuses specifically on three goals: improved gene structure prediction, improved translation initiation site recognition, and reduced false positives (Hyatt et al., 2010). The source code is freely available under the General Public License and the program can be accessed at http://compbio.ornl.gov/prodigal/.

2.4.4 GeneMarkTM

GeneMark is a public access program for gene prediction in eukaryotes. It is a family of gene prediction programs developed at Georgia Institute of Technology, Atlanta, Georgia, USA. GeneMark can operate in two ways: the first one is online, where one can make predictions, using for comparison one of the many available models; the second option is for novel genomes, in this way one can install and run the program locally. The web–based version of GeneMark is available at http://exon.biology.gatech.edu/.

For gene prediction in eukaryotes, GeneMark combines two programs, GeneMark–E* and GeneMark.hmm–E. The GeneMark-E program determines the protein-coding potential of a DNA sequence (within a sliding window) by using species-specific parameters of the Markov models of coding and non-coding regions. This approach allows delineating local variations with coding potential. The GeneMark graph shows details of the protein-coding potential distribution along a sequence, while the GeneMark.hmm-E program predicts genes and intergenic regions in a sequence as a whole. The Hidden Markov models take advantage of the "grammar" of gene organization. The GeneMark.hmm programs identify the most likely parse of the whole DNA sequence into protein coding genes (with possible introns) and intergenic regions.

The statistical model employed in the GeneMark.hmm algorithm is a hidden Markov model. It includes hidden states for initial, internal and terminal exons, introns, intergenic regions and single exon genes. It also includes hidden states for start site (initiation site), stop site (termination site), and donor and acceptor splice sites. The protein-coding states (initial, internal, terminal exons and single exon genes) emit nucleotide sequences modeled by inhomogeneous 3–periodic fifth–order Markov chains. The non-coding states (intron and intergenic regions) emit sequences modeled by homogeneous Markov chains (Lukashin & Borodovsky, 1998).

3. Automated functional annotation

Automated functional annotation of genomes can be quite efficient because it is a computational process based on the alignment of ORF sequences of the organism with sequences from various other organisms (Kislyuk et al., 2010). Public domain databases contain full annotations of thousands of prokaryotic organisms (Benson et al., 2008). Automatic functional annotation takes advantage of knowledge concerning ORFs of homologous organisms, saving considerable time in manual curation (Li et al., 2010). However, care must be taken with fully automated functional annotation, since similarity of sequences can easily incur false positives (Lorenzi et al., 2010). In this section we discuss the advantages and dangers of using fully-automated functional annotation, and we explore some features of tools and services for this purpose.

3.1 Massive sequence alignments must be planned

Algorithms for alignment of biological sequences are intensively used in automatic functional annotation (Aparicio et al., 2006; Meyer et al., 2003). Alignments of ORFs from a newly assembled genome with counterpart ORFs can provide the first hints about the new genome. For an organism with about 2,000 ORFs, analysis of similar sequences against a database of non-redundant (NR) proteins from NCBI can consume several processing hours. For example, assuming that this analysis is done on a computer isolated from the internet, hardware with 24 Gb RAM and eight processors, totaling 24 GHz CPU, this task will consume approximately eight hours of processing time.

Though it is a completely automated computer process, the user has considerable responsibility to set conditions to be utilized in the computation in order to obtain good quality data. These conditions define the quality criteria that best fit the type of organism, for example, the cut–off value for a significant alignment with sequences of other organisms in the NCBI, the number of homologous sequences to be returned as a result and the file format of the output alignment. An additional parameter is required if the sequence search (query) and the targeted search sequences (subject) are in different formats (nucleotides versus amino acids). This parameter determines the most adequate table for translation of codons of the organism in question so that the alignment algorithm of sequences is able to interpret the correspondence between the query sequences and the subject. The number of parameters of an algorithm for aligning sequences can be quite large, justifying training with a heavy workload for optimal utilization. Our objective here is not to explore possible situations, but to alert users that the results of these algorithms can improve these alignments by reading the manual algorithm and consequently adjusting it to a particular situation concerning a query organism or subject. Thus, when beginning a massive alignment sequences project involving a novel genome, with an analysis that will take hours and create high expectations, it is advisable to use not just the basic configurations in these alignment algorithms. It would be useful to take time to weigh and incorporate options that will determine the success or failure of these alignments.

3.2 Knowledge reapplication and time saving

There has been significant growth in the number of DNA sequences available in public databases, because of new genome sequencing technologies, which have made it simpler, more efficient and cheaper to obtain complete genomes (Zhao & Grant, 2010). Fully assembled and annotated genomes of various forms of bacterial life are available to facilitate the processing and inclusion of a newly assembled genome. This wide range of genomes provides the opportunity for new research into large-scale SNPs, DNA methylation and mRNA expression profiles, and resequencing data (Datta et al., 2010). It also allows comparison of annotations from different research groups working with different organisms, some of which may be homologous to a newly-sequenced genome. Just as one can take advantage of knowledge about the function of genes from different organisms, it is also advisable to use the personal knowledge of a researcher on a specific organism in order to accelerate the process of automatic annotation. Based on evidence about a high degree of evolutionary proximity between a newly-assembled genome and a particular organism homolog that already has a fully-assembled and annotated genome, we can choose to use only the annotation of such an organism as a resource for a first automatic annotation.

The problems a researcher would normally encounter when utilizing annotations from various genomes could be resolved by comparison with the annotation of a homologous organism. This situation is common when one examines the pangenome of a species, as it is expected that most of the coding sequences of different strains of bacteria are not very different (Trost et al., 2010). In this case, it appears to be advantageous to identify a small set of target organisms (subject) in a sequence similarity search, with the objective of providing a first genome annotation (query); this may even be a set with only one organism.

3.3 Error propagation: Automated versus manual annotation

It is important to bear in mind that the GenBank is not a fully curated database (Benson et al., 2008); many genomes may have been deposited only as automatic annotations. With current technology, it is not possible to dispense with manual curation of an automatic annotation, or even experimental evidence concerning gene prediction and annotation based on sequence similarities (Poptsova & Gogarten, 2010). Although it is not normally feasible to initially include experimental verification of gene prediction, it seems reasonable to take advantage of expert human annotation of genomes to help determine the outcome of automatic annotation. Assuming one is working on the pangenome of an organism, such a measure can not only reduce false positives in comparisons of sequence similarities, but also determination of homologous genomes based on a particular annotation. During automatic annotation, a measure that has the potential to minimize error propagation would be allocating different weights for the results of sequence similarity to genes from organisms for which there is evidence of expert manual curation.

3.4 Tools

The following are some tools for automatic annotation of entire genomes, with brief descriptions of their core functionality and instructions on how to use them.

3.4.1 GenDB

One of the reasons that GenDB is included among a select set of tools for automatic annotation of genomes is the fact that it was developed for the web platform (Meyer et al., 2003). Geographically dispersed research groups can benefit from web interfaces using standard tools and a centralized database. Version 2.4 of GenDB has three modules: core, web, and gui. The core module has programs written in perl that allow creation of an annotation project, importation of data in fasta / EMBL format, execution pipeline automatic annotation, display of circular genomic maps, data export and annotation project deletion. Implementation of the programs in the module allows a team of curators to work on the web and edit diverse features of various genes. The gui module has editing features that are more sophisticated than those of the web module, allowing execution of tasks performed by the core module, but with a graphical interface. The GenDB program performs sequence alignments using the program Blast (Altschul et al., 1997) and allows incorporation of predictions of conserved domains of protein families based on InterPro-Scan (Hunter et al., 2009),as well as transmembrane domains based on TMHMM (Krogh et al., 2001), and indications of export to the extracellular medium through SignalP (Bendtsen et al., 2004).

3.4.2 BLAST2GO (B2G)

This tool was designed as an interface for Gene Ontology (GO); additional features have transformed it into a more comprehensive annotation platform (Aparicio et al., 2006). The program menus include various steps initiating annotation, with an automatic alignment of genome sequences against a protein-based non-redundant (NR) NCBI database, through prediction of conserved domains (InterPro–Scan), GO annotation ratings against the enzymatic English Enzymatic Code (EC) and subsequent visualization of molecular interactions in a genome by means of maps in the format of the Kyoto Encyclopedia of Genes and Genomes (KEEGO). Being a visually oriented tool, it has graphical tools to help analyze the vast amount of data generated in the predictions. A user of B2G does not necessarily have to perform all the steps of analysis that are offered, but in order to advance to the next phase of analysis it is imperative that the previous phase be performed beforehand. Processing of an entire genome with approximately two thousand ORFs can take several days, as the first step is always sequence alignment against the NCBI NR base. Fortunately, B2G is designed to be a modular analysis tool. If a B2G user has computational resources that are more efficient than the shared resources on the public server, the user can perform alignment of sequences on his own hardware to generate an output in HTML format and continue the alignment processes following annotation with B2G. Should the user be dissatisfied with the efficiency of processes of annotating GO terms of the server's common B2G, there is a version of B2G than he can run separately with his superior hardware. The results generated in the offline mode can be uploaded to the online tool to continue the review process using a variety of tools, including statistical comparisons between two genomes. B2G was developed with Sun Java technology, which can be run on any operating system; however, the B2G offline module is designed to run on the Linux platform.

3.4.3 CpDB relational schema: a practical example

This tutorial has approximately 100 steps, including software installation and configuration, edition of files by Linux commands or through interfaces with biological sequence manipulation programs. The tutorial presumes that the programs Artemis, Java (Sun) and Blast version 2.2.20 or previous were locally installed. Many editions of files are made with the "sed" program of Linux, which is included in most Linux versions. All of the steps in this manual can be automated in order to develop an automatic pipeline for annotation, allowing the *Corynebacterium pseudotuberculosis* DataBase (CpDB), a relational database schema and tools for bacterial genomes annotation and pos-genome research, to become another web-based automatic annotation environment. For now, this tutorial has an instructional character, to help make a student aware of the necessities and difficulties involved in the process of automatic annotation of genomes. In order to obtain the tutorial files, type the following command in Linux, Ubuntu 10.10 or later:

```
svn checkout svn://150.164.37.20/genomes/autoannotation --username=student --password=bioinfo
```

After finalizing the verification of all of the files, this tutorial continues in the document "Tutorial.pdf", which will be in the folder "autoannotation".

4. Manual curation

Genome annotation is a process that consists of adding analyses and biological interpretations to DNA sequence information. This process can be divided (Stein, 2001), into three main categories: annotation of nucleotides, proteins and processes. Annotation of nucleotides can be done when there is information about the complete genome (or DNA segments) of an organism. It involves looking for the physical location (position on the chromosome) of each part of the sequence and discovering the location of the genes, RNAs, repeat elements, etc. In the annotation of proteins, which is done when there is information about the genes (obtained by genome or cDNA sequencing) of an organism, there is a search for gene function. Besides general predictions about gene and protein function, other information can be found in an annotation, such as biochemical and structural properties of a protein, prediction of operons, gene ontology, evolutionary relationships and metabolic cycles (Stothard & Wishart, 2006). Consequently, functional annotation or manual curation is a fundamental part of the process of assembling and annotating a genome, in which the curator is the person responsible for validating the elements. In manual curation, all of the predicted genes will be validated and their products named (Stein, 2001). A more detailed description of the gene or gene family product is obtained through similarity analyses using protein data banks that contain well-characterized and conserved proteins (Overbeek et al., 2005).

4.1 Technical terms used in manual annotation

In functional annotation done with Artemis, several fields should be filled out to increase knowledge about particular genome elements. It is necessary to use annotation terms, which involve an official nomenclature developed for this purpose. Some of these terms and respective examples are given below: "LOCUS-TAG" is the term used to identify all of the genome elements, except for the feature "misc". Generally, one uses an abbreviation to identify the particular species, followed by an underline (_) and numbers, for example: Cp1002_0001 (*Corynebacterium pseudotuberculosis*, strain 1002). For tRNAs, the nomenclature is the abbreviation, followed by underline, a "t" and numbers, with a specific count, which is not included in the total CDS count, among others; for example: Cp1002_t001. For rRNAs, the nomenclature is the symbol followed by underline, an "r" and numbers, with specific counts, not included in the total CDS count; for example: Cp1002_r001. "PROTEIN_ID" is used to designate all of the elements of the genome, except for the feature "misc". It is a standardized form for NCBI to identify e proteins; for example: gnl|gbufpa|Cp1002_0001. "GENE" is one of the most important topics to be informed in manual annotation, indicating the gene symbol of the protein; fore example: pld. The field "SIMILARITY" corresponds to information obtained from the best similarity search result – BLASTp. Various types of information should be entered into this field, such as similarity among organisms, size of the amino-acids sequence analyzed, e-value and also the percentage identity between its own protein and the protein found in the data bank; for example: similar to *Corynebacterium pseudotuberculosis* 1002, hypothetical protein Cp1002_00047 (345 aa), e–value: 0.0, 98% ID in 344 aa. In "PRODUCT", there is a description of the gene product, for which similarity was found in the public domain data bank; for example: Phospholipase D. The tag "PSEUDO" should be added whenever a protein presents one or various breaks, due to insertion of a premature stop codon. These are the famous proteins that have frameshifts or probable pseudogenes. Consequently, the manual annotation window has this pattern:

```
/gene="dnaA"
/product="Chromosomal replication initiation protein"
/locus_tag="Cp1002_0001"
/protein_id="gnl|ufmg|Cp1002_0001"
/colour=3
/similarity="Similar to Corynebacterium pseudotuberculosis FRC41,
Chromosomal replication initiation protein (603 aa), e value: 0.0, 98% id in 599s aa"
```

4.2 Steps for manual curation

Manual curation is a very complex task and is subject to errors for various reasons. One of these is a lack of padronization in the interpretation of BLAST results. Another problem is propagation of errors, which involves prediction of protein function based on proteins that were also predicted but could have imprecise or even incorrect annotation (Gilks et al., 2002). For these reasons, some criteria are suggested in order to obtain reliable functional annotation. The fundamental step for doing this well is mining data obtained from similarity analyses of BLASTp data banks. It is recommended to give greater value to annotation of proteins of individuals of the same species or of species that are phylogenetically close to the organism under study, the protein of which one wants to infer the function of, decreasing in this way the possibility of annotation errors. Another parameter is to observe if there is any consensus among the first 10 hits (the same protein is identified among various). In this case, even if the best hit is not identified as such, it is preferable to identify the sequence as similar to that of an organism that appears various times in the BLASTp results and is within the consensus. In cases where there is no consensus or when the e-value of the best hit (first BLAST result and which corresponds to the best alignment within the data bank that is being researched) is significantly larger than that of the following sequences, it is preferable to transfer the annotation of the best hit (Prosdocimi, 2003), or if necessary, in cases of non-significant alignments, always also run a similarity search at the nucleotide level (BLASTn). Other criteria are also analyzed, such as percentage identity between the sequence being analyzed and the sequence in the data bank, score value and e-value, as well as pair-by-pair alignment evaluation. This evaluation consists of checking the texture of the alignment (evaluating the number of gaps, size of the gaps, and the number of conserved substitutions of amino acids). If doubts remain, research of domain data banks and protein classification are also commonly utilized.

4.3 Frame shifts (Pseudogenes)

Comparisons between non-coding regions of genomes of various prokaryotic species has aided in the identification and characterization of genome segments with regulatory roles (Pareja et al., 2006), contributing to the elucidation of genetic circuits of transcriptional regulation. These non-coding regions, known as pseudogenes, are DNA sequences that are highly similar to functional genes but do not express a functional protein, probably because of deleterious mutations. These degraded genes contain one or more inactivating mutations, such as a nonsense mutation that introduces a premature stop codon, resulting in an incomplete protein and a later change in the open reading frame (Lerat & Ochman, 2005). When found in the genome, the break region is checked with Artemis, and the quality of the bases in that region is also evaluated. Whenever possible, addition or removal of erroneous bases can restore the reading frame. If there is no data that justifies addition or removal of bases, the genes should be classified as pseudogenes (tag /pseudo).

4.4 Tools

4.4.1 Artemis

The program Artemis, (Berriman & Rutherford, 2003), available for download at http://www.sanger.ac.uk/Software/Artemis is a freely-distributed algorithm developed for visualization of genomes and for annotation and manual curation. Artemis allows the curator to visualize various characteristics of the genome sequences, such as: product coded by the predicted gene; presence of tRNAs and rRNAs; search for protein and nucleotide similarity in biological data banks; visualization of probable domains and conserved protein families; visualization of GC / AT content, and misplaced codon use; and various other functions. These data can be visualized in the six phases of translating DNA reads into proteins (Rutherford et al., 2000). Also, the program provides a visualization of BLAST visits between two complete genome sequences, allowing rapid analysis of the degree of synteny (conservation at the level of genes), the main genomic rearrangements and integration of new genomic islands (Field et al., 2005). This algorithm is written in the Java language and is available for the following operating systems: UNIX, Macintosh and Windows. Artemis is capable of processing data in the formats EMBL and GENBANK, or even sequences in the format FASTA.

4.5 Sequence similarity searches

4.5.1 BLAST (Basic Local Alignment Search Tool)

BLAST (Altschul et al., 1990) is a tool that is widely used for the characterization of products coded by genes that are identified by gene prediction. It is able to identify a great majority of the alignments that attend the desired criteria, with a significant gain in performance (Gibas & Jambeck, 2001). This program is available on the NCBI - National Center for Biotechnology Information site `http://www.ncbi.nlm.nih.gov` (Stein, 2001), which is considered the central databank for genome information. As shown in the figure, BLAST has programs for alignment of protein and nucleotide sequences, among others, according to the needs of the work that is to be undertaken:

Program	Entry sequence	Type of sequence target
BLASTp	Protein	Protein
BLASTn	Nucleotide	Nucleotide
BLASTx	Translated nucleotide	Protein
TBLASTn	Protein	Translated nucleotide
TBLASTx	Translated nucleotide	Translated nucleotide

Table 1. Types of BLAST – NCBI programs.

Through this type of algorithm, we can compare any DNA sequence or protein (query) with all of the genome sequences in the public domain (subject) (Altschul et al., 1997). It is important to note that the program BLAST does not try to make a comparison of the full extension of the molecules that are being compared, but rather it identifies in the data bank a sequence that is sufficiently similar to that of the sequence that is being studied.

4.5.2 Interpreting blast results

In the manual annotation of genomes, analysis of BLAST parameters, such as the number of points obtained (score), gap opening/extension penalties, number of expected alignments in the case of scores equal to or superior to the alignment that is being investigated (expectation value), and the normalized score (bitscore), are indispensible for the interpretation of the results. The smaller the value of "E", the smaller the chance of such a comparison being found merely by chance, consequently inferring a greater homology between the sequence being investigated and the data base (Baxevanis & Ouellette, 2001). Among the sequences with identity above 50%, a general approach is to characterize the function of the known sequence and transfer this annotation to the new sequence. Though annotation transfer is a common practice, a high rate of error has been reported when this is done without due caution (Liberman, 2004). Based on this principle, we consider that for sequences with identity above 80%, a simple alignment or a comparison with a protein that has been experimentally characterized using BLAST can be sufficient to infer function, as long as the pair being compared has similar lengths and align end to end without large deletions or insertions. For pairs with identity in the range of 50–80%, the general approach for attributing function includes evaluation of databanks with homologous protein and protein domain families.

4.5.3 PFAM

Proteins generally are composed of one or more functional regions, or domains. Different combinations of domains result in the large variety of proteins found in nature. Identification of the domains that are found in proteins can, therefore, provide insight about protein function (Sanger Institute, 2009). In sequences with an identity of less than 70%, without end to end similarity, the approach that is used is to evaluate the protein domains through a search of the Pfam database, which gives very extensive coverage (Mazumder & Vasudevan, 2008). The Pfam database is accessible via the Web http://pfam.sanger.ac.uk and is available in various formats for download. This databank is contains two complementary groupings; Pfam–A is composed of high–quality protein domains that have been manually verified, while Pfam–B contains data that has been generated automatically from the ProDom databank (Finn et al., 2010). Pfam–B is generally lower in quality, though it can suggest new domains that can be added to the manual annotation, if they are not available in Pfam–A. Basically, in Pfam, the sequences that are in full alignment are identified through a search for a hidden profile using the algorithm Hidden Markov Model (HMM), which is later generated using the software HMMER, based on the UniProt database (UniProt, 2007). These HMMs are statistical models that capture specific information about how much each alignment column is conserved and indicates the residuals in this evaluation.

5. Genomics

A genome is the complete set of DNA sequences of a living organism; it consists of coding and non–coding sequences. Genomics is a discipline of genetics that deals with genomes or DNA sequences. Simply put, genomics is the study of genomes. Computational genomics derives knowledge from genome sequences and related data, including both DNA and RNA sequences as well as experimental data. Computational biology mainly deals with whole genome analysis to understand the DNA mechanisms and molecular biology of a species. As biological datasets are extremely large, computational biology has become an important part of modern biology.

5.1 Pangenomics

The efficient and low cost sequencing technologies that are currently available provide complete genome sequences of pathogenic, industrially useful, and other economically-important organisms. Genome sequences, and information that is coded in these sequences, can help identify pathogenicity and other important genes.

Complete genomic sequences of various strains of a species are important to help us understand pathogenesis mechanisms and to determine how genetic variability affects pathogenesis; it would be difficult to extract such useful information from a single genome (Lefébure & Stanhope, 2007).

A pangenome consists of a "core genome", which contains the gene or sequences present in all strains. In other words, genes that are found in all the genomes in a species of bacteria are called the core genome. A "dispensable genome or accessory genome" consists of genome sequences present in more than two strains but are not part of the core genome. "Unique genomic sequences" or "unique genes" are strain-specific genes. These genes are limited to single strain. The pangenome is important for identification and for designing effective vaccines and drug targets (Mira et al., 2010).

There are many web tools and softwares available to manage and efficiently extract data from genomes of various strains of the same species. These tools recognize the accession numbers allotted to complete genomes submitted to NCBI and to other databanks. Online tools developed by the Computational Genomics group of Bielefeld University, Germany, EDGAR – "Efficient Database framework for comparative Genome Analyses using BLAST score Ratios" http://edgar.cebitec.uni-bielefeld.de are efficient web tools to determine the core genome, along with dispensable and unique genes in the form of colored graphs and tables (Blom et al., 2009)

For example, we analyzed the core genome, dispensable genes and unique genes, using "EDGAR", of three different *Corynebacterium pseudotuberculosis* strains, *C. pseudotuberculosis* Cp–I19, *C. pseudotuberculosis* Cp1002 and *C. pseudotuberculosis* CpC231.

This core genome consists of 1,862 genes, with 48 dispensable genes between Cp–I19 and Cp1002, 52 dispensable genes between Cp-I19 and CpC231, and 103 dispensable genes between Cp1002 and CpC231. There were 208, 46 and 36 unique genes in strains Cp-I19, Cp1002 and CpC231, respectively.

6. Genome plasticity

The high degree of adaptability of bacteria to a wide range of environments and hosts is long known to be influenced by genome plasticity, a dynamic property that involves DNA gain, loss and rearrangement (Maurelli et al., 1998). Various mechanisms can drive these changes, including point mutations, gene conversions, rearrangements (inversion or translocation), deletions and DNA insertions from other organisms (plasmids, bacteriophages, transposons, insertion elements and genomic islands) (Schmidt & Hensel, 2004).

6.1 Plasmids

Plasmids contribute to genomic plasticity through their transfer capability. They are also able to mobilize co-resident plasmids and integrate into the chromosome. Plasmids may harbor antibiotic resistance genes and other genes associated with pathogenicity (Dobrindt & Hacker, 2001); e.g., Rhodococcus equi harbors a virulence plasmid that codes for surface-associated proteins (vap genes) that is absent in avirulent strains (Takai et al., 2000).

6.2 Bacteriophages

Bacteriophages are viruses that infect bacteria and which influence genome plasticity through transduction mechanisms. Functional phages inject DNA from one bacterium into another one without causing damage to the acceptor organism; the DNA can incorporate into the acceptor genome leading to adaptive changes. Additionally, prophages (viral DNA incorporated in the bacterial chromosome) confer protection against lytic infections and they can harbor virulence genes that may be acquired by the acceptor bacterium and directly affect its pathogenicity; this has been reported from various species, including Clostridium botulinum, Streptococcus pyogenes, Staphylococcus aureus, Escherichia coli and C. diphtheriae (Brüssow et al., 2004).

6.3 Genomic islands

Genomic islands (GEIs) affect genome plasticity because of their mobility and their capability of carrying a large number of genes as a single block, including operons and groups of coding genes with related functions. These GEIs can cause dramatic changes that lead the acceptor bacterium to evolve very rapidly compared to wild-type counterparts. GEIs are characterized as large DNA regions acquired from other organisms. They vary in size (10-200 kb), and can harbor sequences derived from phages and/or plasmids, including integrase genes; GEIs are flanked by tRNA genes or direct repeats, which help produce their characteristic instability (Hacker & Carniel, 2001). The instability of GEIs is exemplified by rapid gene acquisition and/or loss and changes in gene composition, as seen in different strains of Burkholderia pseudomallei (Tumapa et al., 2008). Additionally, GEIs can be classified into several classes according to gene content. These include Symbiotic Islands, which are involved in the association of bacterium with Leguminosae hosts (Barcellos et al., 2007); Resistance Islands, which harbor genes related to antibiotic resistance (Krizova & Nemec, 2010); Metabolic Islands, which contain genes associated with secondary metabolite biosynthesis (Tumapa et al., 2008); and Pathogenicity Islands (PAIs), which have a high concentration of virulence genes. PAIs are associated with pathogenic bacteria and have been implicated in the reemergence of various pathogens as causes of serious disease problems (Dobrindt et al., 2000). The first description of a PAI was made in 1990, in vitro (Hacker et al., 1990),. The identification was based on the observation of a close relation between deletion of hemolysin and fimbrial adhesin coding regions and non pathogenic strains of E. coli. This was investigated by gene cloning technique, pulse field electrophoresis and Southern hybridization. Using these procedures, they showed that the hemolysin and fimbrial adhesin coding genes are located in the same chromosomal region in several wild-type strains of E. coli and that they go through deletion events both in vivo and in vitro (Hacker et al., 1990).

6.4 "Black Holes"

Additionally, it is important to keep in mind that gene deletion is just as important as gene acquirement in some organisms. One example of this event is the so called "Black Holes" or deletion events of "antivirulence" genes, i.e. genes whose expression in pathogenic organisms is incompatible with virulence. The concept of evolution through deletion of "antivirulence" genes is based on the premise that genes required for adaptation of one organism in a specific niche may inhibit adaptability in another niche, a potential host, for example (Maurelli, 2007). In E. coli, loss of cadA, the lysine decarboxylase (LDC) coding gene, and ompT, which synthesizes an outer membrane protein, may confer virulence (Suzuki & Sasakawa, 2001). The mechanism of action of cadaverine, produced by decarboxylation of lysine by LDC, is still unknown. However, there are two hypotheses: cadaverine inactivates the synthesized toxin, or cadaverine acts directly on the target cell to protect it. Maurelli et al. (1998) demonstrated that when rabbit mucous cells are pre-treated with cadaverine and then washed, they are protected from enteroxin effects. Absence of Omp-T in Shigella strains and enteroinvasive E. coli strains is crucial for maintaining VirG on the cell surface, a pre-requisite for mobility on mammal cells, including bacterial dispersion through epithelial cells (Suzuki & Sasakawa, 2001).

6.5 Software to identify horizontal gene transfer (HGT) events

Gene acquisition and loss through HGT influence bacterial lifestyles and their physiological versatility (Dobrindt & Hacker, 2001). The increasing number of complete genome sequences available for analysis has stimulated in silico research in an effort to identify HGT events. Horizontally–acquired regions can be identified based on observation G+C content and codon usage patterns, which differ among species and species groups. Sets of genes acquired by HGT events show deviations in these patterns that reflect the genomic signature of the donor genome (Langille et al., 2008). Various softwares can be used to identify HGT events based on base composition patterns (wavelet analysis of G+C content, cumulative GC profile, P–web, IVOM, IslandPath and PAI–IDA) and codon usage deviation (SIGI–HMM and PAI–IDA). However, due to adaptations in codon usage (Karlin et al., 1998), which tend towards homogenous base composition distributions (Hershberg & Petrov, 2009), identification of mobile regions based on genomic signature is only possible for regions that have recently been acquired from phylogenetically distant organisms, i.e. those that have a discrepant genomic signature when compared to the acceptor genome.

Additionally, identification of HGT events may be aided by concentrating on regions that are flanked by tRNA genes, which are "hot spots" for transfer elements since they possess 3'–terminal insertion sequences that are recognized by various integrases (Hou, 1999). The integration of PAIs into these insertion sequences is responsible for their instability, since a single integrase may cause excision of the entire region. Insertion/deletion events have been demonstrated in PAIs I and II of E. coli strain 536, which are flanked by selC and leuX tRNA genes (Blum et al., 1994), and in high pathogenicity islands (HPIs) of several *Yersinia pseudotuberculosis* and *Y. pestis* strains (Lesic et al., 2004), which frequently insert into ASN3 tRNA genes.

However, although efficient in the identification of HGT events, approaches based on genomic signature and flanking tRNAs are not aimed at classification of GEIs, since they do not consider the overall gene content of the region. Additionally, horizontally acquired regions may deviate only in G+C content or codon usage alone, which would be a problem for the identification process if only one of these features is used to identify the event. However, there are tools designed to identify a specific class of GEIs, pathogenicity islands, through a multi-pronged strategy that overcomes such constraints. These tools are named PredictBias (Pundhir et al., 2008), IslandViewer (Waack et al., 2006) and PIPS (unpublished); they perform analyses based on genomic signature deviations that are not found in closely-related organisms and finding of genes coding for virulence factors. Although all of these programs use similar strategies and are complementary, PIPS deserves special attention since it surpasses the others in accuracy and is easy to install.

In analysis of C. diphtheriae strain NCTC 13129, PIPS outperformed the other approaches, identifying 12 out of the 13 PAIs of the reference strain, compared to 10 by IslandViewer and six by PredictBias. In the identification of PAIs of uropathogenic E. coli strain CFT073, PIPS had an overall accuracy of 93.9% (unpublished) against 89.5% for IslandViewer and 88.1% for PredictBias.

7. Reverse vaccinology

Reverse Vaccinology (RV) (Rappuoli, 2000) starts from the genomic sequence of a pathogen, which is an expected coded sequence for all the possible genes expressed during the life cycle of the pathogen. All open reading frames (ORF's) derived from the genome sequence can be evaluated with a computer program in order to determine their aptitude as vaccine candidates. Special attention is given to exported proteins because they are essential in host-pathogen interactions. Examples of this interaction can be cited: (i) adherence to host cells, (ii) invasion of compliant cells, (iii) damage to host tissues, (iv) resistance to environmental stress by the machinery defense of the cell being infected and finally, (v) mechanisms for subversion of the host immune response (Sibbald & van Dij, 2009). The word "Reverse" in RV can be explained by the reverse genetics (RG) technique. Before the dawn of genomics, there were attempts to discover the genes responsible for each phenotype. With Crick's central dogma (DNA > RNA > Protein) the research path was reversed. In possession of the likely gene sequence, several techniques have been developed to identify changes in the phenotype of an organism derived from sequence changes in genes. The principle of Crick's dogma is also used by RV; when a gene sequence is found, one can determine whether a probable protein encoded by this sequence can be an antigen capable of stimulating an immune response in a host organism.

Long before the creation of the term RV, a number of approaches had been considered to determine exported proteins in order to move to the next step of the production of a subunit vaccine (Diaz Romero & Outschoorn, 1994). For example, research on exported proteins was advanced as an alternative to subunit vaccines based on the polysaccharide capsule of meningococci. Vaccines produced with such antigens had a low capacity to induce a satisfactory immune response. This research effort on exported proteins includes almost two decades of work searching for a vaccine against meningococcal serogroup B, which now gives good results. This vaccine currently is the best RV alternative for the production of a subunit vaccine for Neisseria meningitidis serogroup B. Meningitis caused by serogroup B (Men B) is responsible for approximately half of the worldwide incidence of this disease (Diaz Romero & Outschoorn, 1994), and this research result for targeted vaccination is commonly used as a demonstration of the usefulness of RV, because of the excellent results.

Currently, a subunit vaccine against Men B created with antigens targeted by RV is being tested in phase-2 clinical trials (Bambini & Rappuoli, 2009). The advantages of RV continue to be attractive, enabling vaccine research for organisms whose cultivation in the laboratory is difficult or impossible. Reducing the time needed to select target proteins could allow investigation of different species or strains at the same time, for selecting vaccine candidates that can elicit adaptive immune responses. To achieve these benefits all we need is to have a sequenced genome, a personal computer and core software widely available to the scientific community. These conditions demonstrate another advantage of using RV, the low cost. What we call core software is a set of tools for identifying well-known motifs, such as, for example, SignalP, TMHMM, LipoP, and HMMSEARCH. There is still room for innovation in the use of core software; the choice of software strategies can be directed to the identification of vaccine candidates specific to an organism, such as in the case of gram-negative (bilayer) or gram positive (monolayer) bacteria, or also according to heuristics for selection of vaccine candidates with specific characteristics. For example, membrane or exported to the extracellular environment (Barinov et al., 2009).

The concept of RV was adapted to fit a new reality of widespread availability of genomic data (Rinaudo et al., 2009). Instead of researching vaccine targets for a single strain or subspecies of an organism, we can do it simultaneously in dozens of genomes, exploring potential joint antigens or those exclusive to multiple genomes (Lapierre & Gogarten, 2009). The possibility of having a large number of genomes available to implement RV leads to the emergence of the concept of pangenomics RV (PGRV) (Bambini & Rappuoli, 2009). PGRV can also apply the concepts of core, extended, and character genomes. The core genome in PGRV is composed of exported genes (genes that transcribe exported proteins) that are common to all strains, genes that could be candidates for a universal vaccine, while the extended genome consists of genes that are absent in at least one of the strains of the studied species, while the character genome consists of genes that are specific to a strain (Lapierre & Gogarten, 2009). From the standpoint of vaccines, the core and character genomes would be good candidates to develop a vaccine that is suitable for all strains, without losing sight of the particularities of specific genes in each strain.

8. Immunoinformatics

The immune system has considerable diversity in its components, such as, for example, immunoglobulin receptors of lymphocytes, or cytokines, with the principle cell types being B- and T-cells, which have important roles in inflammation, infection and protection (Evans, 2008). Immunoinformatics is very complex and can be characterized as a combinatory science, since it has a great complexity of regulatory cycles and network type interactions, which allows the utilization of computational models to resolve problems that can be converted into biological significant responses (Brusic & Petrovsky, 2003). This leads us to immunoinformatics, which is the application of informatics techniques to immune system molecules, with the main objective of helping develop vaccines through the prediction of immunogenic epitopes (Flower & Doytchinova, 2002).

8.1 Immunological databases

The immunological databases are a source of data used to explore, refine and develop new tools and algorithms (Salimi et al., 2010). There is a large variety of databases that group information relevant to the immune system. The Nucleic Acids Research Molecular Biology Database Collection http://www3.oup.co.uk/nar/database/c/ included 29 immunological databases in March 2011. The International ImMunoGeneTics information system (IMGT), the world reference databank for immunogenetics and immunoinformatics, was created by Marie-Paule Lefranc in 1989 (Lefranc et al., 2009). This databank is specialized in immunoglobulins or antibodies, T-cell receptors (TCR), MHCs, and others. The IMGT is constituted of a variety of databanks, including: structure, monoclonal antibody, sequence and genome databanks. All of these databanks are curated manually and daily by a team that works fulltime, which helps maintain high-quality annotation and standardization of the information. Other databases that house information related to epitopes, such as AntiJen (Toseland et al., 2005) and FIMM (Schonbach et al., 2000), have not been maintained and their data has migrated to other websites. Among these, the most promising epitope database seems to be the Immune Epitope Database (IEDB) (Peters et al., 2005), which is a curated database that has information based on experimental data associated with the target epitope; consequently, it is hoped that all of the information in the various existing databanks also migrates to IEDB within the next few years.

8.2 Epitope prediction

The principal goal of immunoinformatics is the development of algorithms that can both help develop vaccines and analyze the gene products of pathogens, such as viruses and bacteria. This is why it is very important to understand antigen-antibody interactions. Macallum et al. (1996) made a detailed analysis of 26 antigen–antibody complexes; they found that binding between molecules is very complex, and that there are different antibody–antigen classes for different types of molecules. A later study of 59 antigen–antibody interactions (Almagro, 2004) found results similar to those of Macallum. These studies show that tools that can identify molecules and predict their interactions with other molecules need to be very accurate and sensitive.

8.2.1 B cell epitope prediction

Epitopes of B cells are antigenic regions that are recognized by antibodies of the immune system, specifically those that interact with B cell receptors. These epitopes can be continuous or discontinuous (Kumagai & Tsumoto, 2001). B–cell epitopes can be used to design vaccines and new diagnostic tests (Larsen et al., 2006). As with T cells, there are also numerous methodologies to model and predict B–cell epitopes. The classic system to predict B–cell epitopes (Hopp & Woods, 1981) uses propensity scale methods (Parker et al., 1986; Levitt, 1978). This method attributes a propensity value to each amino acid, based on studies of the physical–chemical properties. A combination of various scales can improve the prediction results (Pellequer et al., 1991). This work used hydrophilicity scales (Parker et al., 1986), as well as secondary structure (Levitt, 1978; Chou & Fasman, 1978) and accessibility (Emini et al., 1985). The Immune Epitope Database and Analysis Resource, IEDB (Peters et al., 2005), utilizes parameters such as hydrophilicity, flexibility, accessibility, turns, exposed surface, polarity and antigenic propensity of polypeptides chains, which have been correlated with the location of continuous epitopes. All of the prediction calculations are based on propensity scales. Another methodology that can be used to predict continuous B-cell epitopes combines hidden Markov model (HMM) and propensity scale methods (Parker et al., 1986; Levitt, 1978); it is called Bepipred http://www.cbs.dtu.dk/services/BepiPred/ (Larsen et al., 2006). This methodology has given increased prediction accuracy. Prediction of discontinuous B–cell epitopes has also improved, due to an increase in the number of three–dimensional (3D) structures of antibody–antigen complexes available in PDB and in IMGT/3Dstructure-DB (Kaas et al., 2004) and in the Epitome database (Schlessinger et al., 2006).

8.2.2 T cell epitope prediction

There are two classes of T cells: (1) CD8+ T cytotoxic (Tc) cells, which produce cytotoxins responsible for cell lysis, recognize peptides presented by class I MHCs and (2) CD4+ T helper (Th) cells, which recognize proteins associated with MHC class II. Interferon γ (IFN–γ) and tumor necrosis factor β (TNF–β) are produced by Th1 cells. Th2 cells produce interleukin 4 (IL–4), IL–5, IL–10 and IL–13. Eptitopes that bind to MHC de class I generally are 8–10 amino acids long, with a mean of nine amino acids (Reche et al.., 2002), while epitopes that bind to MHC class II are 13–17 amino acids long (Sercarz & Maverakis, 2003; Chicz et al., 1992). There are various online tools for predicting T–cell epitopes on the basis of MHC class I and class II binding. Prediction of MHC binding is based on motifs associated with epitopes or binders for specific alleles. SYFPEITHI is a tool that is widely used for prediction of T–cell epitopes and MHC binding; however, these predictions have been found to be of low quality (Ruppert et al., 1993). More sophisticated tools that use quantitative matrixes, artificial neural network decision trees, hidden Markov models

(HMM), support vector machines (SVM), homology modeling, protein threading and docking techniques have been developed. The NetMHC 3.2 server http://www.cbs.dtu.dk/services/NetMHC/ predicts binding of peptides to a series of different HLA alleles using artificial neural networks (ANNs) and weight matrixes. All of the previous versions are available online, for comparison and reference. ANNs were trained with 57 different human MHCs (HLA), representing all of the 12 HLA alleles, supertypes A and B (Lund et al., 2004). Also predictions are available for 22 animal alleles (monkey and rat). ANN prediction values are given in nM IC50 values. Weight prediction matrixes use an aptitude score, with a high aptitude score indicating strong binding. Predictions can be made for sizes from 8 to 11 for all of the alleles using an ANNs algorithm trained with 9mer peptides. Probably because of the limited quantity of 10mer data available, this method has better prediction value when an ANNs algorithm is trained with 10mer data. However, one should be careful with 8mer predictions, since some alleles do not link to 8mer to a significant degree. Binding peptides are indicated at output as strongly binding (SB) and weakly binding (WB). The allele for each HLA supertype is indicated in the selection window for HLA alleles (Lundegaard et al., 2008).

The NetMHCII 2.2 server http://www.cbs.dtu.dk/services/NetMHCII/ predicts peptides that bind to MHC classe II alleles HLA-DR, HLA-DQ, HLA-DP and mouse alleles, using ANNs. Predictions can be obtained for the 14 HLA-DR alleles, including the nine HLA-DR, six HLA-DQ, and six HLA-DP supertypes and two H2 class II alleles in mice. The prediction values are given in nM IC50 values, and in %-Rank for a random set of 1,000,000 natural peptides. Strongly and weakly binding peptides are indicated in the output file (Nielsen et al., 2007).

Without a doubt, there is a great variety of predictors, which when they are combined can be quite precise in the prediction of T-cell epitopes; however, this is only possible when well-characterized alleles are available, which is true for some alleles that have been predicted as MHC class I alleles, but much less so for those predicted as MHC class II. This is even more of a problem in the prediction of B cell proteins, for which it is often necessary to have prior knowledge of the structure and sequence of the protein. Nevertheless, it is known that no method can go further than the data used to train it, and only through extensive compilation and by obtaining high quality data, will it be possible to create excellent models that will can be generally applied (Flower & Doytchinova, 2002).

9. References

Allen, J. E., Pertea, M. Salzberg, S. L. 2004. Computational gene prediction using multiple sources of evidence, *Genome Res* 14(1): 142–8.

Almagro, J. C. 2004. Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires, *J Mol Recognit* 17(2): 132–43.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. Lipman, D. J. 1990. Basic local alignment search tool, *J Mol Biol* 215(3): 403–10.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. Lipman, D. J. 1997. Gapped blast and psi-blast: a new generation of protein database search programs, *Nucleic Acids Res* 25(17): 3389–402.

Aparicio, G., Götz,, S., Conesa, A., Segrelles, D., Blanquer, I., García, J. M., Hernandez, V., Robles, M. Talon, M. 2006. Blast2go goes grid: developing a grid-enabled prototype for functional genomics analysis, *Stud Health Technol Inform* 120: 194–204.

Bambini, S. Rappuoli, R. 2009. The use of genomics in microbial vaccine development, *Drug Discov Today* 14(5-6): 252–60.

Barcellos, F. G., Menna, P., da Silva Batista, J. S. Hungria, M. 2007. Evidence of horizontal transfer of symbiotic genes from a bradyrhizobium japonicum inoculant strain to indigenous diazotrophs sinorhizobium (ensifer) fredii and bradyrhizobium elkanii in a brazilian savannah soil, *Appl Environ Microbiol* 73(8): 2635–43.

Barinov, A., Loux, V., Hammani, A., Nicolas, P., Langella, P., Ehrlich, D., Maguin, E. van de Guchte, M. 2009. Prediction of surface exposed proteins in streptococcus pyogenes, with a potential application to other gram-positive bacteria, *Proteomics* 9(1): 61–73.

Baxevanis, A. D. Ouellette, F. F. 2001. A practical guide to the analysis of genes and proteins, *Wiley* (2): 260–2.

Bendtsen, J. D., Nielsen, H., von Heijne, G. Brunak, S. 2004. Improved prediction of signal peptides: Signalp 3.0, *J Mol Biol* 340(4): 783–95.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. Wheeler, D. L. 2008. Genbank, *Nucleic Acids Res* 36(Database issue): D25–30.

Berriman, M. Rutherford, K. 2003. Viewing and annotating sequence data with artemis, *Brief Bioinform* 4(2): 124–32.

Blom, J., Albaum, S. P., Doppmeier, D., Pühler, A., Vorhölter, F.-J., Zakrzewski, M. Goesmann, A. 2009. Edgar: a software framework for the comparative analysis of prokaryotic genomes, *BMC Bioinformatics* 10: 154.

Blum, G., Ott, M., Lischewski, A., Ritter, A., Imrich, H., Tschäpe, H. Hacker, J. 1994. Excision of large dna regions termed pathogenicity islands from trna-specific loci in the chromosome of an escherichia coli wild-type pathogen, *Infect Immun* 62(2): 606–14.

Brown, T. A. 1999. Genes e expressÃ£o gênica., *Genética – um enfoque molecular* 1(2): 124–132.

Bru, C., Courcelle, E., Carrère, S., Beausse, Y., Dalmar, S. Kahn, D. 2005. The prodom database of protein domain families: more emphasis on 3d, *Nucleic Acids Res* 33(Database issue): D212–5.

Brusic, V. Petrovsky, N. 2003. Immunoinformatics–the new kid in town, *Novartis Found Symp* 254: 3–13; discussion 13–22, 98–101, 250–2.

Brüssow, H., Canchaya, C. Hardt, W.-D. 2004. Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion, *Microbiol Mol Biol Rev* 68(3): 560–602.

Chicz, R. M., Urban, R. G., Lane, W. S., Gorga, J. C., Stern, L. J., Vignali, D. A. Strominger, J. L. 1992. Predominant naturally processed peptides bound to hla-dr1 are derived from mhc-related molecules and are heterogeneous in size, *Nature* 358(6389): 764–8.

Choi, G.-E., Eom, S.-H., Jung, K.-H., Son, J.-W., Shin, A.-R., Shin, S.-J., Kim, K.-H., Chang, C. L. Kim, H.-J. 2010. Cysa2: A candidate serodiagnostic marker for mycobacterium tuberculosis infection, *Respirology* 15(4): 636–42.

Chou, P. Y. Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence, *Adv Enzymol Relat Areas Mol Biol* 47: 45–148.

Cole, S. T., Eiglmeier, K., Parkhill, J., James, K. D., Thomson, N. R., Barrell, B. G. 2001. Massive gene decay in the leprosy bacillus, *Nature* 409(6823): 1007–11.

Datta, S., Datta, S., Kim, S., Chakraborty, S. Gill, R. S. 2010. Statistical analyses of next generation sequence data: A partial overview, *J Proteomics Bioinform* 3(6): 183–190.

Diaz Romero, J. Outschoorn, I. M. 1994. Current status of meningococcal group b vaccine candidates: capsular or noncapsular? , *Clin Microbiol Rev* 7(4): 559–75.

Dobrindt, U. Hacker, J. 2001. Whole genome plasticity in pathogenic bacteria, *Curr Opin Microbiol* 4(5): 550–7.

Dobrindt, U., Janke, B., Piechaczek, K., Nagy, G., Ziebuhr, W., Fischer, G., Schierhorn, A., Hecker, M., Blum-Oehler, G. Hacker, J. 2000. Toxin genes on pathogenicity islands: impact for microbial evolution, *Int J Med Microbiol* 290(4-5): 307–11.

Emini, E. A., Hughes, J. V., Perlow, D. S. Boger, J. 1985. Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide, *J Virol* 55(3): 836–9.

Evans, M. C. 2008. Recent advances in immunoinformatics: application of in silico tools to drug development, *Curr Opin Drug Discov Devel* 11(2): 233–41.

Field, D., Feil, E. J. Wilson, G. A. 2005. Databases and software for the comparison of prokaryotic genomes, *Microbiology* 151(Pt 7): 2125–32.

Finn, R. D., Mistry, J., Schuster-Böckler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R., Eddy, S. R., Sonnhammer, E. L. L. Bateman, A. 2006. Pfam: clans, web tools and services, *Nucleic Acids Res* 34(Database issue): D247–51.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E. L. L., Eddy, S. R. Bateman, A. 2010. The pfam protein families database, *Nucleic Acids Res* 38(Database issue): D211–22.

Flower, D. R. Doytchinova, I. A. 2002. Immunoinformatics and the prediction of immunogenicity, *Appl Bioinformatics* 1(4): 167–76.

Gibas, C. Jambeck, P. 2001. Developing bioinformatics computer skills, *O'Reilly* 1(1): 21–22.

Gilks, W. R., Audit, B., De Angelis, D., Tsoka, S. Ouzounis, C. A. 2002. Modeling the percolation of annotation errors in a database of protein sequences, *Bioinformatics* 18(12): 1641–9.

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R. Goebel, W. 1990. Deletions of chromosomal regions coding for fimbriae and hemolysins occur in vitro and in vivo in various extraintestinal escherichia coli isolates, *Microb Pathog* 8(3): 213–25.

Hacker, J. Carniel, E. 2001. Ecological fitness, genomic islands and bacterial pathogenicity. a darwinian view of the evolution of microbes, *EMBO Rep* 2(5): 376–81.

Hershberg, R. Petrov, D. A. 2009. General rules for optimal codon choice, *PLoS Genet* 5(7): e1000556.

Hopp, T. P. Woods, K. R. 1981. Prediction of protein antigenic determinants from amino acid sequences, *Proc Natl Acad Sci U S A* 78(6): 3824–8.

Hou, Y. M. 1999. Transfer rnas and pathogenicity islands, *Trends Biochem Sci* 24(8): 295–8.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Yeats, C. 2009. Interpro: the integrative protein signature database, *Nucleic Acids Res* 37(Database issue): D211–5.

Hyatt, D., Chen, G.-L., Locascio, P. F., Land, M. L., Larimer, F. W. Hauser, L. J. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification, *BMC Bioinformatics* 11: 119.

Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. Madden, T. L. 2008. Ncbi blast: a better web interface, *Nucleic Acids Res* 36(Web Server issue): W5–9.

Kaas, Q., Ruiz, M. Lefranc, M. P. 2004. Imgt/3dstructure-db and imgt/structuralquery, a database and a tool for immunoglobulin, t cell receptor and mhc structural data, *Nucleic Acids Res* 32(Database issue): D208–10.

Karlin, S., Mrázek, J. Campbell, A. M. 1998. Codon usages in different gene classes of the escherichia coli genome, *Mol Microbiol* 29(6): 1341–55.

Kendrew, J. 1999. In: The encyclopedia of molecular biology, *in* B. Science (ed.), *Gene*, Porto Alegre, pp. 343–401.

Kislyuk, A. O., Katz, L. S., Agrawal, S., Hagen, M. S., Conley, A. B., Jayaraman, P., Nelakuditi, V., Humphrey, J. C., Sammons, S. A., Govil, D., Mair, R. D., Tatti, K. M., Tondella, M. L., Harcourt, B. H., Mayer, L. W. Jordan, I. K. 2010. A computational genomics pipeline for prokaryotic sequencing projects, *Bioinformatics* 26(15): 1819–26.

Krizova, L. Nemec, A. 2010. A 63 kb genomic resistance island found in a multidrug-resistant acinetobacter baumannii isolate of european clone i from 1977, *J Antimicrob Chemother* 65(9): 1915–8.

Krogh, A., Larsson, B., von Heijne, G. Sonnhammer, E. L. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes, *J Mol Biol* 305(3): 567–80.

Kumagai, I. Tsumoto, K. 2001. Antigen-antibody binding, *Encyclopedia of Life Sciences - Nature Publishing Group* pp. 1–7.

Langille, M. G. I. Brinkman, F. S. L. 2009. Islandviewer: an integrated interface for computational identification and visualization of genomic islands, *Bioinformatics* 25(5): 664–5.

Langille, M. G. I., Hsiao, W. W. L. Brinkman, F. S. L. 2008. Evaluation of genomic island predictors using a comparative genomics approach, *BMC Bioinformatics* 9: 329.

Lapierre, P. Gogarten, J. P. 2009. Estimating the size of the bacterial pan-genome, *Trends Genet* 25(3): 107–10.

Larsen, J. E., Lund, O. Nielsen, M. 2006. Improved method for predicting linear b-cell epitopes, *Immunome Res* 2: 2.

Lefranc, M. P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene, F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., Regnier, L., Ehrenmann, F., Lefranc, G. Duroux, P. 2009. Imgt, the international immunogenetics information system, *Nucleic Acids Res* 37(Database issue): D1006–12.

Lefébure, T. Stanhope, M. J. 2007. Evolution of the core and pan-genome of streptococcus: positive selection, recombination, and genome composition, *Genome Biol* 8(5): R71.

Lerat, E. Ochman, H. 2005. Recognizing the pseudogenes in bacterial genomes, *Nucleic Acids Res* 33(10): 3125–32.

Lesic, B., Bach, S., Ghigo, J.-M., Dobrindt, U., Hacker, J. Carniel, E. 2004. Excision of the high-pathogenicity island of yersinia pseudotuberculosis requires the combined actions of its cognate integrase and hef, a new recombination directionality factor, *Mol Microbiol* 52(5): 1337–48.

Levitt, M. 1978. Conformational preferences of amino acids in globular proteins, *Biochemistry* 17(20): 4277–85.

Li, L., Shiga, M., Ching, W.-K. Mamitsuka, H. 2010. Annotating gene functions with integrative spectral clustering on microarray expressions and sequences, *Genome Inform* 22: 95–120.

Liberman, F. 2004. *Análise dos fatores determinantes para a qualidade da anotação genÃ´mica automática*, Master's thesis, Universidade Católica de Brasília.

Lorenzi, H. A., Puiu, D., Miller, J. R., Brinkac, L. M., Amedeo, P., Hall, N. Caler, E. V. 2010. New assembly, reannotation and analysis of the entamoeba histolytica genome reveal new genomic features and protein content information, *PLoS Negl Trop Dis* 4(6): e716.

Lukashin, A. V. Borodovsky, M. 1998. Genemark.hmm: new solutions for gene finding, *Nucleic Acids Res* 26(4): 1107–15.

Lund, O., Nielsen, M., Kesmir, C., Petersen, A. G., Lundegaard, C., Worning, P., Sylvester-Hvid, C., Lamberth, K., Roder, G., Justesen, S., Buus, S. Brunak, S. 2004. Definition of supertypes for hla molecules using clustering of specificity matrices, *Immunogenetics* 55(12): 797–810.

Lundegaard, C., Lamberth, K., Harndahl, M., Buus, S., Lund, O. Nielsen, M. 2008. Netmhc-3.0: accurate web accessible predictions of human, mouse and monkey mhc class i affinities for peptides of length 8-11, *Nucleic Acids Res* 36(Web Server issue): W509–12.

Macallum, R. M., Martin, A. C. R. Thornton, J. M. 1996. Antibody-antigen interactions: Contact analysis and binding site topography, *Journal of Molecular Biology* 262: 732–45.

Mathé, C., Sagot, M.-F., Schiex, T. Rouzé, P. 2002. Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Res* 30(19): 4103–17.

Maurelli, A. T. 2007. Black holes, antivirulence genes, and gene inactivation in the evolution of bacterial pathogens, *FEMS Microbiol Lett* 267(1): 1–8.

Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K. Fasano, A. 1998. "black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of shigella spp. and enteroinvasive escherichia coli, *Proc Natl Acad Sci U S A* 95(7): 3943–8.

Mazumder, R. Vasudevan, S. 2008. Structure-guided comparative analysis of proteins: principles, tools, and applications for predicting function, *PLoS Comput Biol* 4(9): e1000151.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. Pühler, A. 2003. Gendb–an open source genome annotation system for prokaryote genomes, *Nucleic Acids Res* 31(8): 2187–95.

Mira, A., Martín-Cuadrado, A. B., D'Auria, G. Rodríguez-Valera, F. 2010. The bacterial pan-genome:a new paradigm in microbiology, *Int Microbiol* 13(2): 45–57.

Nielsen, M., Lundegaard, C. Lund, O. 2007. Prediction of mhc class ii binding affinity using smm-align, a novel stabilization matrix alignment method, *BMC Bioinformatics* 8: 238.

Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H.-Y., Cohoon, M., de Crécy-Lagard, V., Diaz, N., Disz, T. Edwards, e. a. 2005. The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes, *Nucleic Acids Res* 33(17): 5691–702.

Pareja, E., Pareja-Tobes, P., Manrique, M., Pareja-Tobes, E., Bonal, J. Tobes, R. 2006. Extratrain: a database of extragenic regions and transcriptional information in prokaryotic organisms, *BMC Microbiol* 6: 29.

Parker, J. M., Guo, D. Hodges, R. S. 1986. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and x-ray-derived accessible sites, *Biochemistry* 25(19): 5425–32.

Pearson, W. R. Lipman, D. J. 1988. Improved tools for biological sequence comparison, *Proc Natl Acad Sci U S A* 85(8): 2444–8.

Pellequer, J. L., Westhof, E. Van Regenmortel, M. H. 1991. Predicting location of continuous epitopes in proteins from their primary structures, *Methods Enzymol* 203: 176–201.

Peters, B., Sidney, J., Bourne, P., Bui, H. H., Buus, S., Doh, G., Fleri, W., Kronenberg, M., Kubo, R., Lund, O., Nemazee, D., Ponomarenko, J. V., Sathiamurthy, M., Schoenberger, S., Stewart, S., Surko, P., Way, S., Wilson, S. Sette, A. 2005. The immune epitope database and analysis resource: from vision to blueprint, *PLoS Biol* 3(3): e91.

Poptsova, M. S. Gogarten, J. P. 2010. Using comparative genome analysis to identify problems in annotated microbial genomes, *Microbiology* 156(Pt 7): 1909–17.

Prosdocimi, F. 2003. Bioinformática: manual do usuário., *Biotecnologia Ciência & Desenvolvimento* 2(29): 2.

Pundhir, S., Vijayvargiya, H. Kumar, A. 2008. Predictbias: a server for the identification of genomic and pathogenicity islands in prokaryotes, *In Silico Biol* 8(3-4): 223–34.

Rappuoli, R. 2000. Reverse vaccinology, *Curr Opin Microbiol* 3(5): 445–50.

Retter, I., Althaus, H. H., Munch, R. Muller, W. 2005. Vbase2, an integrative v gene database, *Nucleic Acids Res* 33(Database issue): D671–4.

Rinaudo, C. D., Telford, J. L., Rappuoli, R. Seib, K. L. 2009. Vaccinology in the genome era, *J Clin Invest* 119(9): 2515–25.

Ruppert, J., Sidney, J., Celis, E., Kubo, R. T., Grey, H. M. Sette, A. 1993. Prominent role of secondary anchor residues in peptide binding to hla-a2.1 molecules, *Cell* 74(5): 929–37.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. Barrell, B. 2000. Artemis: sequence visualization and annotation, *Bioinformatics* 16(10): 944–5.

Salimi, N., Fleri, W., Peters, B. Sette, A. 2010. Design and utilization of epitope-based databases and predictive tools, *Immunogenetics* 62(4): 185–96.

Salzberg, S. L., Delcher, A. L., Kasif, S. White, O. 1998. Microbial gene identification using interpolated markov models, *Nucleic Acids Res* 26(2): 544–8.

Schellenberg, M. J., Ritchie, D. B. MacMillan, A. M. 2008. Pre-mrna splicing: a complex picture in higher definition, *Trends Biochem Sci* 33(6): 243–6.

Schlessinger, A., Ofran, Y., Yachdav, G. Rost, B. 2006. Epitome: database of structure-inferred antigenic epitopes, *Nucleic Acids Res* 34(Database issue): D777–80.

Schmidt, H. Hensel, M. 2004. Pathogenicity islands in bacterial pathogenesis, *Clin Microbiol Rev* 17(1): 14–56.

Schonbach, C., Koh, J. L., Sheng, X., Wong, L. Brusic, V. 2000. Fimm, a database of functional molecular immunology, *Nucleic Acids Res* 28(1): 222–4.

Sercarz, E. E. Maverakis, E. 2003. Mhc-guided processing: binding of large antigen fragments, *Nat Rev Immunol* 3(8): 621–9.

Servant, F., Bru, C., Carrère, S., Courcelle, E., Gouzy, J., Peyruc, D. Kahn, D. 2002. Prodom: automated clustering of homologous domains, *Brief Bioinform* 3(3): 246–51.

Setúbal, J. Meidanis, J. 1997. *Introduction to Computational Molecular Biology*, Pacific Grove.

Sibbald, M. J. J. B. van Dij, J. M. l. 2009. Secretome mapping in gram-positive pathogens. in karl wooldridge (ed.), bacterial secreted protein: Secretory mechanisms and role in pathogenesis, *Caister Academic Press* pp. 193–225.

Sleator, R. D. 2010. An overview of the current status of eukaryote gene prediction strategies, *Gene* 461(1-2): 1–4.

Smith, T. F. Waterman, M. S. 1981. Identification of common molecular subsequences, *J Mol Biol* 147(1): 195–7.

Stein, L. 2001. Genome annotation: from sequence to biology, *Nat Rev Genet* 2(7): 493–503.

Stothard, P. Wishart, D. S. 2006. Automated bacterial genome analysis and annotation, *Curr Opin Microbiol* 9(5): 505–10.

Suzuki, T. Sasakawa, C. 2001. Molecular basis of the intracellular spreading of shigella, *Infect Immun* 69(10): 5959–66.

Takai, S., Hines, S. A., Sekizaki, T., Nicholson, V. M., Alperin, D. A., Osaki, M., Takamatsu, D., Nakamura, M., Suzuki, K., Ogino, N., Kakuda, T., Dan, H. Prescott, J. F. 2000. Dna sequence and comparison of virulence plasmids from rhodococcus equi atcc 33701 and 103, *Infect Immun* 68(12): 6840–7.

Trost, B., Haakensen, M., Pittet, V., Ziola, B. Kusalik, A. 2010. Analysis and comparison of the pan-genomic properties of sixteen well-characterized bacterial genera, *BMC Microbiol* 10: 258.

Tumapa, S., Holden, M. T. G., Vesaratchavest, M., Wuthiekanun, V., Limmathurotsakul, D., Chierakul, W., Feil, E. J., Currie, B. J., Day, N. P. J., Nierman, W. C. Peacock, S. J. 2008. Burkholderia pseudomallei genome plasticity associated with genomic island variation, *BMC Genomics* 9: 190.

Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. Banfield, J. F. 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature* 428(6978): 37–43.

UniProt 2007. The universal protein resource (uniprot), *Nucleic Acids Res* 35(Database issue): D193–7.

Waack, S., Keller, O., Asper, R., Brodag, T., Damm, C., Fricke, W. F., Surovcik, K., Meinicke, P. Merkl, R. 2006. Score-based prediction of genomic islands in prokaryotic genomes using hidden markov models, *BMC Bioinformatics* 7: 142.

Zhao, J. Grant, S. F. A. 2010. Advances in whole genome sequencing technology, *Curr Pharm Biotechnol*.

6.3 Genômica Estrutural

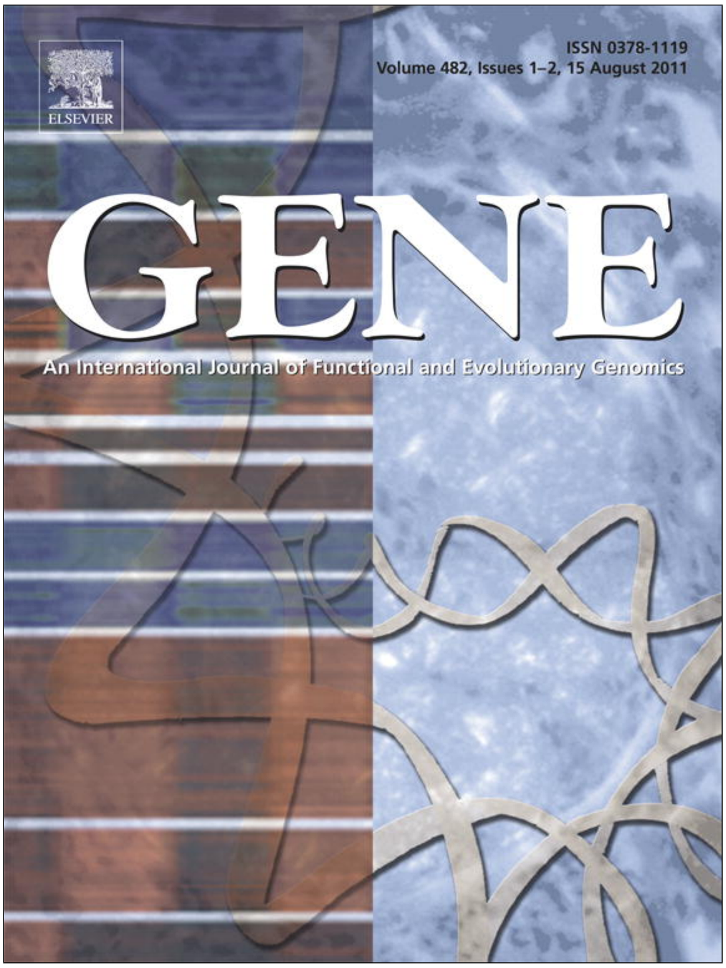
Os artigos científicos a seguir foram realizados em parceria com grupos de pesquisa parceiros do LGCM e portanto não são o foco principal dessa tese de doutorado. O doutorando participiou auxiliando os demais pesquisadores com conhecimentos sobre ferramentas de análises de sequências biológicas, dentre elas, ferramentas para alinhamento de sequências e para predição *in silico* do local subcelular de proteínas.

6.3.1 Reparo de DNA no modelo *Corynebacterium*

Corynebacterium spp. incluem patógenos de plantas e animais, além de bactérias não patogênicas do solo e espécies saprófitas. A compreensão da biologia destes organismos está aquém do desejado ao compararmos com a compreensão de outros organismos bacterianos, mas novas perspectivas oferecidas pelos dados da sequência do genoma e a elucidação dos conteúdos gênicos forneceu pistas sobre a natureza, a estabilidade do genoma, patogenicidade e virulência destes organismos. Foram comparados 15 genomas de *Corynebacterium*, entre espécies patogênicas e não patogênicas, com foco em genes de reparo do DNA.

O reparo do DNA é um mecanismo de grande importância na manutenção da estabilidade do genoma de qualquer organismo; a ineficiência desse sistema pode promover a instabilidade genômica e levar à morte celular. Técnicas que utilizam a ineficiência desses sistemas são uma estratégia interessante no estudo de meios para controlar organismos infecciosos. Descobrimos que o reparo por excisão de nucleotídeos (NER) foi a única via envolvida cujos genes foram encontrados em todas as espécies, sugerindo que a integridade do DNA pode ser mantida principalmente por NER. O reparo por recombinação (RR) é também uma via bem conservada e a maioria dos genes RR existem geralmente no gênero *Corynebacterium*. A ausência de genes *recCD* também foi compartilhada por todas as espécies, contribuindo para evitar inversões do genoma e o favorecimento da estabilidade genômica. Alguns genes da via *mismatch repair* (MMR), como *Mut* (*mutY* e *mutL*), estão presentes apesar da ausência de outros. O reparo por excisão de base (BER) e vias de reparo diretos não são vias conservadas, uma vez que os genes não são compartilhados por todas as espécies. No entanto, a existência de alguns genes parece ser suficiente para garantir a atividade da via. Um fato interessante é a persistência/aquisição de alguns genes de reparo em algumas espécies, sugerindo um papel importante na manutenção do DNA e evolução. Estes genes podem ser alvos importantes na investigação do papel de reparo do DNA na patogenicidade de espécies do gênero *Corynebacterium* e serem usados como alvos de intervenção terapêutica. A análise filogenética dos genes *uvrABC* NER mostrou um padrão de agrupamentos, em que a maioria dos agrupamentos era compartilhada. Em geral, a presença ou a inexistência de genes de reparo foi compartilhada por todas as espécies analisadas. A perda ou a aquisição de certos genes de DNA *repair* é sugestivo de ter sido um evento ancestral.

Provided for non-commercial research and education use.
Not for reproduction, distribution or commercial use.


ISSN 0378-1119
Volume 482, Issues 1–2, 15 August 2011
ELSEVIER
GENE
An International Journal of Functional and Evolutionary Genomics

This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

http://www.elsevier.com/copyright

Author's personal copy

Gene 482 (2011) 1–7

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Review

DNA repair in *Corynebacterium* model

B.C. Resende [a], A.B. Rebelato [a], V. D'Afonseca [b], A.R. Santos [b], T. Stutzman [b], V.A. Azevedo [b], L.L. Santos [a], A. Miyoshi [b], D.O. Lopes [a,*]

[a] *Laboratório de Genética Molecular, Universidade Federal de São João Del-Rei, CCO, Divinópolis, MG, Brazil*
[b] *Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, ICB, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil*

ARTICLE INFO

Article history:
Accepted 28 March 2011
Available online 8 April 2011

Received by A.J. van Wijnen

Keywords:
DNA repair
Bioinformatic
Corynebacterium
Orthologous genes

ABSTRACT

Corynebacterium spp. are a group of Gram-positive bacteria that includes plant and animal pathogens, nonpathogenic soil bacteria, and saprophytic species. Our understanding of these organisms is still poor compared with that of other bacterial organisms, but new insights offered by genome sequence data and the elucidation of gene content has provided clues about the nature, genome stability, pathogenicity and virulence of these organisms. We compared 15 *Corynebacterium* genomes, from pathogenic and nonpathogenic species, focusing on DNA repair genes. DNA repair is a mechanism of great importance in the maintenance of the genomic stability of any organism; inefficiency of this system can promote genomic instability and lead to death. This vulnerability makes it an interesting target in the study of means to control infectious organisms. We found that nucleotide excision repair (NER) was the only pathway whose involved genes were found in all species, suggesting that DNA integrity can be primarily maintained by NER. Recombination repair (RR) is also a well conserved pathway and most RR genes exist commonly in *Corynebacterium* spp. Absence of *recCD* genes was also shared by all species, contributing to prevent genome inversions and favoring genomic stability. Mismatch repair (MMR) appeared to be missing, although some genes in this pathway, such *mutT*, *mutY* and *mutL*, are present. Base excision repair (BER) and direct repair pathways are not conserved pathways, since the genes are not shared by all members; however, the existence of some seems to be enough to ensure pathway activity. An interesting fact is the persistence/acquisition of some repair genes in some species, suggesting an important role in DNA maintenance and evolution. These genes can be important targets in the investigation of the role of DNA repair in the pathogenicity of *Corynebacteirum* species and be used as targets in therapeutic intervention. Phylogenetic analysis of *uvrABC* NER genes showed a pattern of clusters, in which most groups remained fixed. In general, the presence or inexistence of repair genes was shared by all the species we analyzed, and the loss or acquisition of certain DNA repair genes seems to have been an ancestral event.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Cells are constantly exposed to hostile conditions and DNA-damaging agents, both of which cause DNA lesions. To prevent the potential deleterious effects of these lesions, cells have developed an elaborate system of DNA repair that is composed of a group of enzymes that verify the type of injury and repair it. Given the vast number of mutagens present in the environment, as well as the intrinsic error rate of DNA replication, cells have evolved in several ways to counteract these adverse effects. DNA repair as a whole is a highly complex process of metabolism and is indispensable for genome maintenance and organism survival (Carvalho et al., 2005).

DNA repair pathways are well established and are activated according to the type of damage. They function by eliminating or reducing the deleterious effects of the lesions. There is evidence that repair pathways complement one another or require two genes for crucial repair steps. DNA repair is achieved by several methods. Initially, there can involve direct reversal (DR) of damage, which reverses the damage without excision of DNA (Drables et al., 2004). Then, there is base excision repair (BER), which uses DNA glycosylase to recognize and excise the damaged base (Xu et al., 2008). A third option is nucleotide excision repair (NER), which recognizes DNA distortion and uses a protein complex to perform the repair by excising a small group of bases. A fourth system is mismatch repair (MMR), which recognizes bases that are poorly matched and corrects them (Antony and Hingorani, 2003). A fifth system is recombination repair (RR), which acts on double-stranded breaks.

Abbreviations: 16S rDNA, 16S subunit ribosomal RNA coding gene; DNA, deoxyribonucleic acid; BER, base excision repair; DR, direct reversal; DDR, direct damage reversal; MMR, mismatch repair; NER, nucleotide excision repair; RR, recombination repair.

* Corresponding author at: Sebastião Gonçalves Coelho, 400-Chanadour, CEP 35501-296, Divinópolis, MG, Brazil.
E-mail address: debora@ufsj.edu.br (D.O. Lopes).

0378-1119/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2011.03.008

Author's personal copy

If the number of injuries exceeds a threshold, cells activate an important pathway called the SOS system. At this critical moment, several proteins are expressed, among them, polymerases, ligases and endonucleases, whose goal is to ensure cell survival, even though there may be some incorporated errors (Aravind et al., 1999).

The fact that DNA repair genes are generally highly conserved in different organisms demonstrates their importance for genomic stability. Their ubiquitousness means that they could be used as target for therapeutic intervention in various species, including pathogenic *Corynebacterium* spp. (Xu et al., 2008). This genus is closely related to *Mycobacterium* spp. and consists of a large number of Gram-positive bacteria that are pleiomorphic, asporogenous, have a high G + C content in their genomes (Deb and Nath, 1999) and include nonpathogens, animal and plant pathogens, and saprophyte (Collins and Bradbury, 1986). Progress in bacterial genome projects and the advances in the development of bioinformatics tools have made it possible to study the evolution of genome structure, allowing a comparative genomic analysis between these closely related species. (Nakamura et al., 2003; Fudou, 2002).

We performed a study of the DNA repair pathways in 15 *Corynebacterium* genomes, including two nonpathogenic species: *C. efficiens*, and *C. glutamicum*, and 13 pathogenic species: *C. accolens, C. ammoniagenes, C. aurimucosum*, *C. diphtheriae*, *C. genitalium, C. glucuronolyticum, C. jeikeium*, *C. kroppenstedtii*, *C. lipophiloflavum, C. matruchotii, C. pseudotuberculosis*, *C. striatum,* and *C. urealyticum*, that infect the respiratory and urogenital tracts, skin, and visceral organs of man and various animals of economic interest. Then, we conducted a comparative study of the DNA repair systems of these organisms; the main focus was identifying genes or pathways important to ensure stability of the genome, conserved pathways, and major differences in DNA repair among the species.

The elucidation of the repair pathways in *Corynebacterium* spp. is indispensable for understanding the biology of these bacteria and can provide clues about pathogenicity and virulence, essential information for designing antibacterial strategies (D'Afonseca et al., 2009; Dorella et al., 2006).

2. Material and methods

2.1. Complete genome sequences

The genome complete sequences of *C. diphtheria* NCTC13129 (Cerdeno-Tarraga et al., 2003), *C. aurimucosum* ATCC700975 (Trost et al., 2010, *C. efficiens* YS314 (Nishio et al., 2003), *C. glutamicum* ATCC13032 (Kalinowski et al., 2003), *C. jeikeium* K411 (Tauch et al., 2005), *C. kroppenstedtii* DSM44385 (Tauch et al., 2008a, 2008b), and *C. urealyticum* DSM7109 (Tauch et al., 2008a, 2008b) were obtained from the Bielefeld University Center of Biotechnology in the CoryneRegNet 6.0 database and from the National Center of Biotechnology (http://www.ncbi.nlm.nih.gov) (accession numbers: BX248353, CP001601, BA000035, BA000036, CP001620, CR931997, AM942444, respectively).

The genome sequences *C. accolens* ATCC49725 (ACGD00000000), *C. ammoniagenes* DSM20306 (ADNS00000000), *C. genitalium* ATCC33030 (ACLJ00000000), *C. glucuronolyticum* ATCC51866 (ACHF00000000), *C. lipophiloflavum* DSM44291 (ACHJ00000000), *C. matruchotii* ATCC33806 (ACEB00000000), *C. striatum* ATCC6940 (ACGE00000000) and *C. pseudotuberculosis* CP1002 were deposited in the National Center of Biotechnology as draft genome (unpublished).

2.2. Presence of orthologous genes

To find orthologous genes, we conducted a FASTA search between 15 species, using each protein sequence as a query in a given genome, against the database composed of all genes in the other genome (value cutoff $= 10^{-8}$). Next, we conducted the FASTA search by alternating the query genome and the database genome. These sequences were also analyzed with other genomic tools, including InterProScan (http://ebi.ac.uk/interproscan/) and ScanProsite tool (http://expasy.org/tools/scnpsit3.html).

2.3. Phylogenic analyses

The evolutionary history was inferred using the neighbor-joining method (Saitou and Nei, 1987) and NER enzymes of 21 *Corynebacterium* spp. Along with the original 15, we also included the sequences of *C. pseudogenitalium, C. tuberculostearicum, C. pseudotuberculosis C231 and FRC41, C. amycolatum* and *C. resistens*. The optimal tree with the sum of branch lengths is shown. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (500 replicates) is shown next to the branches. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree (Tamura et al., 2004). The evolutionary distances were computed using the Poisson correction method (Zuckerkandl and Pauling, 1965) and the units are the number of amino acid substitutions per site. All positions containing gaps, and missing data was eliminated from the dataset (complete deletion option). Phylogenetic analyses were conducted with MEGA4 (Tamura et al., 2007). We obtained the 16S rRNA sequences to verify the evolutionary distances of *Corynebacterium* spp. and to compare with NER trees. Phylogenetic analyses were conducted as previously described to NER proteins trees.

3. Results

3.1. Genome content

After obtaining the sequences of the genome projects of the 15 *Corynebacterium* spp., we found that this genus have a high G + C content in their genomes and the mean genome size of the species was 2.6 Mb, the smallest is from *C. lipophiloflavum* with 2.2 Mb and the largest is from *C. pseudotuberculosis* with 3.5 Mb (data not show).

3.2. Nucleotide excision repair (NER)

NER is able to recognize and repair many types of damage that causes DNA double-helix distortion, including pyrimidine dimers induced by UV light and DNA intrastrand cross-links (Zhang et al., 2009). In *E. coli,* NER is mediated by *UvrABC* and *UvrD*. The UvrABC pathway is responsible for the identification of damage and cleavage of oligonucleotides that contain lesions, and *UvrD* encodes a helicase that removes the lesion, generating a DNA gap that is filled by DNA polymerase and then sealed by ligase (Carvalho et al., 2005).

The *uvrABCD* genes were identified in all the *Corynebacterium* spp. and the sequences showed a high level of conservation (data not show). This high degree of conservation qualifies the NER genes as candidates for the construction of phylogenetic models. This pathway seems to be the most generic method of DNA repair; it is normally used as a complementary backup system for other more specific pathways, such as BER, involving the use of endonucleases with relatively low specificity.

3.3. Base excision repair (BER)

BER is a multiprotein pathway composed of four proteins; these proteins include a DNA glycosylase, an AP endonuclease or AP DNA lyase, a DNA polymerase, and a DNA ligase. The proteins involved in this core reaction work together in a coordinated fashion to remove a damaged DNA base and replace it with the correct base (Robertson et al., 2009; Huffman, et al., 2005; Lindahl, 2001). Unlike in NER, these enzymes act independently and not in consecutive steps in a pathway (Lau et al., 2000; Dizdaroglu, 2005).

All the genomes that we analyzed were found to have various DNA glycosylases (*ung, mug, nth, fpg, nei, tag1, alkA*) and one AP

Author's personal copy

endonuclease (*xth*), which seem to be sufficient to ensure the functioning of this pathway (Table 1). The existence of different types of this enzyme is explained by the specificity in recognizing different types of injury, due to oxidative stress within the host (Doublie et al., 2004; Wallace et al., 2003).

3.4. Recombination (RER)

Double-strand breaks of DNA and other lesions, including cross-links, can stop the replication fork and activate cellular apoptosis if they are not repaired. There are approximately 40 known enzymes involved in recombination repair (Dos Vultos et al., 2009), making this one of the most complex systems of repair; it is divided into four steps: initiation, strand exchange, Holliday junction migration, and resolution (Carvalho et al., 2005).

As with NER, the recombination pathways in *Corynebacterium* were found to be highly conserved. In the recBCD pathway, only *recB,* which encodes a helicase, was found. On the other hand, all the ORFs were found in the recF pathway, with the exception of *recJ*, an exonuclease, found only in *C. glutamicum*. Two of the three exonucleases, *sbcC* and *sbcD*, were found in the SbcBCD pathway. The AddAB pathway is absent, though other related recombination genes, including *recG, ruvABC, xerCD,* lig, *recX*, *polA*, *ssb*, *radA lexA*, and *recA*, were found in all the *Corynebacterium* spp.; the latter two are directly related to the SOS response (Table 2).

3.5. Direct damage reversal (DDR)

Direct repair is one of the simplest forms of repair; it consists of the removal, in one step, of only the base-modifying agent, without the need to remove the base, which is needed in BER (Nieminuszczy and Grzesiuk, 2007).

This pathway is not conserved in *Corynebacterium*; however, we observed at least two direct repair enzymes in the species that we analyzed. The gene *ada,* which encodes a methyltransferase, was found in all, whereas *alkB*, which encodes a demethylase, *ogt,* which encodes an alkyltranferase, and *phrB*, which encodes a photolyase, were found in most. The gene *phr*, a second photolyase enzyme that repairs pyrimidine dimers by photoreaction, seems to have been lost during evolution (Table 3) (Goosen and Moolenaar, 2008; Nieminuszczy and Grzesiuk, 2007).

3.6. Mismatch repair (MMR)

MMR in prokaryotes is initiated when mismatches are recognized by a highly conserved MMR protein, MutS. This protein and a second conserved protein, MutL, act in concert to enable the excision repair pathway by activating endonucleolytic cleavage by a third MMR protein, MutH, which directs its nicking activity to the unmethylated strand at transiently hemimethylated CATC sites shortly after replication. This ethyl-directed nicking by MutH ensures that MMR in *E. coli* is directed to the newly synthesized DNA strand containing the error. Then, the exonucleases, in the presence of a helicase, promote excision of the fragment; the resulting gap is filled by a specific DNA polymerase (Razin et al., 1998).

In these *Corynebacterium* spp., the mismatch repair pathways lacked two important MMR genes, *mutS* and *mutH*, responsible for mismatch recognition and incision, respectively. Curiously, *mutL*, *uvrD* and exonuclease VII, genes involved in the MMR pathway, were found (Table 4). The exonucleases *exoI*, *rec*J and enzymes, encoded by the genes *dhs1* and *vsr*, were absent. The other MMR genes *mutT* and *mutY*, which encode enzymes that recognize adenine mispaired with oxidized guanine, were found in almost all species (Hall and Matson, 1999; Fowler et al., 2003).

3.7. Other repair genes

Other genes involved in DNA repair were analyzed; among them we found conservation of *LexA*, which encodes a protein involved in SOS system activation, *ssb*, which encodes proteins that stabilize single strand DNA, and at least one of the three helicases that we analyzed (Table 5).

3.8. Phylogenetic analysis of DNA repair proteins in Corynebacterium

Repair genes are among the sequences with a high level of conservation probably because they are important for the conservation and maintenance of genetic stability. We did, through the same methodology, filogenetic trees of NER and observed a similar pattern of clusters. We found two major branches, a minor one that includes *C. amycolatum, C. jeikeium, C. resistens, C. urealyticum* and *C. kroppenstedtii*, and a major one containing three sub-branches. The first sub-branch contains *C. pseudogenitalium, C. tuberculostearicum, C. accolens, C. aurimucosum, C. striatum,* and *C. amoniagenes.* The second sub-branch contains *C. pseudotuberculosis* 1002, C231, FRC41 and *C. diphtheria*, that share a common ancestor, and the third sub-branch contains nonpathogenic *C. efficiens* and *C. glutamicum,* that also share a common ancestor in all trees. The species *C. matruchotii, C. genitalium, C. lipophiloflavum* and *C. glucuronolyticum* did not remain in fixed groups, changing their position according to the gene analyzed (Fig. 1).

The phylogenetic tree obtained from 16S rRNA showed similar topology to that observed in NER trees. Like these, there are two mean branches sharing almost the same species; however, we observed that

Table 1
Predicted base excision repair genes in *Corynebacterium* spp.

Gene	Description	Cacc	Camm	Caur	Cdip	Ceff	Cgen	Cglc	Cglu	Cjei	Ckro	Clip	Cmat	Cpse	Cstr	Cure[a]
ung	Uracil glycosylase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
mug	T:G, T:U glycosylase	−	+	−	+	−	+	+	−	−	−	−	−	+	−	−
ogg	8-oxoG glycosylase	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
nth/mutY	Endonuclease III	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
fpg/mutM	Formamidopyrimidine-DNA glycosylase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
nei	Endonuclease VIII	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
mpg	3-meA glycosylase	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
tag1	3-meA glycosylase	+	+	+	+	+	−	+	+	+	+	+	+	+	+	+
xth	Exo III	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
nfo	Endo IV	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
alkA	3-meA glycosylase	+	−	+	−	+	+	−	+	+	+	+	+	−	+	+
dnlI	NAD-dependent DNA ligase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ligII	ATP-dependent DNA ligase	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−

[a] Cacc: *C. accolens*; Camm: *C. ammoniagenes*; Caur: *C. aurimucosum*; Cdip: *C. diphteriae*; Ceff: *C. efficiens*; Cgen: *C. genitalium*; Cglc: *C. glucuronolyticum*; Cglu: *C. glutamicum*; Cjei: *C. jeikeium*; Ckro: *C. kroppenstedtii*; Clip: *C. lipophiloflavum*; Cmat: *C. matruchotii*; Cpse: *C. pseudotuberculosis*; Cstr: *C. striatum*; Cure: *C. urealyticum*. The symbol + indicates likely present and − indicates likely absent.

Author's personal copy

Table 2
Predicted recombination repair genes in *Corynebacterium* spp.

Gene	Descripition	Cacc	Camm	Caur	Cdip	Ceff	Cgen	Cglc	Cglu	Cjei	Ckro	Clip	Cmat	Cpse	Cstr	Cure
RecBDC pathway																
recB	ExoV helicase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
recC	ExoV nuclease	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
recD	ExoV helicase	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
RecF pathway																
recF	ss and dsDNA binding	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
recJ	5′–3′ ssDNA exonuclease	−	−	−	−	−	−	−	+	−	−	−	−	−	−	−
recN (radB)	ATP binding	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
recO	ssDNA/RecA loading	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
recQ	3′–5′ DNA ligase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
recR	ssDNA/RecA stabilization	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
RecE pathway																
recE	5′–3′ dsDNA exonuclease	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
recT	Recombinase	−	−	+	−	−	−	−	−	−	−	−	−	−	−	−
SbcBCD pathway																
sbcB	3′–5′ ssDNA exonuclease	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
sbcC	dsDNA exonuclease	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
sbcD	dsDNA exonuclease	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
AddAB pathway																
addA	Exo, helicase with addB	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
addB	Exo, helicase with addA	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
Related genes																
recA	Recombinase, strandexchange	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
recG	Resolvase, 3′–5′ helicase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rus	Junction endonuclease	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
ruvA	Holliday junction helicase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ruvB	Holliday junction helicase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ruvC	Junction endonuclease	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
xerC	Recombinase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
xerD	Recombinase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
lexA	SOS transcription repressor	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
lig	DNA ligase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
recX	Regulatory protein for RecA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
polA	DNA Polymerase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
priA	Replication factor Y	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
radA	Recombinase	+	+	+	+	+	+	+	+	+	−	+	+	+	+	+
radC	Recombinase	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
ssb	SS binding protein	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

nonpathogenic corynebacteria branch and *C. pseudotuberculosis* and *C. diphtheriae* branch are evolutionarily closer to the branch containing *C. amycolatum, C. jeikeium, C. resistens, C. urealyticum* and *C. kroppenstedtii* in 16S rRNA tree.

4. Conclusions

The main DNA repair pathways, responsible for cell vitality, are conserved in *Corynebacterium* spp. In general, presence or inexistence of repair genes was shared, suggesting that the loss of a majority of DNA repair genes was an ancestral event, occurring before the divergence of *Corynebacterium*. The most conserved repair pathway seems to be nucleotide excision repair, which allows recognition of lesions that cause structural distortions, such as photoproducts induced by UV light and chemical adducts. The conservation and the possibility of interaction with other repair enzymes, suggest that genomic stability can be primarily maintained by nucleotide excision repair (NER) (Carvalho et al., 2005).

When we examined BER, all bacteria genomes were found to have at least one glycosylase, along with AP endonucleases, which seems sufficient to ensure the functioning of this pathway. We observed the acquisition or retention of genes in some species, like *mug*, which encodes a uracil glycosilase, and *alkA*, which encodes a 3-metil adenine glycosilase. Acquisition and loss of genetic material are key mechanisms in bacterial evolution, as they are essential for adaptation to new lifestyles and pathogenicity. Studies of pathogenic *Corynebacterium* spp. knockouts, like *mug* gene, not found in pathogenic bacterias and present in some pathogenic species, such as *C. pseudotuberculosis* and *C. diphtheriae*, may reveal the importance of DNA repair in the survival of these parasites in host and can give us clues about pathogenicity and virulence mechanisms.

Most of the recombinational repair genes that we examined are common among *Corynebacterium* spp. Genome rearrangement in an

Table 3
Predicted direct repair genes in *Corynebacterium* spp.

Gene	Description	Cacc	Camm	Caur	Cdip	Ceff	Cgen	Cglc	Cglu	Cjei	Ckro	Clip	Cmat	Cpse	Cstr	Cure
alkB	Oxidative demethylase	+	+	−	+	+	−	−	+	−	−	+	−	+	+	+
ada	Methyltransferase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ogt	Alkyltranferase	+	−	+	+	+	+	+	+	+	+	−	+	+	−	+
phr	Photolyase	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
phrB	Photolyase	+	−	−	+	+	+	−	+	−	−	+	−	−	+	−

Author's personal copy

Table 4
Predicted mismatch repair genes in *Corynebacterium* spp.

Gene	Description	Cacc	Camm	Caur	Cdip	Ceff	Cgen	Cglc	Cglu	Cjei	Ckro	Clip	Cmat	Cpse	Cstr	Cure
mutS1	Methyl-directed mismatch repair	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
mutS2	Methyl-directed mismatch repair	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
hnh	Endonuclease	+	+	+	+	+	+	+	+	+	+	+	+	+	+	–
mutT	Oxoguanine-triphosphatase	+	+	+	–	+	+	+	+	–	+	+	+	+	+	+
mutY	A:G adenine glycosylase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
mutL	T:G mismatch endonuclease	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
vsr	GATC endo	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
mutH	GATC methyl-directed/endo	–	+	–	–	–	–	–	–	–	–	–	–	–	–	–
dam	Methyltransferase	+	+	+	–	–	–	–	–	–	–	–	–	–	+	–
exoI	5′–3′ ssDNA exo	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
recJ	5′–3′ ssDNA exo with xseB	–	–	–	–	–	–	–	+	–	–	–	–	–	–	–
xseA	Deoxyribonuclease VII large subunit	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
xseB	Deoxyribonuclease VII small subunit	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
dhs1	Precursor	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
uvrD	Helicase II	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

organism involves exchanges and shuffles of DNA segments in the chromosome. This phenomenon, related to the recombination repair system, is useful for understanding the evolution and genome stability of these organisms. Absence of *recCD* and presence of *recB* genes was found to be shared by all the species. RecBCD, also known as Exonuclease V, is an enzyme initially described in *E. coli* that initiates recombinational repair from potentially lethal double strand breaks in DNA which may result from ionizing radiation, replication errors, endonucleases, and a host of other factors. This enzyme is a helicase that unwinds or separates the strands of DNA and a nuclease that makes single-stranded nicks in DNA. The permanence of a single gene of this pathway in all analyzed genomes suggests a possible participation in another route or protein complex, since it is known that pathways can complement each other.To test this possibility will need to conduct functional heterologous complementation tests. The same was observed in sbcBCD pathway. The genes *sbcC* and *sbcD* that encode exonucleases were found while *sbcB*, which encodes an exonuclease–deoxyribophosphodiesterase in this pathway, was absent.

DNA mismatch repair proteins are ubiquitous players in a diverse array of important cellular functions. In its role in post-replication repair, MMR safeguards the genome by correcting base mispairings that arise as a result of replication errors (Brown et al., 2003). We found that absence of *mutS* and *mutH* genes was shared among the *Corynebacterium* spp. Loss of MMR can result in greatly increased rates of spontaneous mutation, which is important for organisms that have a parasitic life, such as some *Corynebacterium* spp., since it can allow for better adaptation to new environments (Razin et al., 1998). Furthermore, we found that *C. diphteriae* and *C. jeikeim* lack *mutT*, one of the mutator genes, whereas the other species retain it (Horst et al., 1999), Defective mutation of mutT in *E. coli* increases the frequency of transversion from A–T to C–G and increases GC content, a feature of the *Corynebacterium* genus.

Another interesting fact is the persistence or acquisition of the *dam* gene for only four species. The *dam* gene encodes a DNA methyltransferase that methylates adenine in –GATC– sequences in double-stranded DNA. Mutant strains that lacking this enzyme display a pleiotropic phenotype including increased mutability, hyper-recombination and increased sensitivity to DNA-damaging agents, making it an important target in the study of variability, stability genome maintenance, pathogenicity and virulence observed in some strains.

The main enzymes that activate the SOS system are present and guarantee its activation in *Corynebacterium*; this helps ensure the survival of cells, even though it can lead to mutation accumulation. Finally, in direct repair system, the simplest pathway, we verified that the *alkB*, which encodes a demethylase, *ogt* gene, which encodes an alkyltransferase and *phrB,* that encodes a photolyase, are present only in some of analyzed Corynebacteria. The permanence of these genes can ensure an increase of genomic stability and decrease the chances of errors perpetuation in these species.

DNA repair is critical for the survival of pathogens inside the host because of the DNA lesions introduced in the genome of the pathogen by harmful agents released from the host. An understanding of the precise molecular functions of the enzymes participating in DNA metabolism and in the maintenance of pathogen genome integrity can be the key to search for potential targets to be used in therapeutic intervention.

Acknowledgments

This work received financial support of the CNPq/MAPA and Fapemig.

Table 5
Others predicted repair genes in *Corynebacterium.*

Gene	Description	Cacc	Camm	Caur	Cdip	Ceff	Cgen	Cglc	Cglu	Cjei	Ckro	Clip	Cmat	Cpse	Cstr	Cure
umuC	SOS mutagenesis	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
umuD	SOS transcription repressor	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
lexA	SOS transcription repressor	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
priA	Primosome helicase	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p53	Txn, tumor suppressor	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
ssb	Binds ssDNA	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
hepA	Helicase	–	+	+	+	+	–	+	+	+	+	–	+	+	–	+
hepA2	Helicase	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
spl	Repair spore UV dimers	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
lon	Protease	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
uvde	UV damage endo	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Author's personal copy

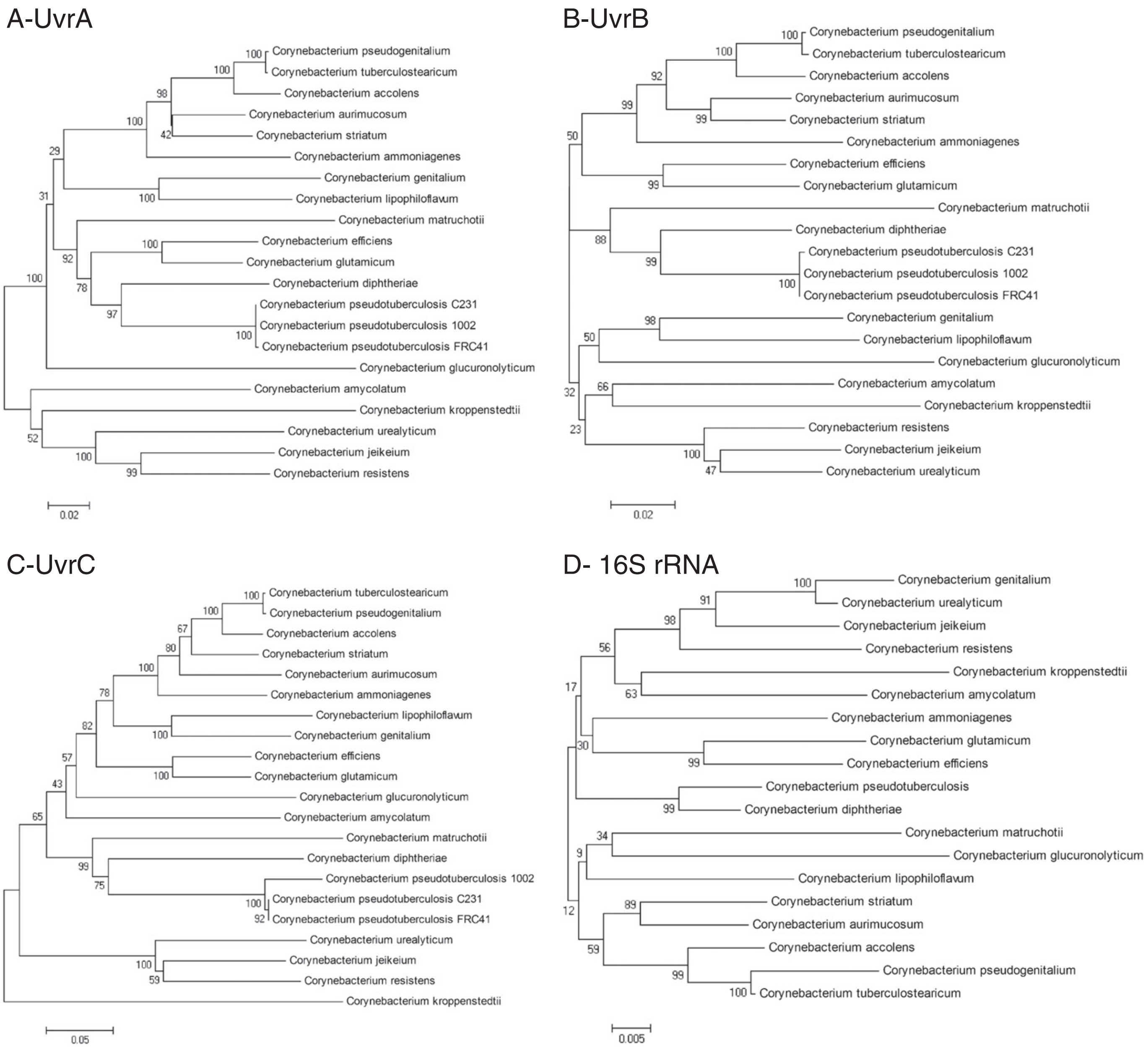

A-UvrA
100
Corynebacterium pseudogenitalium
100
Corynebacterium tuberculostearicum
98
Corynebacterium accolens
100
Corynebacterium aurimucosum
42
Corynebacterium striatum
29
Corynebacterium ammoniagenes
Corynebacterium genitalium
31
100
Corynebacterium lipophiloflavum
Corynebacterium matruchotii
100
Corynebacterium efficiens
92
Corynebacterium glutamicum
100
78
Corynebacterium diphtheriae
97
Corynebacterium pseudotuberculosis C231
Corynebacterium pseudotuberculosis 1002
100
Corynebacterium pseudotuberculosis FRC41
Corynebacterium glucuronolyticum
Corynebacterium amycolatum
Corynebacterium kroppenstedtii
52
Corynebacterium urealyticum
100
Corynebacterium jeikeium
99
Corynebacterium resistens
0.02
B-UvrB
100
Corynebacterium pseudogenitalium
100
Corynebacterium tuberculostearicum
92
Corynebacterium accolens
99
Corynebacterium aurimucosum
99
Corynebacterium striatum
50
Corynebacterium ammoniagenes
Corynebacterium efficiens
99
Corynebacterium glutamicum
Corynebacterium matruchotii
Corynebacterium diphtheriae
88
Corynebacterium pseudotuberculosis C231
99
Corynebacterium pseudotuberculosis 1002
100
Corynebacterium pseudotuberculosis FRC41
98
Corynebacterium genitalium
50
Corynebacterium lipophiloflavum
Corynebacterium glucuronolyticum
66
Corynebacterium amycolatum
32
Corynebacterium kroppenstedtii
23
Corynebacterium resistens
100
Corynebacterium jeikeium
47
Corynebacterium urealyticum
0.02
C-UvrC
100
Corynebacterium tuberculostearicum
100
Corynebacterium pseudogenitalium
67
Corynebacterium accolens
80
Corynebacterium striatum
100
Corynebacterium aurimucosum
78
Corynebacterium ammoniagenes
82
Corynebacterium lipophiloflavum
100
Corynebacterium genitalium
57
Corynebacterium efficiens
43
100
Corynebacterium glutamicum
Corynebacterium glucuronolyticum
65
Corynebacterium amycolatum
Corynebacterium matruchotii
99
Corynebacterium diphtheriae
75
Corynebacterium pseudotuberculosis 1002
100
Corynebacterium pseudotuberculosis C231
92
Corynebacterium pseudotuberculosis FRC41
Corynebacterium urealyticum
100
Corynebacterium jeikeium
59
Corynebacterium resistens
Corynebacterium kroppenstedtii
0.05
D- 16S rRNA
100
Corynebacterium genitalium
91
Corynebacterium urealyticum
98
Corynebacterium jeikeium
56
Corynebacterium resistens
Corynebacterium kroppenstedtii
17
63
Corynebacterium amycolatum
Corynebacterium ammoniagenes
30
Corynebacterium glutamicum
99
Corynebacterium efficiens
Corynebacterium pseudotuberculosis
99
Corynebacterium diphtheriae
34
Corynebacterium matruchotii
9
Corynebacterium glucuronolyticum
Corynebacterium lipophiloflavum
12
89
Corynebacterium striatum
Corynebacterium aurimucosum
59
Corynebacterium accolens
99
Corynebacterium pseudogenitalium
100
Corynebacterium tuberculostearicum
0.005

Fig. 1. Phylogenetic tree generated based on NER genes, *uvrA*(A)*, uvrB* (B) and *uvrC* (C) from *Corynebacterium* spp. The evolutionary history was inferred using the neighbor-joining method and NER enzymes of 21 *Corynebacterium* spp. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree. The evolutionary distances were computed using the Poisson correction method and the units are the number of amino acid substitutions per site. All positions containing gaps and missing data was eliminated from the dataset (complete deletion option). Phylogenetic analyses were conducted with MEGA4.

References

Antony, E., Hingorani, M.M., 2003. Mismatch recognition-coupled stabilization of Msh2-Msh6 in an ATP-bound state at the initiation of DNA repair. Biochemistry 42, 7682–7693.

Aravind, L., Roland, W.D., Eugene, V.K., 1999. Conserved domains in DNA repair proteins and evolution of repair systems. Nucleic Acids Res. 27, 1223–1242.

Brown, K.D., et al., 2003. The mismatch repair system is required for S-phase checkpoint activation. Nat. Genet. 33, 80–84.

Carvalho, F.M., et al., 2005. DNA repair in reduced genome: the Mycoplasma model. Gene 360, 111–119.

Cerdeno-Tarraga, A.M., et al., 2003. The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. Nucleic Acids Res. 31, 6516–6523.

Collins, M.D., Bradbury, J.F., 1986. Plant pathogenic species of *Corynebacterium.* Bergey's manual of systematic. Bacteriology 2, 276–1283.

D'Afonseca, V., et al., 2009. Survey of genome organization and gene content of *Corynebacterium pseudotuberculosis*. Microbiol. Res. 165 (4), 312–320.

Deb, J.K., Nath, N., 1999. Plasmids of *Corynebacterium*. FEMS Microbiol. Lett. 75, 11–20.

Dizdaroglu, M., 2005. Base-excision repair of oxidative DNA damage by DNA glycosylases. Mutat. Res. 591, 45–59.

Dorella, F.A., Pacheco, L.G.C., Oliveira, S.C., Miyoshi, A., Azevedo, V., 2006. *Corynebacterium pseudotuberculosis*: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet. Res. 37, 201–218.

Dos Vultos, T., Mestre, O., Tonjum, T., Gicquel, B., 2009. DNA repair in Mycobacterium tuberculosis revisited. FEMS Microbiol. Rev. 33, 471–487.

Doublie, S., Bandaru, V., Bond, J.P., Wallace, S.S., 2004. The crystal structure of human endonuclease VIII-like 1 (NEIL1) reveals a zincless finger motif required for glycosylase activity. Proc. Natl. Acad. Sci. USA 101, 10284–10289.

Drables, F., et al., 2004. Alkylation damage in DNA and RNA repair mechanisms and medical significance. DNA Repair 3, 1389–1407.

Fowler, R.G., et al., 2003. Interactions among the Escherichia coli mutt, mutM and mutY damage prevention pathways. DNA repair (Amst). 2, 159–173.

Fudou, R., 2002. *Corynebacterium efficiens* sp. nov., a glutamic-acid-producing species from soil and vegetables. Int. J. Syst. Evol. Microbiol. 52 (Pt 4), 1127–1131.

Goosen, N., Moolenaar, G.F., 2008. Repair of UV damage in bacteria. DNA Repair 7, 353–379.

Hall, M.C., Matson, W., 1999. The *Escherichia coli* MutL protein physically interacts with MutH and stimulates the MutH associated endonuclease activity. J. Biol. Chem. 264, 1306–1312.

Horst, J.P., Wu, T.H., Marinus, M.G., 1999. Escherichia coli mutator genes. Trends Microbiol. 7, 29–36.

Author's personal copy

Huffman, J.L., Sundheim, O., Tainer, J.A., 2005. DNA base damage recognition and removal: new twists and grooves. Mutat. Res. 577, 55–76.

Kalinowski, J., et al., 2003. The complete *Corynebacterium glutamicum* ATCC 13032 genome sequence and its impact on the production of L-aspartate-derived amino acids and vitamins. J. Biotechnol. 104, 5–25.

Lau, A.Y., Wyatt, M.D., Glassner, B.J., Samson, L.D., Ellenberger, T., 2000. Molecular basis for discriminating between normal and damaged bases by the human alkyladenine. glycosylase, AAG. Proc. Natl. Acad. Sci. USA 97, 13573–13578.

Lindahl, T., 2001. Keynote: past, present, and future aspects of base excision repair. Prog. Nucleic Acid Res. Mol. Biol. 68, 17–30.

Nakamura, Y., Nishio, Y., Ikeo, K., Gojobori, T., 2003. The genome stability in *Corynebacterium* species due to lack of the recombinational repair system. Gene 317, 149–155.

Nieminuszczy, J., Grzesiuk, E., 2007. Bacterial DNA repair genes and their eukaryotic homologues: 3. AlkB dioxygenase and Ada methyltransferase in the direct repair of alkylated DNA. Acta Biochim. Pol. 54, 459–468.

Nishio, Y., et al., 2003. Comparative complete genome sequence analysis of the amino acids replacements responsible for thermostability of *Corynebacterium efficiens*. Genome Res. 13, 1572–1579.

Razin, S., Yogev, D., Naot, Y., 1998. Molecular biology and pathogenicity of mycoplasmas. Microbiol. Mol. Biol. Rev. 62, 1094–1156.

Robertson, B., Klungland, A., Rognes, T., Leirosc, I., 2009. Base excision repair: the long and short of it. Cell. Mol. Life Sci. 66, 981–993.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.

Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc. Natl. Acad. Sci. USA. 101, 11030–11035.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.

Tauch, A., et al., 2005. Complete genome sequence and analysis of the multiresistant nosocomial pathogen *Corynebacterium jeikeium* K411, a lipid-requiring bacterium of the human skin flora. J. Bacteriol. 187, 4671–4682.

Tauch, A., et al., 2008a. Ultrafast pyrosequencing of *Corynebacterium kroppenstedtii* DSM44385 revealed insights into the physiology of a lipophilic *Corynebacterium* that lacks mycolic acids. J. Biotechnol. 136, 22–30.

Tauch, A., et al., 2008b. The lifestyle of *Corynebacterium urealyticum* derived from its complete genome sequence established by pyrosequencing. J. Biotechnol. 136, 11–21.

Trost, E., et al., 2010. Complete genome sequence and lifestyle of black-pigmented *Corynebacterium aurimucosum* ATCC 700975 (formerly C. nigricans CN-1) isolated from a vaginal swab of a woman with spontaneous abortion. BMC Genomics 11–91.

Wallace, S.S., Bandaru, V., Kathe, S.D., Bond, J.P., 2003. The enigma of endonucleaseVIII. DNARepair (Amst). 2, 441–453.

Xu, G., Herzig, M., Rotrekl, V., Walter, C.A., 2008. Base excision repair, aging and health span. Mech. Ageing Dev. 129, 366–382.

Zhang, Y., Rohde, L.H., Wu, H., 2009. Involvement of nucleotide excision and mismatch repair mechanisms in double strand break Repair. Curr. Genom. 10, 250–258.

Zuckerkandl, E., Pauling, L., 1965. Evolutionary divergence and convergence in proteins. Evol. Genes Prot. 1, 97–166.

6.3.2 Genômica subtrativa *in silico* para a identificação de alvos em patógenos bacterianos de seres humanos

A identificação do alvo é o primeiro passo no processo de descoberta de drogas e vacinas, sendo que a genômica subtrativa *in silico* é amplamente utilizada neste processo. Por meio desta abordagem, nos anos recentes, um grande número de alvos foram identificados em agentes patogênicos bacterianos que são resistentes a drogas ou para os quais nenhuma vacina adequada esteja disponível. O método *in silico* reduz o tempo, bem como o custo de rastreamento do alvo. Embora seja uma técnica poderosa que pode ser aplicada a uma vasta gama de agentes patogênicos, há muitas armadilhas na análise e interpretação dos dados. Revisou-se esta abordagem, incluindo metas que foram identificadas com esta técnica, incluindo vantagens e desvantagens. Discutiu-se também as nossas próprias experiências utilizando esta abordagem.

Este foi o primeiro trabalho na área com bactérias do gênero *Corynebacterium*, incluindo a *C. pseudotuberculosis*.

DRUG DEVELOPMENT RESEARCH 72:162–177 (2011)

Research Overview

In Silico Subtractive Genomics for Target Identification in Human Bacterial Pathogens

Debmalya Barh,[1,4]* Sandeep Tiwari,[2] Neha Jain,[2] Amjad Ali,[3]
Anderson Rodrigues Santos,[3] Amarendra Narayan Misra,[4] Vasco Azevedo,[3] and Anil Kumar[2]

[1]*Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur WB-721172, India*

[2]*School of Biotechnology, Devi Ahilya University, Khandwa Rd., Indore 452001, India*

[3]*Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CP 486, CEP 31270-901, Belo Horizonte, Minas Gerais, Brazil*

[4]*Department of Biosciences and Biotechnology, School of Biotechnology, Fakir Mohan University, Jnan Bigyan Vihar, Balasore, 756020 Orissa, India*

Strategy, Management and Health Policy				
Enabling Technology, Genomics, Proteomics	Preclinical Research	Preclinical Development Toxicology, Formulation Drug Delivery, Pharmacokinetics	Clinical Development Phases I-III Regulatory, Quality, Manufacturing	Postmarketing Phase IV

ABSTRACT Target identification is the first step in the drug and vaccine discovery process; *in silico* subtractive genomics is widely used in this process. Using this approach, in recent years, a large number of targets have been identified for bacterial pathogens that are either drug resistant or for which no suitable vaccine is available; most such reports concern a specific pathogen. The *in silico* method reduces the time as well as the cost of target screening. Although a powerful technique that can be applied to a wide range of pathogens, there are many pitfalls in the analysis and interpretation of the data. We review this approach, including targets that have been identified with this technique and various other aspects, including advantages and disadvantages. We also discuss our own experiences using this technology. Drug Dev Res 72:162–177, 2011. © 2010 Wiley-Liss, Inc.

Key words: drug target; essential genes; subtractive genomics; bacterial pathogen

INTRODUCTION

Although high-throughput techniques and synthetic chemistry are an integral part of today's drug discovery process, accelerating the process manifold, the introduction of a new drug on the market still takes 10–15 years and therefore requires a huge investment [Plotkin, 2005]. Technological advancements, along with improved and innovative strategies, could reduce the cost and the time required to develop a new drug.

Most infectious diseases are caused by bacterial pathogens. An increase of 58% in the mortality rate due to such infectious diseases has been reported from 1980 to 1992 in the United States [Pinner et al., 1996]. According to the 2004 World Health Organization Report [www.who.int/whr/2004/annex/topic/en/annex_2_en.pdf], 16.4 million people died worldwide in that year from bacterial infectious diseases. Although several antibiotics are currently available for each bacterial pathogen, the emerging drug-resistant strains of such pathogens make them difficult to control,

*Correspondence to: Debmalya Barh, Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology (IIOAB), Nonakuri, Purba Medinipur WB-721172, India. E-mail: dr.barh@gmail.com

Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/ddr.20413

© 2010 Wiley-Liss, Inc.

perhaps due to decades-old uses of such drugs in human patients [Arias and Murray, 2009].

Drug target identification is the first step in the drug discovery process [Chan et al., 2010]. Because of the availability of both pathogen and host–genome sequences, it has become easier to identify drug targets at the genomic level for any given pathogen [Allsop, 1998; Stumm et al., 2002; Meinke et al., 2004; Owa, 2007]. In recent years, the strategies are shifting progressively from a generic approach to genomic and metabolomic approaches [Ishii et al., 2004; Lin and Qian, 2007] to identify novel drug targets that are required to design new defenses against antibiotic-resistant pathogens [McDevitt and Rosenberg, 2001; Mills, 2006; Fischbach and Walsh, 2009]. Tremendous advancements have been made in target identification and drug discovery since the human genome sequence became available [Lander et al., 2001; Venter et al., 2001].

Use of computational approaches, with integrated genomics, proteomics, transcriptomics, interactomics, signalomics, and metabolomics, are current trends in target discovery for most human diseases, especially for cancer, cardiovascular, neuroendocrine, and infectious diseases; they make the discovery process faster and more cost effective. Currently, genomics and more specifically *in silico* comparative, subtractive, and functional genomics are being widely used to identify novel drug and vaccine targets in order to develop effective antibacterial agents and vaccines against bacterial pathogens that are either resistant to existing antibacterial regimens or for which a suitable vaccine is not available [Ji, 2002; Pucci, 2006].

Computational metabolic flux modeling, along with systems approaches, have been found to be a great aid for understanding and manipulating microbial metabolism [Downs, 2006; Thykaer et al., 2009]. They can help in the identification of key essential or survival proteins (the targets) of the organism that can be inhibited by using appropriate lead molecule(s) identified by *in silico* virtual screening. Additionally, *in silico* comparative genomics-based subtraction analysis using host and pathogen genomes is a powerful approach for the identification of genus- or species-specific genes, or groups of genes that are responsible for a unique phenotype as well as the virulence factors of the pathogen [Huynen et al., 1997, 1998]. It is then necessary to determine whether these genes are essential survival genes of the pathogen and whether there are non-host homologues. Simultaneously, metabolic pathway subtraction is required to identify metabolic pathways that host and pathogen have in common and pathogen-specific pathways. Once the essential non-host homologue survival genes of a pathogen are identified, they need to be allocated to known pathways. If an essential non-host homologue survival gene is found crucial in any of the pathogen's metabolic pathways, it is considered a putative target. To identify a vaccine target, additional analyses, such as localization, antigenicity, and membrane topology, are required in order to design epitopes.

In this review we present an overview of *in silico* subtractive genomics approaches used to identify genomic targets in various human pathogenic bacteria, along with information from our own experiments using this approach. We also discuss various aspects, including advantages, disadvantages, and future prospectives for this approach.

STRATEGIES FOR SUBTRACTION-BASED TARGET DISCOVERY

Concept of Essential Non-Host Homologue Genes

Subtractive genomics-based target identification is based on essential genes and the non-host homologue. Essential genes are genes that are required for growth, adaptability and survival of an organism. Therefore, deficiency of any such gene should be lethal to the organism. Essential genes are likely to have a common function across all organisms [Mushegian and Koonin, 1996]. Often such essential genes are evolutionally conserved in different taxa [Itaya, 1995; Tatusov et al., 1997; Koonin et al., 1998; Jordan et al., 2002; Kobayashi et al., 2003]. Essential genes can be identified through random mutagenesis of bacterial genomes [Hood, 1999]. The Database of Essential Genes (DEG) [Zhang et al., 2004] is the main resource that lists experimentally validated essential genes in bacteria, fungi, plants, and animals. DEG can readily be used for target identification through comparative and subtractive genomics approaches. A non-host homologue (not present in the host but present in the pathogen) essential gene of a pathogen is considered a good target against the pathogen [Sakharkar et al., 2004]. Ideally, a target should fulfill four properties: (1) it must be an essential gene for survival or pathogenesis of the target organism; (2) druggability, i.e., having protein structure characteristics that make it amenable to bind to small inhibitor molecules; (3) functional and structural characterization, with established assays for screening small molecule inhibition; and (4) distinctness from current drug targets to avoid cross-resistance [Holman et al., 2009]. Both experimental and computational methods are available for essential gene-based target prediction; however, the computational methods are preferable as they require less time, labor, and are less expensive [Itaya, 1995; Kobayashi et al., 2003].

Genome Subtraction

Subtraction literally means "removed from below," more precisely, taking a smaller piece from a larger one. It is a mathematical approach to determine the difference between two amounts in the same category. Subtractive genomics is based on a comparative genomics approach; generally, we use two genomes and subtract the genomic data set of one from the other to obtain genus-, species-, and unique phenotype-specific genes. In target identification, the pathogen genome, within which target(s) have to be identified, is subtracted from the host genome, and subtracted genome sequences or genes (non-host homologues) are further analyzed to determine whether they are essential for pathogen survival. These essential and non-host homologues must be a critical component in vital physicochemical and metabolic pathways, so that a designed drug or a lead compound specific to such target(s) will only impact on the pathogen's system, without hampering host physiology or any aspect of host biology. Identified targets may be used to design and develop drugs, vaccines, or dual-purpose targets [Sakharkar et al., 2004; Dutta et al., 2006; Barh and Kumar, 2009; Barh et al., 2009]. In general, enzyme targets located in the cytoplasm are good candidates for drug development; exo-membrane (surface-exposed) and secreted protein targets, based on their antigenicity, can be used for peptide vaccine design. Exo-membrane enzyme or transporter targets are most suitable for dual purpose [Barh et al., 2009].

METHODS FOR IN SILICO SUBTRACTIVE GENOMICS FOR TARGET DISCOVERY

Subtractive genomics provides new opportunities for finding optimal targets among unexplored cellular functions, based on an understanding of related biological processes in bacterial pathogens and their hosts [Dong et al., 2009]. The *in silico* method follows a similar strategy of subtractive hybridization, suppressive subtractive hybridization, positional cloning, and comparative genomics that can be used for the identification of drug targets in the wet lab. This strategy was first applied to *Helicobacter pylori* [Huynen et al., 1997, 1998]. A differential genome display approach was used in this case; it relies on the fact that parasitic bacterial genomes are smaller and encode fewer proteins than a closely related free-living bacterial organisms. Hence, genes which are present in the parasitic bacterium, but absent in closely related free-living taxa, are responsible for adaptability and pathogenicity and therefore may be considered candidate targets. This strategy has evolved over time and has become much faster and more sensitive as a result of the availability of the complete genome sequence of several pathogenic bacteria, improved computational tools, and various databases. The efficiency of this method was further boosted manifold with the development and availability of DEG. This approach was successfully used for the first time to identify essential genes and targets in *Pseudomonas aeruginosa* by Saharker et al. [2004], using DEG. Since then, this approach has been widely applied with slight modifications to identify targets in several pathogenic bacteria, including *P. aeruginosa* [Sakharkar et al., 2004; Perumal et al., 2007], *H. pylori* [Dutta et al., 2006], *B. pseudomallei* [Chong et al., 2006], *A. hydrophila* [Sharma et al., 2008], *N. gonorrhoeae* [Barh and Kumar, 2009], *N. meningitides* [Sarangi et al., 2009], *M. tuberculosis* [Asif et al., 2009], *S. typhi* [Rathi et al., 2009], *M. leprae* [Shanmugam and Natarajan, 2010], and *M. pneumonia* [Gupta et al., 2010]. Table 1 lists bacterial pathogens affecting humans to which this strategy has been applied in order to identify targets in these pathogens.

Current Methodology

The NCBI Genome database (www.ncbi.nlm.nih.gov/genome), the Swiss-Prot protein database (http://us.expasy.org/sprot) [Bairoch and Apweiler, 1997], DEG (http://tubic.tju.edu.cn/deg), KEGG [Ogata et al., 1999], BLAST tools (http://blast.ncbi.nlm.nih.gov/Blast.cgi), VFDB [Chen et al., 2005], cellular localization prediction tools, such as CELLO [Yu et al., 2004], PSLpred [Bhasin et al., 2005], PSORTb [Gardy et al., 2005], and SOSUI-GramN [Imai et al., 2008], are integral parts of current subtractive genomics-based bacterial target identification strategies. In general, the host (human) and the pathogen (in which the target is to be identified) genomes and proteomes are collected from the NCBI genome server. The pathogen genome is then subjected to NCBI human BLAST to subtract the non-human homologous genes of the bacteria. Each identified non-human homologue gene and protein sequence of the pathogen is then subjected to BLASTx and BLASTp, using the bacterial BLAST option in DEG. A BLAST hit with significant cutoff values against any bacterial sequence listed in DEG gives an indication that the query sequence of the bacteria under study is a putative essential gene in the organism. Identified putative, essential non-human homologues genes (targets) are then mapped in metabolic pathways (pathogen unique and host–pathogen common) in which they are involved, using comparative pathway analysis for humans and the pathogen, available in the KEGG database. Essential non-human homologues that are crucial in pathways are identified and subsequently analyzed to determine their localization (cytoplasmic, membrane, exo-membrane, or secreted), using appropriate localization prediction tools, and enzymatic activity-related information is

TABLE 1. Human Bacterial Pathogens to Which In Silico Subtractive Genomics Strategy Has Been Applied to Identify Drug Targets*

Sl. No	Pathogens	Disease caused	Associated diseases	Prevalent countries	Pathogenesis and epidemiology	Available drugs/ antibiotics	Available vaccines	Challenges in vaccine/drug development
1	*M. tuberculosis*	Tuberculosis		Russia, Israel, China, Asia, Africa	An infectious disease that affects lungs and kills young and middle-aged adults faster than any other disease	Isoniazid, rifampicin, ethambutol, streptomycin	Bacille Calmette–Guérin (BCG)	The vaccine is efficient in preventing the disease, but the efficacy in adults is doubtful. Worldwide emergence of extensively drug-resistant tuberculosis is a serious challenge
2	*M. leprae*	Leprosy		Central Africa, Southeast Asia	An infectious disease that primarily affects the skin, mucous membranes, and peripheral nerves causing deformities. Estimated number of existing leprosy patients in the world is 12–15 million	Quinolones, refampicin, dapsone, ofloxacin	BCG	Multi-drug resistance of the strain hinders the use of antibiotics against the pathogen. Biology of the pathogen is poorly understood, hence effective drug discovery is not a priority
3	*H. pylori*	Gastric ulcer	Coronary artery, liver diseases, and MALT-type lymphoma	US, China, Korea, developing countries	Highly infectious and present in approximately 50% of world's population. Pathogen is of serious concern in developing countries. Infected individuals are at high risk of gastric cancer	Clarithromycin, rifabutin, furazolidone	*H. pylori* whole-cell (HWC) vaccine	No improved vaccine available but new vaccines are under development. Multi-drug-resistant species are predominant
4	*V. cholerae*	Endemic and epidemic cholera, secretory diarrhea		South and Central America, Asia	Highly infectious water-born disease that causes rapid loss of body fluids, leads to dehydration and shock. Without treatment, death can occur within hours. Severe cases require intravenous fluid replacement	Tetracycline, azithromycin, ciprofloxacin, erythromycin	Dukoral, Mutacol	No long-term effective vaccine available. Requires new vaccine against the pathogen. Tetracycline and Ciprofloxacin are used, but many resistant strains are reported
5	*N. gonorrhoeae*	Gonorrhoea	Conjunctivitis, pharyngitis, proctitis, prostatitis, and orchitis	US and in underdeveloped countries	Most common sexually transmitted diseases. Every year about sixty million new cases are reported. Severity causes infertility, PID, and ectopic pregnancy	Cefotaxime, cefoperazone, moxalactam, piperacillin, mezlocillin	Not available	The pathogen is highly adaptive and antibiotic resistant
6	*A. hydrophila*	Water-associated traumatic secondary wound infection	Septicaemia, cellulitis, pneumonitis, necrotizing fasciitis, and gastroenteritis	US	Infects through contaminated refrigerated animal products. Causes food poising. It can be fatal if untreated	Cefotaxime, cephalosporins	Not available	The pathogen is resistant to chlorination of water and to a variety of antibiotics

TABLE 1. Continued

Sl. No	Pathogens	Disease caused	Associated diseases	Prevalent countries	Pathogenesis and epidemiology	Available drugs/ antibiotics	Available vaccines	Challenges in vaccine/drug development
7	*B. pseudomallei*	Melioidosis, septicemia	Acute pulmonary infection, subacute and chronic diseases	Tropical Australia, Southeast Asia, East Asia and northern Australia, northeastern Thailand, Brazil	The pathogen is a potential bioterrorism agent; mortality from melioidosis septic shock remains high despite appropriate antimicrobial therapy	Ceftazidime, chloramphenicol, doxycycline, trimethoprim-sulfamethoxazole	Not available currently but in under-developing stage	Antibiotics are recommended, but death rate is >40% of treated patients due to drug-resistant strains
8	*P. aeruginosa*	Pneumonia, septicemia, urinary tract infection, gastrointestinal, and skin and soft tissue infections,	Chronic lung infection of cystic fibrosis, and contact lens-associated pseudomonal keratitis	Europe, Germany, Bulgaria, Malta	Leading cause of nosocomial infections. An important pathogen among debilitated, burned, and immunocopromized individuals	Amikacin, aminoglycoside, piperacillin, fluoroquinolones	Whole-cell *P. aeruginosa* vaccine	Though there are many drugs available but the pathogen exhibits multi-drug resistance
9	*S. typhi*	Typhoid fever		Morocco, Algeria, Tunisia, Libya, Egypt, England	Is one of the most highly host-adapted pathogens; 150 to 300 deaths occur each year in the UK	Ciprofloxacin, trimethoprim	M-01ZH09	Multi-drug resistance is of great concern. Vaccine is in an experimental stage
10	*N. meningitides*	Meningitis	Meningococcemia	America, Asia, Africa	2500 to 3500 cases reported every year in US. Children aged <5 years are at greatest risk. Severity leads to shock and death	Ceftriaxone, rifampicin	MeNZBTM vaccine, Anti-MenB vaccine	Ceftriaxone is more effective and cheaper than rifampicin, but its acceptability by patients may limit its use as a first-line prophylactic agent
11	*C. perfringens*	Gangrene and gastrointestinal disease (food poisoning and necrotic enteritis)	Enterocolitis, dysenteria, and enterotoxemia	Underdeveloped and developing countries as well as in some parts of UK	It is the most prolific producer of toxins	Clindamycin, penicillin, metronidazole	Not available	Strain is resistant to penicillin, erythromycin, and chloramphenicol
12	*M. pneumoniae*	Pneumonia	Meningoencephalitis	Denmark, US, Europe, Japan	Frequently causes community-acquired respiratory infections in children and adults. Severity of illness may vary from severe pneumonia to asymptomatic infection	Clarithromycin, quinolone, tetracyclines, macrolides, ketolides	Not available	No specific vaccine is available until now. Polysaccharides or whole attenuated cell-based vaccines are of limited scope. The biology and molecular pathogenesis is suntil unknown to a certain extent
13	*S. pneumoniae*	Pneumonia	Meningitis, otitis media, and sinusitis	US (Atlanta)	The disease is fatal if untreated. Incidence of disease is highest in children <2 years of age and in adults >65 years of age	Vancomycin, cefotaxime, meropenem, trimethoprim-sulfamethoxazole, clindamycin	Heptavalent protein-polysaccharide conjugate vaccine, 23-valent capsular polysaccharide vaccines	Multi-drug-resistant strains of bacteria have complicated treatment approaches

*Also indicates the diseases caused by the pathogen, their prevalence, epidemiology, available drugs, and vaccines against the pathogen, and challenges for control of the infection and treatment.

collected from www.expacy.org. Both the localization and enzyme-related information may also be collected from Swiss-Port (if available). Cytoplasmic and membrane channel proteins are selected for drug targeting, whereas membrane, exo-membrane, and secreted proteins are used to design peptide vaccines. The overall approach is shown in Figure 1. Drug and vaccine targets for several human bacterial pathogens have been reported, using this innovative strategy. Table 2 presents a list of pathogens along with the applied cutoff values for BLAST.

Our Strategy

In a modified approach that we developed [Barh and Kumar, 2009; Barh and Misra, 2009], we first screen the essential genes of the pathogen using DEG and then identify the non-human homologues to reduce the number of BLASTs. We also use pathway subtraction instead of using all pathways present in host and pathogen. Our pathway analysis-based target identification is based on criteria such as: (1) the target must be an essential non-host homologue; (2) the target must be a core gene of the pathogen; (3) the pathogen's unique pathway related targets are more favorable and will be superior if the target is involved in multiple pathways; (4) pathways having multiple targets are superior to those having single targets; (5) in the case of enzyme targets in host–pathogen common pathways, it should not be of the same class of protein, and the EC. no. of the target should not match that of any protein product

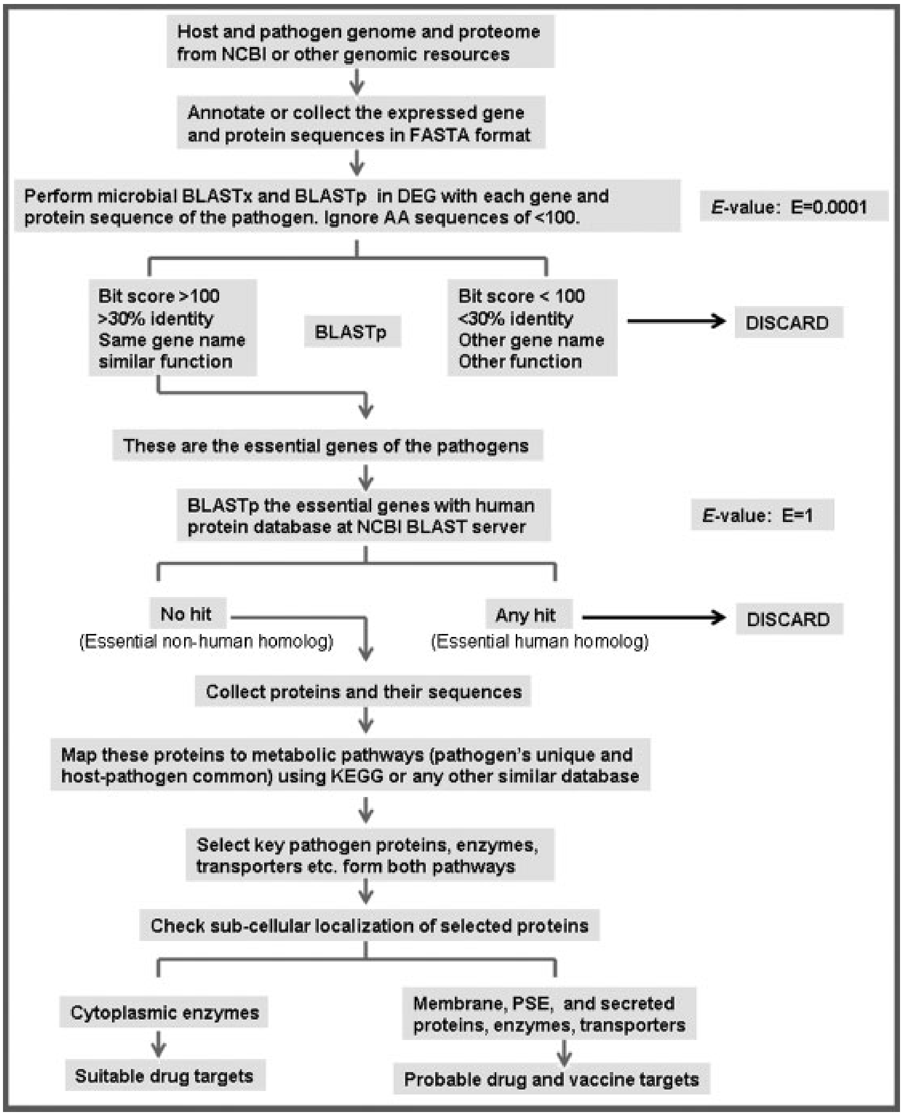

Host and pathogen genome and proteome from NCBI or other genomic resources
Annotate or collect the expressed gene and protein sequences in FASTA format
Perform microbial BLASTx and BLASTp in DEG with each gene and protein sequence of the pathogen. Ignore AA sequences of <100.
E-value: E=0.0001
Bit score >100
>30% identity
Same gene name
similar function
BLASTp
Bit score < 100
<30% identity
Other gene name
Other function
DISCARD
These are the essential genes of the pathogens
BLASTp the essential genes with human protein database at NCBI BLAST server
E-value: E=1
No hit
(Essential non-human homolog)
Any hit
(Essential human homolog)
DISCARD
Collect proteins and their sequences
Map these proteins to metabolic pathways (pathogen's unique and host-pathogen common) using KEGG or any other similar database
Select key pathogen proteins, enzymes, transporters etc. form both pathways
Check sub-cellular localization of selected proteins
Cytoplasmic enzymes
Suitable drug targets
Membrane, PSE, and secreted proteins, enzymes, transporters
Probable drug and vaccine targets

Fig. 1. Schematic representation of steps involved in *in silico* subtractive genomics-based target identification in bacterial pathogens. Identified targets can be used to develop drugs or vaccines, depending on their localization, exo-membrane topology, or secreted protein properties.

TABLE 2. Pathogens, Along With the Applied Cutoff Values for BLAST*

			BLASTp cutoff at amino acid (AA) level								
Sl. No.	Pathogen name	Genes in genome	Essential gene prediction (DEG)	Non-human homologue (NCBI)	No. of essential genes	No. of non-human homologues	No. of drug targets	No. of cytoplasmic targets	No. of membrane targets	Suggested targets	References
1	*B. pseudomallei*	5,855	E-value = 10^{-10}, 30% identity.	E-value = 10^{-3}, 30% identity	312	3,723	312	79.2%	20.8%	rpoE, OmpR	Chong et al. [2006]
2	*H. pylori*	1,590	E-value = 10^{-100}, bit score > 100, AA length > 100	E-value = 10^{-10}	178	40	40	30	10	rlpA, fecA fecA, dppA, nhaA, dppC, ftsX	Dutta et al. [2006]
3	*M. pneumoniae* (M129 strain)	693	E-value = 10^{-100}, bit score > 100, AA length > 100	E-value = 10^{-3}	220	375	112		12		Gupta et al. [2010]
4	*P. aeruginosa*	5,567	E-value = 10^{-3}	E-value = 10^{-10}	306	3,841				Genes Involved in transport of small molecules. Translation, post-translational modification and degradation	Sakharka et al. [2004]
5	*S. typhi*	4718	E-value = 10^{-4}, bit score > 100	E-value = 10^{-100}	300	149	149	138	11	ddl, trpB. motA, CheR, ppc	Rathi et al. [2009]
6	*M. leprae*	2,770	NA	E-value = 10^{-4}		179	62			Alr, rmlC, murC, murD, murE, murF, murG, murY	Shanmugam and Natarajan [2010]
7	*N. meningitides* (serogroup B)	2,001	E-value = $10^{-10,}$ 30% identity, bit-score > 100	E-value = 10^{-4}	362	1,413	35	26	9	Ppc, PilF, trpA, rpB, trpC, trpD, trpE	Sarangi et al. [2009]
8	*N. gonorrhoeae* (FA 10990)	2,002	E-value = $10^{-10,}$ 35% identity, bit-score > 100 AA length > 100	E-value = 10^{-10}	537	106	106	67	40	alf/tsr, ptsN, ddl, TbpA, afuB/fbpB ComL, cysW, PilF, pilV	Barh and Kumar [2009]
9	*A. hydrophila* (ATCC 7966)	4,287	E-value = $10^{-10,}$ bit-score > 100 AA length > 100	E-value = $10^{-10,}$ bit-score > 100	379	2,047	87			ddl, alr, Uroporphyrinogen-III synthase, glutathione S-transferase, biotin synthase	Sharma et al. [2008]
10	*Brugia malayi*	805		E = 10^{-25}	250						Holman et al. [2009]
11	*M. tuberculosis*	3,989	NA	NA	628	304	135			Genes related with Amino-acid biosynthesis	Asif et al. [2009]
12	*S. pneumoniae*	2,355		E-value cutoff < 0.005			161			Genes related with metabolism and cell wall biosynthesis	Singh et al. [2007]
13	*C. perfringens*	2,558	E-value = $10^{-10,}$ bit-score > 100 AA length > 100	E-value = $10^{-10,}$ bit-score > 100 AA length > 100	726	426				ABC transporter-ATP binding protein, FtsZ, RpoD, 50S ribosomal protein L13, and 30S subunit S5	Chhabra et al. [2010]

*Cutoff values for screening essential genes and non-human homologous genes are given separately. Genome size, number of essential genes, number of non-human homologues, number of cytoplasmic and membrane localized targets, and important targets are also presented for each pathogen.

of the host; and (6) pathogenic island-related or virulence proteins are considered superior targets.

TOOLS DEVELOPED FOR GENOME SUBTRACTION

Initial Approaches

The entire process of genome subtraction can be carried out using an *in silico* approach; to our knowledge, only two complementary *in silico* methods have been developed that allow genome subtraction. These are based on computed clusters of homologous proteins or on pairwise protein comparisons. First, all proteins of a sequence database, including those of complete genomes, are compared with each other using similarity search software, such as BLAST [Altschul et al., 1997, 1990] or FASTA [Pearson and Lipman, 1988]. Corresponding search outputs are then processed according to default constraints to extract significant hits. Finally, protein families are constructed using single transitive links. If proteins A and B are similar according to the constraints, and proteins B and C are also similar, proteins A, B, and C are then stored in the same cluster. Software tools and databases, such as CluSTr [Kriventseva et al., 2001], COG [Tatusov et al., 2001], Hobacgen [Perrière et al., 2000], ProtoMap [Yona et al., 1999], and Systers [Krause et al., 2002] provide access to such sets of homologous proteins. However, COG contains a tool called the "phylogenetic pattern search," which allows genome subtraction to select protein families. The second approach does not use fixed constraints. The user declines the similarity thresholds to decide whether a coding sequence is present or absent in a genome. The software Seebugs belongs to this category; it is based on a protein sequence comparison, using the FASTA program [Bruccoleri et al., 1998].

Current Approach-Based Tools

The target identification method involves a number of steps; therefore we need to develop bioinformatics tools that can perform the entire process on a single platform. Bruccoleri et al. [1998] developed a simple but efficient *in silico* tool that can predict putative targets based on subtraction of conserved sequences of essential genes in user-specified genomes. To develop an automated computational tool, FindTarget was built based on BLASTp comparative proteomes [Chetouani et al., 2001]. However, this tool cannot perform the entire process. In a further advancement, Singh et al. [2006] designed the T-iDT tool, which finds essential bacterial genes as well as non-human homologues, by using DEG and a human protein database. This tool can predict both the essential genes and potential targets in a pathogen genome at the same time. However, a pathway-based approach is not integrated into this tool. Recently, efforts have been made to integrate several parameters to enhance the efficacy of the prediction. The mGenomeSubtractor is one of such tools; it performs a rapid analysis of core, accessory, and essential genes, virulence factors, species-specific genes, and targets, using a mpiBLAST-based *in silico* subtractive genome hybridization method [Shao et al., 2010]. This tool can be accessed from http://bioinfo-mml.sjtu.edu.cn/mGS/. A list of available tools and databases useful for subtractive genomics-based bacterial target identification is given in Table 3.

Although in recent years several targets have been reported from various pathogenic bacteria, using genomic subtraction, no database has listed all such targets, except the Genomic Target Database (GTD) (www.iioab.webs.com/GTD.htm), which we started to develop in 2009 [Barh et al., 2009]. This database is a readily available resource that can be used to design mutagenesis studies to validate essential genes as well as the targets of pathogens listed in the database. However, currently, the database is not enriched with all targets available in the literature, as the number of reported pathogens and their targets is huge.

IDENTIFIED TARGETS

Target identification using subtractive genomics has generated a large number of targets from various pathogens. As this method is based on comparative genomics and DEG is most commonly used, several targets are found to be common in many bacteria; however, species or strain-specific, and novel targets have also been reported [Barh and Kumar, 2009; Sarangi et al., 2009]. Even though metabolic pathways related to both cytoplasmic and membrane-associated targets have been used to design both drugs and vaccines for many pathogens [Sakharkar et al., 2004; Sharma et al., 2008; Barh and Kumar, 2009], in some cases membrane-localized and secreted proteins were focused to identify potential vaccine targets [Dutta et al., 2006; Barh and Misra, 2009]. We have given more importance to targets related to pathways unique to bacteria [Barh and Kumar, 2009]; others have given equal importance to all targets [Sakharkar et al., 2004; Sharma et al., 2008]. Targets related to pathogens' unique pathways (e.g., D-alanine metabolism, two-component system, type II and III secretion systems, bacterial chemotaxis, and lipopolysaccharide and peptidoglycan biosynthesis) are ~60–70% in common, regardless of the genotype of the pathogen, and ~30–40% of the targets are genus or strain specific. Table 4 presents a list of targets from pathways unique to bacteria. The number of targets reported in the literature from host–pathogen common pathways is

TABLE 3. Databases and Tools Used in Subtractive Genomics-Based Bacterial-Target Identification†

	Utility	Website	References
Database			
NCBI bacterial genomes	Recourse of bacterial genomes	http://www.ncbi.nlm.nih.gov/genomes/genlist.cgi?taxid = 2&type = 0&name = Complete%20Bacteria	
GOLD: Genomes	Recourse of genome projects	http://www.genomesonline.org/	Bernal et al. [2001]
Swiss-port	Proteome database	http://www.expasy.org/sprot/	Bairoch and Apweiler [1997]
Database of Essential Genes (DEG)	Screening of essential genes	http://tubic.tju.edu.cn/deg/	Zhang et al. [2004]
Kyoto Encyclopedia of Genes and Genomes (KEGG)	Pathway comparison and subtraction	http://www.genome.jp/kegg/	Ogata et al. [1999]
Genomic Target Database (GTD)	List of bacterial targets based on subtractive genomics	www.iioab.webs.com/GTD.htm	Barh et al. [2009]
Virulence Factors of Pathogenic Bacteria Database (VFDB)	Resource of virulence factors of various medically significant bacterial pathogens	http://www.mgc.ac.cn/VFs/main.htm	Chen et al. [2005]
Tools			
CELLO	Subcellular localization prediction for bacteria and eukaryotes	http://cello.life.nctu.edu.tw/	Yu et al. [2004]
PSORTb	Subcellular localization prediction for gram-negative and gram-positive bacterial proteins	http://www.psort.org/psortb/	Gardy et al. [2005]
SOSUI-GramN	Subcellular localization prediction for gram-negative bacterial proteins	http://bp.nuap.nagoya-u.ac.jp/sosui/sosuigramn/sosuigramn_submit.html	Imai et al. [2008]
PSLpred	Subcellular localization prediction for gram-negative bacterial proteins	http://www.imtech.res.in/raghava/pslpred/	Bhasin et al. [2005]
NCBI human BLAST	Subtraction of non-human homologue genes	http://www.ncbi.nlm.nih.gov/genome/seq/BlastGen/BlastGen.cgi?taxid = 9606	Altschul et al. [1990]
FindTarget	Subtractive genomics (link is not working)	http://bioweb.pasteur.fr/seqanal/findtarget	Chetouani et al. [2001]
T-iDT	Platform for identification of subtractive genomics based essential non-human homologue (link is not working)	http://www.milser.co.in/research.htm	Singh et al. [2006]
mGenomeSubtractor	in silico subtractive hybridization	http://bioinfo-mml.sjtu.edu.cn/mGS/	Shao et al. [2010]
SignalP*	Signal peptide prediction	http://www.cbs.dtu.dk/services/SignalP/	Bendtsen et al. [2004]
TMHMM*	Transmembrane domain prediction	http://www.cbs.dtu.dk/services/TMHMM/	Krogh et al. [2001]
LipoP*	Lipoprotein prediction	http://www.cbs.dtu.dk/services/LipoP/	Juncker et al. [2003]
SurfG*	Bacterial protein subcellular localization	http://genome.jouy.inra.fr/surfgplus/	Barinov et al. [2009]

†FindTarget and T-iDT web addresses mentioned in the references are currently not working. Tools used in reverse vaccinology are marked with an asterisk (*).

higher than the number of unique pathway targets. Some important targets from such common pathways are listed in Table 5.

ADVANTAGES OF IN SILICO SUBTRACTIVE GENOMICS FOR TARGET DISCOVERY

The importance of *in silico* subtractive genomics in drug-target identification is a function of its rapid and cost-effective screening of targets at the genome level. It also shortens the time required to develop immunomics-based antigens and thereby speeds up peptide vaccine design [Barh et al., 2010a,b]. Another major advantage is identification of putative essential genes in pathogens, which can be validated via mutagenesis studies [Sakharkar et al., 2004]. GTD has been developed with subtractive genomics-based targets to

TABLE 4. Selected Targets From Pathogen-Specific Metabolic Pathways*

	Pathways unique to bacteria	Genes	EC No.	Localization
1	Bacterial chemotaxis			
	Methyltransferase PilK	*pilK*	2.1.1.80	Cytoplasm
	Two-component sensor PilS	*pilS*	2.7.3.-	Membrane
	Chemotaxis-specific methylesterase		3.1.1.61	Cytoplasm
	Sensor histidine kinase		2.7.13.3	Cytoplasm
2	Polyketide sugar unit biosynthesis			
	Glucose 1-phosphate thymidylyltransfease	*rmlA*	2.7.7.24	Cytoplasm
	dTDP-D-Glucose 4,6 dehydratase	*rmlB*	4.2.1.46	Cytoplasm
	dTDP-4-dehydrorhamnose 3,5 epimerease	*rmlC*	5.1.3.13	Cytoplasm
	dTDP-4-dehydrorhamnose reductase	*rmlD*	1.1.1.133	Cytoplasm
3	Lipopolysaccharide biosynthesis			
	Probable glucosyltransferases		2.4.-	Cytoplasm
	3-deoxy-manno-octulosonate cytidylyltransferase	*kdsB*	2.7.7.38	Cell wall
	Putative 3-deoxy-D-manno-octulosonate 8-phosphate phosphatase		3.1.3.45	Cytoplasm
	Tetraacyldisaccharide 4′-kinase	*lpxK*	2.7.1.130	Cell wall
	Lipid A-disaccharide synthase	*lpxB*	2.4.1.182	Cytoplasm
	Lipopolysaccharide core biosynthesis protein WaaP	*waaP*	2.7.-.-	Cytoplasm
	Poly(3-hydroxyalkanoic acid) synthase 1	*phaC1*	2.3.1.-	Cytoplasm
	UDP-3-*O*-[3-hydroxymyristoyl] glucosamine *N*-acyltransferase		2.3.1.-	Cytoplasm
	UDP-glucose:(heptosyl) LPS α 1,3-glucosyltransferase WaaG	*waaG*	2.4.1.-	Cytoplasm
	UDP-2,3-diacylglucosamine hydrolase		3.6.1.-	Cytoplasm
	UDP-3-*O*-acyl-*N*-acetylglucosamine deacetylase	*lpxC*	3.5.1.-	Cytoplasm
	UDP-*N*-acetylglucosamine acyltransferase	*lpxA*	2.3.1.129	Cytoplasm
	Putative sugar kinase/ADP heptose synthase	*rfaE*	2.7.1.-	Cytoplasm
	Lipopolysaccharide heptosyltransferase I	*rfaC*	2.4.-.-	Cytoplasm
	ADP-heptose–LPS heptosyltransferase II	*rfaF*	2.4.-.-	Cytoplasm
	ADP-L-glycero-D-mannoheptose 6-epimerase	*rfaD*	5.1.3.20	Cytoplasm
	2-dehydro-3-deoxyphosphooctonate aldolase (KDO 8-P-synthase)	*kdsA*	2.5.1.55	Cytoplasm
4	D-alanine metabolism			
	D-alanine-D-alanine ligase A	*Ddl*	6.3.2.4	Cell wall
	Biosynthetic alanine racemase	*Alr*	5.1.1.1	Cytoplasm
5	Carbon fixation in photosynthetic organisms			
	Fructose-1,6-bisphosphate aldolase	*alf/tsr*	4.1.2.13	Cytoplasm
6	Two-component system			
	Nitrite two-component system transcriptional response regulator	*narL*		Intracellular
	Two-component sensor PilS	*pilS*	2.7.3.-	Membrane
	Probable 2-(5′-triphosphoribosyl)-3′-dephosphocoenzyme-A synthase		2.7.8.25	Cytoplasm
	Serine protease MucD precursor	*mucD*	3.4.21.-	Cytoplasm
	Probable acyl-CoA thiolase		2.3.1.9	Cytoplasm
	Glutamine synthetase	*glnA*	6.3.1.2	Cytoplasm
	Citrate lyase β chain		4.1.3.6	Cytoplasm
	Putative nitrogen regulatory protein P-II	*glnB*		Cytoplasm
	Protein-PII uridylyltransferase	*glnD*	2.7.7.59	Cytoplasm
	β-lactamase precursor	*ampC*	3.5.2.6	Extracellular
	Anthranilate synthase component II	*trpG*	4.1.3.27	Membrane
	Anthranilate phosphoribosyltransferase	*trpD*	2.4.2.18	Cytoplasm
	Indole-3-glycerol-phosphate synthase	*trpC*	4.1.1.48	Cytoplasm
	Tryptophan synthase subunit β	*trpB*	4.2.1.20	Cytoplasm
	Tryptophan synthase α chain	*trpA*	4.2.1.20	Cytoplasm
	Potassium-transporting ATPase	*kdpA*	3.6.3.12	Membrane
	Probable methylesterase		3.1.1.61	Cytoplasm
	Alkaline phosphatase	*phoA*	3.1.3.1	Membrane
	Respiratory nitrate reductase α chain	*narG*	1.7.99.4	Cytoplasm
	Sensor histidine kinase		2.7.13.3	Cytoplasm
7	Type II secretion system			
	Two-component sensor PilS	*pilS*	2.7.3.-	Membrane
	Leader peptidase (prepilin peptidase)/*N*-methyltransferase	*pilD*	3.4.23.43	Membrane
	Methyltransferase PilK	*pilK*	2.1.1.80	Membrane
	Sensor histidine kinase		2.7.13.3	
	Type IV pilus assembly protein	*PilF*		Membrane

TABLE 4. Continued

	Pathways unique to bacteria	Genes	EC No.	Localization
	Putative type IV pilin protein	*PilV*		Fimbrium
8	Type III secretion system			
	Flagellum-specific ATP synthase	*fliI*	3.6.3.14	Cytoplasm
	ATP synthase F0, B subunit		3.6.3.14	Membrane
9	Flagellar assembly			
	ATP synthase F0, B subunit		3.6.3.14	Membrane
10	Phosphotransferase system (PTS)			
	Phosphotransferase system, fructose-specific IIBC component	*fruA*	2.7.1.69	Membrane
	Putative two-component system transcriptional response regulator	*pstN*		Cytoplasm
	Probable phosphotransferase system enzyme I		2.7.3.9	Cytoplasm
11	Biosynthesis of siderophore group nonribosomal peptides			
	Isochorismate synthase	*pchA*	5.4.4.2	Cytoplasm
	Isochorismate pyruvate lyase	*pchB*	4.1.99.-	Cytoplasm
12	1,2-Dichloroethane degradation			
	Quinoprotein alcohol dehydrogenase	*exaA*	1.1.99.8	Periplasm
	Probable aldehyde dehydrogenase	*calB*	1.2.1.3	Cytoplasm
13	Toluene and xylene degradation			
	Catechol 1,2-dioxygenase	*catA*	1.13.11.1	Cytoplasm
14	Peptidoglycan biosynthesis			
	UDP-*N*-acetyl glucosamine 1-carboxyvinyltransferase	*murA*	2.5.1.7	Cytoplasm
	UDP-*N*-acetyl muramyl tripeptide synthase	*murD*	6.3.2.9	Cytoplasm
	UDP-*N*-acetyl muramoyl alanyl-D-glutamyl-2,6-diamino pimelate–D-alanyl-D-alanyl ligase	*murF*	6.3.2.10	Cytoplasm

*Targets have been selected from *P. aeruginosa* [Sakharkar et al., 2004; Perumal et al., 2007], *H. pylori* [Dutta et al., 2006], *B. pseudomallei* [Chong et al., 2006], *A. hydrophila* [Sharma et al., 2008], *N. gonorrhoeae* [Barh and Kumar, 2009], *N. meningitides* [Sarangi et al., 2009], *M. tuberculosis* [Asif et al., 2009], *S. typhi* [Rathi et al., 2009], *M. leprae* [Shanmugam and Natarajan, 2010], and *M. pneumonia* [Gupta et al., 2010]. None of these targets are found in any single pathogen indicated here. The EC nos. and localization information of targets are also presented.

obtain a readily available resource of putative essential genes as well as drug targets in human bacterial pathogens [Barh et al., 2009]. Selective and essential genes of pathogens that are non-homologous to the host are considered putative targets. Such target sequences are not present in the host, making the identified target unique to the pathogen. Thus, inhibition of such targets with appropriate drug(s) should avoid cytotoxicity issues in the host and will reduce the cost of ADMET validation of newly designed drugs [Sakharkar et al., 2004, Barh and Kumar, 2009].

SCOPE IN REVERSE VACCINOLOGY

Reverse vaccinology (RV) is a computational approach that takes a path different from conventional approaches for the development of vaccines [Bambini and Rappuoli, 2009; Serruto and Rappuoli, 2006]. Rather than start from a set of proteins that have experimentally been proven antigenic, RV explores previously unconsidered possibilities. RV seeks candidate proteins to elicit immune responses in the entire genome of an organism against which a vaccine is required; however, special attention is given to proteins that are secreted or exposed on the cell wall of the organism [Rappuoli, 2000]. Subtractive genomics-based target discovery is applicable to both drug and vaccine targets. It is preferable that a vaccine candidate be a non-human homologue. In bacteria, it is known that exported proteins are the main forms of interaction with cells infected by such organisms; therefore they are potential candidates for vaccine targets [Sibbald and van Dij, 2009; Simeone et al., 2009; Stavrinides et al., 2008; Bhavsar et al., 2007]. Hence, a non-human homologue secreted, or exo-membrane, or exported protein will be a better option for developing vaccine following RV. Table 6 lists bacterial pathogens for which RV methods are used for developing vaccines.

RV-related tools (see Table 3, marked with an asterisk) are widely used by the scientific community to ensure viability and increase reliability. As an example, we can cite the software SignalP [Bendtsen et al., 2004] for protein motif identification, which indicates the existence of signal peptides; the software TMHMM [Krogh et al., 2001] indicates transmembrane motifs. RV makes use of large-scale software, analyzing all the proteins derived from the genome and combining results. An example of combined results is to determine whether a protein is secreted because of a signal peptide (SignalP), taking into account that there is only one possible transmembrane domain (TMHMM). Otherwise, even though there is a signal peptide, the protein remains anchored to the cell membrane, which

TABLE 5. Selected Targets From Metabolic Pathways That Host and Pathogen Have in Common*

	Host–pathogen shared pathways	Gene	EC No.	Localization
1	DNA replication, repair, and recombination			
	DNA polymerase III subunit ε		2.7.7.7	Cytoplasm
	Holliday junction DNA helicase motor protein	*ruvA*	3.6.1	Membrane
2	Cell cycle			
	Cell division protein	*MraZ*	NA	Cytoplasm
	Cell division membrane protein	*FtsW*	NA	Membrane
3	Pyrimidine metabolism			
	FAD-dependent thymidylate synthase	*thyX*	2.1.1.148	Cytoplasm
	Dihydroorotase		3.5.2.3	Cytoplasm
4	Purine metabolism			
	DNA-directed RNA polymerase subunit α		2.7.7.6	Cytoplasm
	DNA polymerase III, alpha subunit		2.7.7.7	Cytoplasm
	DNA polymerase III subunit β		2.7.7.6	Cytoplasm
5	Transcription and translation			
	Transcription anti-termination protein	*NusB*		Cytoplasm
	50S ribosomal protein L30	*rpmD*		Cytoplasm
	50S ribosomal protein L35	*rpmI*		Cytoplasm
	50S ribosomal protein L34	*rpmH*		Cytoplasm
	50S ribosomal protein L1	*rplA*		Cytoplasm
	50S ribosomal protein L28	*rpmB*		Cytoplasm
	elongation factor P	*efp*		Cytoplasm
	tRNA guanine-N 1-methyltransferase	*trmD*	2.1.1.31	Cytoplasm
6	Histidine metabolism			
	Imidazole glycerol-phosphate dehydratase	*hisB*	4.2.1.19	Cytoplasm
	Histidinol dehydrogenase	*hisD*	1.1.1.23	Cytoplasm
	ATP Phosphoribosyl transferase	*hisG*	2.4.2.17	Cytoplasm
	Imidazole glycerol phosphate synthase subunit	*hisH*	2.4.2.-	Cytoplasm
	Phosphoribosyl-AMP cyclohydrolase		3.5.4.19	Cytoplasm
7	Thiamin biosynthesis			
	Thiamine monophosphate kinase	*thiL*	2.7.4.16	Cytoplasm
	Phosphomethylpyrimidine kinase		2.7.4.7	Cytoplasm
	Cysteine desulfurase		2.8.1.7	Cytoplasm
8	Aminosugars metabolism			
	UDP-*N*-acetylglucosamine 1-carboxyvinyltransferase		2.5.1.7	Cytoplasm
	UDP-*N*-Acetylenolpyruvoylglucosamine reductase		1.1.1.158	Cytoplasm
9	Phenylalanine, tryptophan, porphyrin and chlorophyll metabolism			
	Glutamyl-tRNA reductase		1.2.1.-	Cytoplasm
	Chorismate mutase		4.2.1.51	Cytoplasm
	3-dehydroquinate synthase		4.6.1.3	Cytoplasm
	Shikimate dehydrogenase		1.1.1.25	Cytoplasm
	Phospho-2-dehydro-3-deoxyheptonate aldolase	*aroH*	2.5.1.54	Cytoplasm
10	Glycine, isoleucine, serine, threonine, lysine metabolism			
	Homoserine dehydrogenase	*thrA*	1.1.1.3	Cytoplasm
	Gycyl-tRNA synthetase subunit β	*glyS*	6.1.1.14	Cytoplasm
	Homoserine kinase	*thrB*	2.7.1.39	Cytoplasm
	Diaminopimelate epimerase		5.1.1.7	Cytoplasm
	Aspartate-semialdehyde dehydrogenase		1.2.1.11	Cytoplasm
11	Terpenoid backbone biosynthesis			
	4-hydroxy-3-methyl but-2-enyl diphosphate reductase	*ispH*	1.17.1.2	Cytoplasm
12	Riboflavin metabolism			
	Riboflavin synthase subunit β	*ribH*	2.5.1.9	Cytoplasm
	Riboflavin synthase subunit β		2.5.1.-	Cytoplasm
	GTP cyclohydrolase II		3.5.4.25	Cytoplasm
13	Biotin biosynthesis			
	Biotin synthase family transferase		2.8.1.6	Cytoplasm
	Biotin synthase	*bioB*	2.8.1.6	Cytoplasm
	Dethiobiotin synthetase	*bioD1*	6.3.3.3	Cytoplasm
14	Folate biosynthesis			
	Dihydropteroate synthase	*FolP*	2.5.1.15	Cytoplasm

TABLE 5. Continued

	Host–pathogen shared pathways	Gene	EC No.	Localization
	p-Aminobenzoate synthase component		2.6.1.85	Cytoplasm
15	Oxidative phosphorylation			
	F0F1 ATP synthase subunit A		3.6.1.34	Membrane
	F0F1 ATP synthase subunit B		3.6.3.14	Membrane
16	Environmental information processing and membrane transport			
	ABC transporter iron-uptake	*fbpB*		Membrane
17	Protein export			
	Preprotein translocase subunit SecA	*secA*		Membrane
	Preprotein translocase subunit SecY	*secY*		Membrane
	Preprotein translocase subunit SecD	*secD*		Membrane
18	Pyruvate, propanoate, taurine, and hypotaurine metabolism			
	Acetate kinase		2.7.2.1	Intracellular
	Phosphotransacetylase		2.3.1.8	Cytoplasm
19	Steroids and isoprene biosynthesis			
	1-deoxy-D-xylulose 5-phosphate reductoisomerase	*dxr*	1.1.1.267	Cytoplasm
	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase		4.6.1.12	Cytoplasm
20	Glutamate, D-glutamine, and D-glutamate metabolism			
	Glutathione synthetase		6.3.2.3	Cytoplasm
	UDP-*N*-acetylmuramate-L-alanine ligase		6.3.2.8	Cytoplasm
	Glutamate racemase		5.1.1.3	Cytoplasm

*Targets have been selected from *P. aeruginosa* [Sakharkar et al., 2004; Perumal et al., 2007], *H. pylori* [Dutta et al., 2006], *B. pseudomallei* [Chong et al., 2006], *A. hydrophila* [Sharma et al., 2008], *N. gonorrhoeae* [Barh and Kumar, 2009], *N. meningitides* [Sarangi et al., 2009], *M. tuberculosis* [Asif et al., 2009], *S. typhi* [Rathi et al., 2009], *M. leprae* [Shanmugam and Natarajan, 2010], and *M. pneumonia* [Gupta et al., 2010]. All these targets are not found in any single pathogen indicated here. EC nos. and localization information of targets are also presented.

TABLE 6. Bacterial Pathogens for Which Reverse Vaccinology Approaches Have Been Adopted to Develop Vaccines*

Bacteria	Disease	Vaccine approach	Vaccine development stage
B. anthracis	Anthrax	Reverse vaccinology, CGH microarray, microarray proteomics, immunoproteomics	Discovery/preclinical
C. pneumoniae	Pneumonia, meningitis, middle era infections	Reverse vaccinology, proteomics	Discovery/preclinical
H. pylori	Ulcer, atrophicgastritis, adenocarcinoma, lymphoma	Reverse vaccinology, immunoproteomics	Discovery/preclinical
M. tuberculosis	Tuberculosis	Reverse vaccinology	Discovery/preclinical
N. meningitidis Serogroup B	Bacterial meningitis, septicemia	Reverse vaccinology, microarray, proteomics	Phase II clinical trials
S. aureus	Variety of infections, including, pelvic syndrome, rapidly progressive pneumonia, ocular infections, septic thrombophlebitis	CGH microarray Immunoproteomics	Discovery/preclinical
S. pyogenes (GAS)	Many systemic invasive infections, including necrotizing fasciitis, myositis, pneumonia, sepsis, arthritis	Genome-wide analysis, proteomics	Discovery/preclinical
S. agalactiae (GBS)	Bacterial sepsis, pneumonia, meningitis	Reverse vaccinology, classical or comparative	Discovery/preclinical
S. pneumoniae	Bacterial pneumonia, sepsis, sinusitis, otitis media, bacterial meningitis	Classical or comparative, reverse vaccinology, proteomics	Discovery/preclinical

*Adapted from Bambini and Rappuoli [2009].

leads to the classification of a membrane protein. There is also the software based on Hidden Markov Models (HMM) to check whether a protein has classic signs of retention and the software LipoP [Juncker et al., 2003] to check whether it is a lipoprotein.

We can make a rational analysis in RV, increasing the speed and reliability of results. Electron microscopy can be used to measure the thickness of cell walls. Data on cell wall thickness is used as the cutoff in the TMHMM output. A recently developed tool, SurfG plus, takes into account transmembrane domain positive prediction, and the estimated size of transmembrane domains is confronted with the estimated measure for the cell wall [Barinov et al., 2009]. In this way, it is possible to arrive at a more reliable estimate of the probability of a protein being characterized as an integrated membrane protein versus exposed on the surface. Besides developing a list of predicted proteins that could be exported, we can also make an analysis of possible B- and T-cell epitopes, in order to create an additional filter and minimize the list of targets that can be experimentally proven through this immunoinformatics approach [Serruto and Rappuoli, 2006].

DISADVANTAGES OF THE METHOD

Although the method has many advantages, there are certain concerns about the use of this technique. To perform subtraction, both the host and pathogen genomes are required; if one is not available, the analysis is difficult to perform. Similar to other *in silico* methods, targets derived from such analyses require experimental validation. In recent years, in almost all reports, DEG BLAST has been used for identification of essential genes of the pathogen based on gene or amino acid sequence similarities. The DEG is continuously enriched with mutagenesis-based new essential genes, also including new pathogens. An obvious concern is the consistency of the number of screened essential genes for a given pathogen with respect to time. We found that the number increases dramatically as a result of data enrichment of the DEG [Barh and Kumar, 2009]. Second, researchers, including ourselves, have not considered proteins with less than 100 amino acids [Dutta et al., 2006; Sharma et al., 2008; Barh and Kumar, 2009]. However, it has been found with DEG BLAST that many proteins listed in this essential gene database are <100 amino acids long. Therefore, when we exclude such small proteins, we may purge out some novel targets. This may not always be true, because it has been observed in mutagenesis studies that when there are insertion mutations in a nucleotide sequence of <300 bp, expression of nearby genes is altered, resulting in lethality, giving a false-positive result concerning the essentiality of the target gene. Also pathogen genes that are essential but non-homologous to any DEG-listed essential gene may be missed. Sakharkar et al. [2004] cautioned that because the method is based on BLAST results and does not consider specific growth conditions, care should be taken in interpreting the BLAST results. Otherwise, a conditional essential gene may be screened and selected. Hence a parallel method and tool independent of DEG should be developed. We found that if we increase the number of different species/strains within the same genus of the pathogen and use more than one host, the number of targets is considerably reduced (unpublished data). Hence, it is advisable to use multiple strains of a pathogen and all strain-specific hosts in the analysis to identify common targets for all strains as well as for a broad host range. Therefore, a pangenomics approach, including distant gene relationships, should be considered.

CONCLUSIONS AND PERSPECTIVES FOR THE FUTURE

In silico subtractive genomics is a rapid, powerful, and cost-effective approach for screening of drug and vaccine targets for any given pathogen, provided both the pathogen and host genomes are available. However, the identified targets require experimental validation. This approach requires multiple analyses at different stages that mostly use BLAST. Parameters of BLAST at different stages require optimization to standardize the method. Similarly, an efficient integrated platform needs to be developed to perform the entire analysis at the same time. A parallel method independent of DEG-based screening of essential genes is also required. Pangenomics-based conserved essential genes as the targets may be considered in such analyses. An *in silico* mutagenesis approach and other computational validation methods could be included in the analysis to improve the efficacy of the original method.

ACKNOWLEDGMENTS

D.B., S.T., and N.J. acknowledge the motivation and encouragement of all IIOAB members. S.T., N.J., A.N.M., and A.K. acknowledge facilities of DBT's Bioinformatics subcenters at their respective schools. A.A., A.R.S., and V.A. received research scholarships from CNPq.

AUTHOR CONTRIBUTIONS

D.B., S.T., and N.J. collected data and wrote the paper; A.A. and A.R.S. provided inputs on reverse vaccinology; A.N.M., A.K., and V.A. reviewed the article and provided technical guidance.

FINANCIAL DISCLOSURE

This work was carried out without any grant or other financial support, other than scholarships. There is no conflict of interest regarding this work.

REFERENCES

Allsop AE. 1998. Bacterial genome sequencing and drug discovery. Curr Opin Biotechnol 9:637–642.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Arias CA, Murray BE. 2009. Antibiotic-resistant bugs in the 21st century—a clinical super-challenge. N Engl J Med 360:439–443.

Asif SM, Asad A, Faizan A, Anjali MS, Arvind A, Neelesh K, Hirdesh K, Sanjay K. 2009. Dataset of potential targets for *Mycobacterium tuberculosis* H37Rv through comparative genome analysis. Bioinformation 4:245–248.

Bairoch A, Apweiler R. 1997. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 25: 31–36.

Bambini S, Rappuoli R. 2009. The use of genomics in microbial vaccine development. Drug Discov Today 14:252–260.

Barh D, Kumar A. 2009. In silico identification of candidate drug and vaccine targets from various pathways in *Neisseria gonorrhoeae*. In Silico Biol 9:225–231.

Barh D, Misra AN. 2009. In silico identification of membrane associated candidate drug targets in *Neisseria gonorrhoeae*. Int J Integr Biol 6:65–67.

Barh D, Kumar A, Misra AN. 2009. Genomic Target Database (GTD): a database of potential targets in human pathogenic bacteria. Bioinformation 4:50–51.

Barh D, Misra AN, Kumar A, Azevedo V. 2010a. A novel strategy of epitope design in *Neisseria gonorrhoeae*. Bioinformation 5: 77–85.

Barh D, Misra AN, Kumar A. 2010b. In silico identification of dual ability of *N. gonorrhoeae* ddl for developing drug and vaccine against pathogenic *Neisseria* and other human pathogens. J Proteomics Bioinform 3:082–090.

Barinov A, Loux V, Hammani A, Nicolas P, Langella P, Ehrlich D, Maguin E, van de Guchte M. 2009. Prediction of surface exposed proteins in *Streptococcus pyogenes*, with a potential application to other Gram-positive bacteria. Proteomics 9:61–73.

Bendtsen JD, Nielsen H, von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340:783–795.

Bernal A, Ear U, Kyrpides N. 2001. Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. Nucleic Acids Res 29:126–167.

Bhasin M, Garg A, Raghava GPS. 2005. PSLpred: prediction of subcellular localization of bacterial proteins. Bioinformatics 21: 2522–2524.

Bhavsar AP, Guttman JA, Finlay BB. 2007. Manipulation of host-cell pathways by bacterial pathogens. Nature 449:827–834.

Bruccoleri RE, Dougherty TJ, Davison DB. 1998. Concordance analysis of microbial genomes. Nucleic Acids Res 26:4482–4486.

Chan JN, Nislow C, Emili A. 2010. Recent advances and method development for drug target identification. Trends Pharmacol Sci 31:82–88.

Chen L, Yang J, Yu J, Yao Z, Sun L, Shen Y, Jin Q. 2005. VFDB: a reference database for bacterial virulence factors. Nucleic Acids Res 33:325–328.

Chetouani F, Glaser P, Kunst F. 2001. FindTarget: software for subtractive genome analysis. Microbiology 147:2643–2649.

Chhabra V, Sharma P, Anant A, Deshmukh S, Kaushik H, Gopal K, Srivastava N, Sharma N, Garg LC. 2010. Identification and modeling of a drug target for *Clostridium perfringens* SM101. Bioinformation 4:278–289.

Chong CE, Lim BS, Nathan S, Mohamed R. 2006. In silico analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. In Silico Biol 6:341–346.

Dong QJ, Wang Q, Xin YN, Li N, Xuan SY. 2009. Comparative genomics of *Helicobacter pylori*. World J Gastroenterol 15: 3984–3991.

Downs DM. 2006. Understanding microbial metabolism. Annu Rev Microbiol 60:533–559.

Dutta A, Singh SK, Ghosh P, Mukherjee R, Mitter S, Bandyopadhyay D. 2006. In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. In Silico Biol 6:43–47.

Fischbach MA, Walsh CT. 2009. Antibiotics for emerging pathogens. Science 28:1089–1093.

Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FSL. 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics 21:617–623.

Gupta SK, Singh S, Gupta MK, Pant KK, Seth PK. 2010. Identification of potential targets in *Mycoplasma pneumoniae* through subtractive genome analysis. J Antivir Antiretrovir 2: 038–041.

Holman AG, Davis PJ, Foster JM, Carlow CK, Kumar S. 2009. Computational prediction of essential genes in an unculturable endosymbiotic bacterium, *Wolbachia of Brugia malayi*. BMC Microbiol 9:243.

Hood DW. 1999. The utility of complete genome sequences in the study of pathogenic bacteria. Parasitology 118:S3–S9.

Huynen MA, Diaz-Lazcoz Y, Bork P. 1997. Differential genome display. Trends Genet 13:389–390.

Huynen M, Dandekar T, Bork P. 1998. Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. FEBS Lett 426:1–5.

Imai K, Asakawa N, Tsuji T, Akazawa F, Ino A, Sonoyama M, Mitaku S. 2008. SOSUI-GramN: high performance prediction for sub-cellular localization of proteins in Gram-negative bacteria. Bioinformation 2:417–421.

Ishii N, Robert M, Nakayama Y, Kanai A, Tomita M. 2004. Toward large-scale modeling of the microbial cell for computer simulation. J Biotechnol 113:281–294.

Itaya M. 1995. An estimation of minimal genome size required for life. FEBS Lett 362:257–260.

Ji Y. 2002. The role of genomics in the discovery of novel targets for antibiotic therapy. Pharmacogenomics 3:315–323.

Jordan IK, Rogozin IB, Wolf YI, Koonin EV. 2002. Essential genes are more evolutionarily conserved than are nonessential genes in bacteria. Genome Res 12:962–968.

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A. 2003. Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12:1652–1662.

Kobayashi K, Ehrlich SD, Albertini A, Amati G, Andersen KK, Arnaud M, Asai K, Ashikaga S, Aymerich S, Bessieres P, et al. 2003. Essential *Bacillus subtilis* genes. Proc Natl Acad Sci USA 100:4678–4683.

Koonin EV, Tatusov RL, Galperin MY. 1998. Beyond complete genomes: from sequence to structure and function. 8:355–363.

Krause A, Haas SA, Coward E, Vingron M. 2002. SYSTERS, GeneNest, SpliceNest: exploring sequence space from genome to protein. Nucleic Acids Res 30:299–300.

Kriventseva EV, Fleischmann W, Zdobnov EM Apweiler R. 2001. CluSTr: a database of clusters of SWISSPROT TrEMBL proteins. Nucleic Acids Res 29:33–36.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305:567–580.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

Lin J, Qian J. 2007. Systems biology approach to integrative comparative genomics. Expert Rev Proteomics 4:107–119.

McDevitt D, Rosenberg M. 2001. Exploiting genomics to discover new antibiotics. Trends Microbiol 9:611–617.

Meinke A, Henics T, Nagy E. 2004. Bacterial genomes pave the way to novel vaccines. Curr Opin Microbiol 7:314–320.

Mills SD. 2006. When will the genomics investment pay off for antibacterial discovery? Biochem Pharmacol 30:1096–1102.

Mushegian AR, Koonin EV. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci USA 93:10268–10273.

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. 1999. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27:29–34.

Owa T. 2007. Drug target validation and identification of secondary drug target effects using DNA microarrays. Tanpakushitsu Kakusan Koso 52:1808–1809.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444–2448.

Perrière G, Duret L, Gouy M. 2000. HOBACGEN: database system for comparative genomics in bacteria. Genome Res 10:379–385.

Perumal D, Lim CS, Sakharkar KR, Sakharkar MK. 2007. Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. In Silico Biol 7: 453–465.

Pinner RW, Teutsch SM, Simonsen L, Klug LA, Graber JM, Clarke MJ, Berkelman RL. 1996. Trends in infectious diseases mortality in the United States. JAMA 275:189–193.

Plotkin SA. 2005. Why certain vaccines have been delayed or not developed at all. Health Aff (Millwood) 24:631–634.

Pucci MJ. 2006. Use of genomics to select antibacterial targets. Biochem Pharmacol 71:1066–1072.

Rappuoli R. 2000. Reverse vaccinology. Curr Opin Microbiol 3: 445–450.

Rathi B, Aditya N, Sarangi AN, Trivedi N. 2009. Genome subtraction for novel target definition in *Salmonella typhi*. Bioinformation 4:143–150.

Sakharkar KR, Sakharkar MK, Chow VT. 2004. A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. In Silico Biol 4: 355–360.

Sarangi AN, Aggarwal R, Rahman Q, Trivedi N. 2009. Subtractive genomics approach for in silico identification and characterization of novel drug targets in *Neisseria meningitidis* serogroup B. J Comput Sci Syst Biol 2:255–258.

Serruto D, Rappuoli R. 2006. Post-genomic vaccine development. FEBS Lett 580:2985–2992.

Shanmugam A, Natarajan J. 2010. Computational genome analyses of metabolic enzymes in *Mycobacterium leprae* for drug target identification. Bioinformation 4:392–395.

Shao Y, He X, Harrison EM, Tai C, Ou HY, Rajakumar K, Deng Z. 2010. mGenomeSubtractor: a web-based tool for parallel in silico subtractive hybridization analysis of multiple bacterial genomes. Nucleic Acids Res 38:194–200.

Sharma V, Gupta P, Dixit A. 2008. In silico identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*. In Silico Biol 8:331–338.

Sibbald MJJB, van Dij JML. 2009. Secretome mapping in gram-positive pathogens. In: Wooldridge K, editor. Bacterial secreted protein: secretory mechanisms and role in pathogenesis. Norfolk, UK: Caister Academic Press. p 193–225.

Simeone R, Bottai D, Brosch R. 2009. ESX/type VII secretion systems and their role in host–pathogen interaction. Curr Opin Microbiol 12:4–10.

Singh NK, Selvam SM, Chakravarthy P. 2006. T-iDT: tool for identification of drug target in bacteria and validation by *Mycobacterium tuberculosis*. In Silico Biol 6:485–493.

Singh S, Malik BK, Sharma DK. 2007. Metabolic pathway analysis of *S. pneumoniae*: an in silico approach towards drug-design. J Bioinform Comput Biol 5:135–153.

Stavrinides J, McCann HC, Guttman DS. 2008. Host–pathogen interplay and the evolution of bacterial effectors. Cell Microbiol 10:285–292.

Stumm G, Russ A, Nehls M. 2002. Deductive genomics: a functional approach to identify innovative drug targets in the post-genome era. Am J Pharmacogenom 2:263–271.

Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. Science 278:631–637.

Tatusov RL, Natale DA, Garkavtsev IV, Tatusova TA, Shankavaram UT, Rao BS, Kiryutin B, Galperin MY, Fedorova ND, Koonin EV. 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res 29:22–28.

Thykaer J, Andersen MR, Baker SE. 2009. Essential pathway identification: from in silico analysis to potential antifungal targets in *Aspergillus fumigatus*. Med Mycol 47:S80–S87.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001.The sequence of the human genome. Science 291:1304–1351.

Yona G, Linial N, Linial M. 1999. ProtoMap: automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. Proteins 37:360–378.

Yu CS, Lin CJ, Hwang JK. 2004. Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 13:1402–1406.

Zhang R, Ou HY, Zhang CT. 2004. DEG: a database of essential genes. Nucleic Acids Res 32:D271–D272.

6.3.2 Genômica subtrativa *in silico* para a identificação de alvos em patógenos bacterianos de seres humanos

A identificação do alvo é o primeiro passo no processo de descoberta de drogas e vacinas, sendo que a genômica subtrativa *in silico* é amplamente utilizada neste processo. Por meio desta abordagem, nos anos recentes, um grande número de alvos foram identificados em agentes patogênicos bacterianos que são resistentes a drogas ou para os quais nenhuma vacina adequada esteja disponível. O método *in silico* reduz o tempo, bem como o custo de rastreamento do alvo. Embora seja uma técnica poderosa que pode ser aplicada a uma vasta gama de agentes patogênicos, há muitas armadilhas na análise e interpretação dos dados. Revisou-se esta abordagem, incluindo metas que foram identificadas com esta técnica, incluindo vantagens e desvantagens. Discutiu-se também as nossas próprias experiências utilizando esta abordagem.

Este foi o primeiro trabalho na área com bactérias do gênero *Corynebacterium*, incluindo a *C. pseudotuberculosis*.

Chem Biol Drug Des 2011

Research Article

© 2011 John Wiley & Sons A/S

doi: 10.1111/j.1747-0285.2011.01118.x

A Novel Comparative Genomics Analysis for Common Drug and Vaccine Targets in *Corynebacterium pseudotuberculosis* and other CMN Group of Human Pathogens

Debmalya Barh[1,2,*], Neha Jain[1,3], Sandeep Tiwari[1,3], Bibhu Prasad Parida[1,2,], Vivian D'Afonseca[4], Liwei Li[5], Amjad Ali[4], Anderson Rodrigues Santos[4], Luís Carlos Guimarães[4], Siomar de Castro Soares[4], Anderson Miyoshi[4], Atanu Bhattacharjee[6], Amarendra Narayan Misra[2], Artur Silva[7], Anil Kumar[3] and Vasco Azevedo[4]

[1]*Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India*

[2]*Department of Biosciences and Biotechnology, School of Biotechnology, Fakir Mohan University, Jnan Bigyan Vihar, Balasore, Orissa, India*

[3]*School of Biotechnology, Devi Ahilya University, Khandwa Rd., Indore, India*

[4]*Laboratório de Genética Celular e Molecular, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, CP 486, CEP 31270-901, Belo Horizonte, Minas Gerais, Brazil*

[5]*Department of Biochemistry and Molecular Biology, Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN, USA*

[6]*Department of Biotechnology and Bioinformatics, Bioinformatics Laboratory, North Eastern Hill University, Shillong, India*

[7]*Instituto de Ciências Biológicas, Universidade Federal do Pará, Belém-PA, Brazil*

**Corresponding author: Debmalya Barh, dr.barh@gmail.com*

Caseous lymphadenitis is a chronic goat and sheep disease caused by *Corynebacterium pseudotuberculosis* (*Cp*) that accounts for a huge economic loss worldwide. Proper vaccination or medication is not available because of the lack of understanding of molecular biology of the pathogen. In a recent approach, four *Cp* (CpFrc41, *Cp1002*, *CpC231*, and *CpI-19*) genomes were sequenced to elucidate the molecular pathology of the bacteria. In this study, using these four genome sequences along with other eight genomes (total 12 genomes) and a novel subtractive genomics approach (first time ever applied to a veterinary pathogen), we identified potential conserved common drug and vaccine targets of these four *Cp* strains along with other *Corybacterium, Mycobacterium* and *Nocardia* (CMN) group of human pathogens (*Corynebacterium diphtheriae* and *Mycobacterium tuberculosis*) considering goat, sheep, bovine, horse, and human as the most affected hosts. The minimal genome of *Cp1002* was found to consist of 724 genes, and 20 conserved common targets (to all *Cp* strains as well as CMN group of pathogens) from various metabolic pathways (13 from host-pathogen common and seven from pathogen's unique pathways) are potential targets irrespective of all hosts considered. ubiA from host-pathogen common pathway and an ABC-like transporter from unique pathways may serve dual (drug and vaccine) targets. Two *Corynebacterium*-specific (mscL and resB) and one broad-spectrum (rpmB) novel targets were also identified. Strain-specific targets are also discussed. Six important targets were subjected to virtual screening, and one compound was found to be potent enough to render two targets (cdc and nrdL). We are currently validating all identified targets and lead compounds.

Key words: *Corynebacterium pseudotuberculosis*, drug and vaccine targets, minimal genome, subtractive genomics

Received 13 December 2010, revised 3 March 2011 and accepted for publication 6 March 2011

Corynebacterium pseudotuberculosis (*Cp*) is a gram-positive bacteria and an important veterinary pathogen under the genus *Corynebacterium*. Other species under this genus are *Corynebacterium diphtheriae*, an important human pathogen, and *Corynebacterium glutamicum*, an important bacterium widely used in biotechnology (1). Owing to the pathogenic impacts and biological relevance, several *Corynebacterium* genomes including *C. diphtheriae*, *C. efficiens*, *C. urealyticum*, *C. aurimucosum* have been sequenced long back. The genus *Corynebacterium* belongs to the CMN group (2,3) that harbors species physiologically and ecologically heterogeneous although they share some common characteristics including a specific cell wall organization composed of peptidoglycan, arabinogalactan, and mycolic acid polymers (1,4), and having high G + C content in their genome (5,6).

Corynebacterium pseudotuberculosis infection causes the disease known as caseous lymphadenitis (CLA) in goat and sheep, a chronic contagious disease characterized by abscess formation in superficial lymph nodes and in subcutaneous tissues, and in severe cases, it infects the lungs, kidneys, liver, and spleen, threatening the life of the infected animal (7,8). CLA is prevalent around the world but extensively present in regions of intensive husbandry (9,10) including Australia (11), Brazil (12), New Zealand, South Africa, the USA, Israel (13), and the UK (14–16). CLA accounts for a significant economic loss by hindering the production of wool, leather, meat, and milk yields (8,11), decreasing reproductive efficiencies of affected animals, condemnation of carcasses and skins (17,18), culling of affected animals, and mortality from the internal environment (8).

Although *C. pseudotuberculosis* was originally identified as the causative microorganism of CLA in sheep and goats, this bacterium has also been isolated from other species, including horses, in which it causes ulcerative lymphangitis and pigeon fever in cattle, camels, swine, buffaloes, and humans (7,9,19). Ulcerative lymphangitis is one of the most common and economically deleterious infectious diseases of horses in California and is increasing in other dry, western states of the USA. Its onset is slow, leading to painful inflammation, nodules, and ulcers, especially in the regions of fetlock.

The pathogen also infects humans although very few cases are reported and most are because of occupational exposure with symptoms of lymphadenitis and abscesses. About 25 human cases have been reported in Australia (6,20).

In cattle, the pathogen is transmitted through the ingestion of contaminated food and water, through wounds on the skin surface, or by aerosol infection of the lungs (21). Early diagnosis is quite difficult, and no specific diagnostic test is available, increasing the severity of the disease owing to delay in treatment. Notably, an enzyme-linked immunoassay for the detection of phospholipase D is used for diagnosis, although sensitivity of the test remains unsatisfactory (22). There is also no available vaccine that can effectively prevent CLA. Some available bacterin-toxoid vaccines are capable of decreasing the prevalence and number of abscesses in the host; however, it is very difficult to maintain the efficiency of the vaccine for a long time, and the immunization efficacy also varies depending on the type of host (23,24).

Similarly, there is no drug available to effectively control the infection. Although the pathogen shows sensitivity to many drugs *in vitro*, *Cp* is reported to be highly resistant to penicillin (25) and several other drugs. The treatment approaches are difficult as the bacteria remains protected inside the host abscesses. Therefore, antibiotics are generally ineffective and are not recommended. Approach with antimicrobial chemotherapy is also not fully effective because of an inadequate drug delivery system that cannot cross the abscesses layer (26).

In a recent approach to elucidate the molecular biology and the pathogenicity of *Cp*, three strains (*Cp1002*:isolated from Brazilian goat, *CpC231*: isolated from Australian sheep, and *Cp I-19*: isolated from Israel dairy cow) were completely sequenced, and based on comparative genomics study, it is reported that both strains share common features including similar G + C content, gene density, and some distinguishing features including genome size, number of genes, and pseudogenes (27,28). The human isolate *CpFrc41* genome is already available in NCBI (NC_014329) that was also sequenced by the same group.

Dorella *et al.* in 2006 (6) have employed some genomic approaches to develop effective control measures to prevent the disease, but CLA still remains uncontrolled nevertheless. Therefore, we took advantage of the recently sequenced genome to identify common and effective drug and vaccine targets at the genomic level against strains of *Cp* that could be useful to design drugs and vaccines against the pathogen and to prevent or treat the disease especially in goat, sheep, bovine, horse, and human hosts. We have also included CMN group of two human pathogens (*C. diphtheriae* and *Mycobacterium tuberculosis*) and one industrially important microbe (*C. glutamicum*) in this analysis to identify broad-spectrum conserved potential targets that can be useful to develop a common drug and/or vaccine regardless of the pathogen or host.

Subtractive genomics approaches have successfully been used to identify targets in various human bacterial pathogens including *Pseudomonas aeruginosa* (29), *Helicobacter pylori* (30), *Burkholderia pseudomalleii* (31), *M. tuberculosis* (32), *Neisseria gonorrhoeae* (33), and *Salmonella typhi* (34) among others. This study is the first ever to apply the subtraction approach at a genomic level to identifying drug and vaccine targets specifically in *Cp*, a veterinary pathogen.

Materials and Methods

Genomes and identification of essential genes in *Cp1002*

We employed comparative and subtractive genomics approaches following a modified method as described by Barh and Kumar (33) on the strategy that a target will be an essential survival gene for the pathogen, which is non-homologous to its hosts. Twelve genomes were used in this study where the *Cp1002*, *CpC231,* and *CpI-19* were new, and other genomes (*CpFrc41*, *C. diphtheriae*, *C. glutamicum*, *M. tuberculosis*, goat, sheep, bovine, horse, and human) were accessed from NCBI genome server. We took the advantage of comparatively smallest genome of Brazilian strain *Cp1002* among the three *Cp* strains to identify essential genes and targets of the pathogen. In brief, each gene and protein sequence of the *Cp 1002* were subjected to BLASTX (35) and BLASTP (36), respectively, against the Database of Essential Genes (DEG: http://tubic.tju.edu.cn/deg) (37) to identify all essential genes of the strain and to map the minimal genome. Essential genes were shortlisted based on cutoff values for bit score, *E*-value, and percentage of identity at amino acid level, respectively, >100, *E* = 0.0001, and >35%. *E*-value is the 'Expect value' that describes the expected number of 'hits' to see by chance when searching a database of a particular size. The lower the *E*-value, the more significant the match is. In few cases, genes having <100 bits score and >25% to <35% identity were also selected where the query gene of *Cp1002* showed same gene name and functioned against a DEG listed essential gene hit. Proteins <100 amino acids were also included in the selection criteria. Selected essential genes were classified according to Clusters of Orthologous Groups of Proteins (COGs) nomenclature based on the comparative genomics with *CpFrc41*, *C. diphtheriae,* and *M. tuberculosis* using corresponding pathogen genomes available in NCBI.

Localization, pathogenic island (PAI), and core gene prediction

Membrane, potentially surface exposed (PSE), secreted, and cytoplasmic localization prediction of essential *Cp1002* proteins was carried out using SurfG+ (http://genome.jouy.inra.fr/surfgplus/) (a new tool under evaluation), and the results were cross-checked with tools used by Barh and Kumar (33). List of PAI-related *Cp* proteins and pangenomics-based identified core, accessory, and dispensable genes of *Cp* were prepared based on the study of D'Afonseca *et al.* (27) and PIPS software (http://www.genoma.ufpa.br/lgcm/pips) developed by Soares, S.C.; Abreu, V.A.C.; McCulloch, J.A.; D'Afonseca, V.; Ramos, R.T.J.; Silva, A.; Baumbach, J.; Trost, E.; Tauch, A.; Hirata-Jr., R.; Mattos-Guaraldi, A.L.; Miyoshi, A.; Azevedo, V. (unpublished data).

Genome subtraction for target identification in Cp1002

To subtract essential non-host homologs (potential targets) of *Cp1002,* we performed BLASTp against sheep, goat, and bovine genomes in NCBI BLAST server. Additionally, GoSh DB (http://www.itb.cnr.it/gosh) was used for goat and sheep. BLASTp was performed using each selected essential protein sequence of *Cp* at *E*-value cutoff $E = 1$ (for GoSh DB, 1e−1). Sequences that showed similarity with any of the selected hosts were eliminated, and sequences without homology (non-host homologs) were considered as putative targets at this initial stage of screening.

Identified targets were also screened against horse and human genomes using horse and human BLASTp at NCBI server with default parameters (*E*-value cutoff $E = 1$) to identify sequence similarity, respectively. The human genome was considered to avoid possible off targeting side-effects. In the results section, goat-, sheep-, and bovine-specific common targets have been grouped together and horse- and human-specific targets are represented separately as appropriate.

Common targets identification in Cp1002, CpC231, CpFrc41, CpI-19*, and other CMN group of pathogens*

To identify targets from the Australian sheep isolate *CpC231*, human isolate *CpFrc41*, and bovine isolate *CpI-19*, we employed a strategy to find whether the identified targets of goat isolate *Cp1002* were similar or identical to *CpC231*, *CpFrc41,* and *CpI-19* by aligning the amino acid sequences of identified essential proteins of *Cp1002* with the corresponding *CpC231* and *CpFrc41* sequences based on names and using BLAST. We also used the BLAST program available in http://corynecyc.cebio.org database for the same purpose. The selected *Cp1002* targets in the previous step that showed high similarity (∼80% identity at $E = 0.0001$) with corresponding *CpC231*, *CpFrc41,* and *CpI-19* protein sequence were selected as common targets for all these four strains (*Cp1002*, *CpC231*, *CpFrc41*, and *CpI-19*), while *Cp1002* proteins that did not show such homology were selected as putative targets for only *Cp1002*. Each identified *CpC231*, *CpFrc41*, and *CpI-19* target sequence was further subjected to DEG BLAST for cross-check. To assess whether the identified common *Cp* targets were essential genes or targets in other *Corynebacterium* and CMN group of species, the non-pathogenic *C. glutamicum* and human pathogens *C. diphtheriae* and *M. tuberculosis* genomes were analyzed following the method applied in the case of *CpC231*, *CpFrc41,* and *CpI-19*. Therefore, in this way, identified targets are common to all pathogens considered having a broad host range.

Metabolic pathway analysis

As goat, sheep, and horse metabolic pathways are not available, we presumed that the bovine pathways were sufficiently similar to these hosts. Host-pathogen common and pathogen-specific unique metabolic pathway–related targets were identified using a cross-species pathway comparison module available at http://corynecyc.cebio.org, selecting pathways for bovine, human, *Cp1002*, and *CpC231*. Owing to high similarities in genomic context with *Cp*, *C. diphtheriae* pathways from kyoto encyclopedia of genes and genomes (KEGG) (38) were utilized as reference for *Cp*. Bovine and human metabolic pathways from KEGG were also used as references for hosts. Pathways and related *Cp* targets were selected based on the following selection criteria: (1) The target must be an essential non-host homolog where hosts are goat, sheep, bovine, horse, or human. (2) Target should be a core gene of the pathogen. (3) A target is preferable if it is involved in pathogen's unique pathway. (4) A better target will be involved in more than one pathogen's unique pathways. (5) A pathway will be considered better if it consists of multiple targets. (6) An enzyme target should not be of same class of protein, and the EC. No. of the target should not match with any protein product of the host in host-pathogen's common pathways. (7) Pathogen-specific unique pathway targets that are common to all *Cp* strains as well as other pathogens considered are better for broad-spectrum targets. (8) Targets that are only present in *Cp1002*-specific pathway but not in *CpC231* or *CpFrc41* can be considered as *Cp1002*-specific targets and *vise versa*. (9) Non-host homolog PAI-related or virulence proteins are better targets. (10) Secreted, PSE, membrane-exposed enzymes or transporter targets can be considered for duel purpose, i.e., developing drug and vaccine where enzyme targets are more preferable. (11) Non-human homolog targets are considered to minimize possible off target side-effects and to avoid residual drug effect and absorption, distribution, metabolism, excretion, and toxicity (ADMET) as the products of all *Cp* hosts (except horse) are human consumable, and *CpFrc41* is a human isolate. (12) Targets should be common to most of the pathogen strains as well as its related species.

3D modeling and virtual screening

The three-dimensional (3D) protein structures for *C. pseudotuberculosis* genes were built using PRIME (Version 22), a protein modeling program from Schrodinger Inc. New York, NY, USA (http://www.schrodinger.com). *Cp1002* protein sequences were used. BLAST search was carried out against RCSB PDB (http://www.rcsb.org/pdb) to identify crystal structures that have high sequence similarity to CP proteins, which will be used as potential templates for model building. In addition to BLAST sequence alignment, secondary structures of CP proteins were predicted using the PSIPRED (39) program, which were further employed by the *Prime* program to adjust and optimize the alignment between CP protein and structural templates. The built CP models were energy minimized to remove any steric clashes.

To carry out structure-based virtual screening, a compound library containing lead-like small molecules was prepared. The compound structures were obtained from the ZINC (http://zinc.docking.org/) website (40), which are commercially available from the ChemDiv Inc. (San Diego, CA, USA) (http://www.chemdiv.com). The compounds were further processed in CANVAS (Version 13) from Schrodinger Inc. New York, NY, USA to eliminate structurally similar analogs and produce a structurally diverse set of 10 000 molecules. Compounds were then docked onto protein structures using GLIDE SP (Version 56) from Schrodinger Inc. with a rigid-receptor flexible-ligand protocol. Docking was focused on the pockets identified on protein models. The docked protein/ligand complex was scored using Schrodinger's proprietary *GlideScore* scoring function. It consists of eight empirical terms that are considered essential for the binding of a ligand to a protein, which includes van der Waals energy, Coulomb energy, lipophilic term for hydrophobic effects, hydrogen-bonding interaction, metal-binding term, penalty for buried polar groups, penalty for freezing rotatable bonds, and polar interaction excluding H-bonding. The scoring function was parameterized to best correlated with the experimentally determined thermodynamic binding data. The compounds were ranked by the Glide score. The ones showing on top of the list, which have the strongest predicted binding affinities, were considered hits from virtual screening.

Results

Minimal genome of Cp1002

Using DEG-based comparative genomics, we predicted the minimal genome of *Cp1002* to consist of 724 genes (*Cp1002* has 2098 genes); therefore, 34.0% of total protein coding sequences was found to be essential for the pathogen. The number can be further reduced using various criteria, but as it is not the goal of this analysis, we chose not to do so. Screened essential genes can be categorized into 19 functional groups based on COG classification (Figure 1). While translation machinery–related genes were found to be the largest group (113 genes), RNA processing and modification class were found to be the smallest (one gene). Using subtractive genomics, a total number of 118 non-host (goat, sheep, and bovine) essential genes belonging to various classes of COGs were predicted to be targeted in this pathogen. Essential genes to non-host homolog ratios within a functional group were highest (32 genes to 17) for the unknown function class and were lowest for the energy production and conversion group (67 genes to 1) (Figure 1).

Targets in the Cp1002 *genome*

At initial target screening, considering goat, sheep, and bovine as hosts, we identified 118 targets from *Cp1002* genome. However, after we screened targets based on our criteria 2 and 6, core gene, and EC numbers, only 100 targets were selected. Among them, 48 and 32 proteins, respectively, from host-pathogen common pathways and pathogen-specific unique pathways were found as potential targets. Two conserved membrane proteins (considered as other group, as they do not fall under any pathway) and a total of 18 hypothetical proteins were identified but not involved in any pathway (data not shown).

Common targets in Cp1002, CpC231, CpFrc41, *and* CpI-19 *with respect to goat, sheep, and bovine hosts*

Following the comparative genomics approach as described in the method, using goat, sheep, and bovine as hosts, we identified 76 putative targets common to *Cp1002* and *CpC231*. When we included the *CpFrc41*, the number of common targets further reduced to 56. Three targets from common as well as unique pathways, one from other group, and all hypothetical proteins that are present in *Cp1002* were absent in *CpC231*. Similarly, 13 and six targets, respectively, from common and pathogen's unique pathways of *Cp1002* are absent in *CpFrc41*. All 18 hypothetical and two other groups of targets of *Cp1002* were also not found in *CpFrc41*. Next, we added CpI-19 genome in pathogen list and found only 15 targets were common to all these four *Cp* isolates. However, two *Cp1002* proteins are named differentially in the case of *CpI-19* (Cp1002_1094 ABC-type transporter is CpI-19 putative membrane protein, and Cp1002_1959 phosphoribose diphosphate is *CpI-19* 4-hydroxybenzoate polyprenyltransferase-like prenyltransferase). Therefore, at this initial level of target screening, 51 targets were selected that are common to all three *Cp* strains with respect to goat, sheep, and bovine as hosts (data not shown).

Common Cp *targets with respect to human and horse*

As per our selection criteria 1, we next screened these 51 targets against horse and human genomes to identify targets that are common to all *Cp* strains with respect to all five hosts (goat, sheep, bovine, horse, and human) considered in this analysis. As found earlier, there was a decrease in the number of common targets with increase in the number of strains, and the similar trend was observed with increase in number of hosts. While we considered horse along with goat, sheep, and bovine, the total number of common targets decreased to 46 and when we further included human in the host list, the number was further reduced to 38 (26 in common and 13 in unique pathways). These 38 targets can be considered to develop drug for any *Cp* strains used in this analysis (Table S1).

Conserved common targets in other CMN species

Next, as per the selection criteria 12, to determine whether all these 38 targets were common in other species of *Corynebacterium*

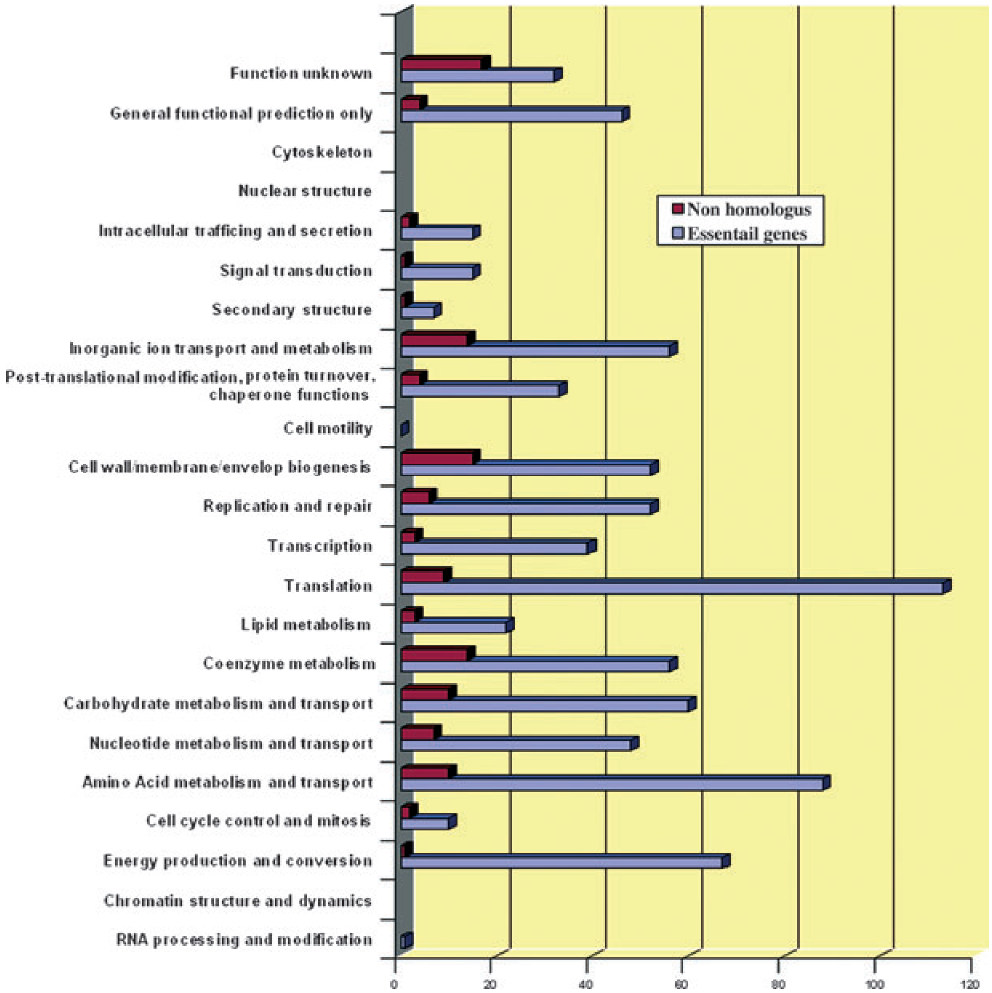


Figure 1: Clusters of Orthologous Groups of Proteins (COG) functional classification of *Cp1002* essential genes. The figure also shows the ratio of essential genes and non-host homolog genes of the species under each functional COG classes.

and CMN group of pathogens, we used similar comparative subtractive genomics approach that was used for target identification in *CpC231* and *Cp Frc41* from the list of targets identified in *Cp1002*. Selected species include the human pathogens *C. diphtheriae* (Cd) and *M. tuberculosis* (Mt). We have also considered non-pathogenic *Corynebacterium* species, *C. glutamicum* (Cg), for the same purpose. A drastic reduction in the number of targets was counted. Of 38 targets, only 20 targets were found to be common to all *Cp* strains and other CMN species. These 20 targets are non-homologous to any of the hosts (goat, sheep, bovine, horse, and human). Therefore, these targets may be used to develop broad-spectrum anti-*Cp* drugs irrespective of any host considered. Among these 20 targets, 13 (nine cytoplasmic enzymes, three ribosomal proteins, and one membrane enzyme, ubiA) belong to host-pathogen common pathways and rest (four cytoplasmic enzymes, one iron regulator ABC transporter (sufB), one membrane-located ABC-like transporter and one membrane protein) are involved in pathogens' unique metabolic pathways (Table S1).

Conserved common targets in pathogens' unique pathways

Of the 20 targets, seven targets are found to be involved in pathogen-specific unique metabolic pathways. When we applied our target selection criteria 5, as mentioned in the method, we found that peptidoglycan biosynthesis pathway was the most important pathogens' unique pathway that could be effectively targeted because of the presence of four cytoplasmic enzyme targets, namely murA, murD, murE, and murF. The next significant pathway was the transport system mainly ABC transporters. Important targets in this pathway are iron-regulated ABC-type transporter (sufB), a cytoplasmic ABC iron III transporter, and membrane-bound ABC-type transporter (*Cp1002*_1094). Membrane-localized enzyme putative lipoprotein signal peptidase (EC No: 3.4.23.36) that plays a crucial role in cell membrane/wall biogenesis and membrane transport was found to be an attractive target in all pathogens. This enzyme has been reported to be a putative target in *Aeromonas hydrophila* (41) and is also conserved in *Corynebacterium*; therefore, it can be better suited to develop anti-*Cp* drugs. As the enzyme is membrane localized, it can also be a good candidate to develop anti-*Cp* vaccine.

Conserved common targets in host-pathogen common pathways

Cytoplasmic translation machinery proteins constituted the highest number of targets (four of 13). These proteins are rpmB, rpmD, rpmL, and ribonuclease-P (rnpA). Among other targets, the most attractive one was homoserine dehydrogenase (thrA) from homoserine and lysine biosynthesis pathway. thrA is also a key enzyme in glycine, serine, and threonine metabolism pathways. Therefore, targeting thrA might block multiple essential metabolic pathways of the pathogen.

Imidazole glycerol-phosphate dehydrogenase (hisB) was identified from histidine metabolism pathway. Similarly, phosphoribose diphosphate (ubiA) in glycan metabolism pathway was found to be a broad-spectrum target. Being a membrane-located enzyme, ubiA may also serve dual purpose, i.e., drug and vaccine target.

Although, from biotin biosynthesis pathway, biotin synthase family transferase and biotin synthase (bioB) were identified as targets for all three *Cp* strains, only bioB was qualified to be a broad-spectrum target considering all pathogen genomes used in this analysis. Thiamine monophosphate kinase (thiL), dihydropteroate synthase (folP), and precorrin-4 c 11-methyl transferase (cobM), respectively, from thiamine, tetrahydrofolate, and adenosylcobalamine biosynthesis pathways were found to be attractive targets regardless any pathogen and host range considered. Two other important targets in common pathways were ribonucleotide reductase stimulatory protein (nrdL) and decxycitidine triphosphate deaminase (dcd) from, respectively, nucleotide metabolism and pyrimidine biosynthesis pathways.

Common novel targets in Cp strains

Extensive literature search was performed to identify novel targets. We considered novel targets that are not reported in any other pathogen but are common in all *Cp* strains with respect to all hosts considered. Therefore, we screened such novel targets from the list of 38 targets. In host-pathogen common metabolic pathways, such targets were cytoplasmic rplA, rpmB (from translation machinery), and membrane-located putative H+ antiporter subunit-c from ATP synthesis–coupled electron transport pathway. Although rplA and H+ antiporter subunit-c are absent in *M. tuberculosis*, rpmB was found to be a universal novel target for any pathogen considered in this analysis.

From pathogens' unique pathways, three novel targets, namely amino acid career protein (sodium and amino acid transport), mscL (cell wall biogenesis and transport), and resB (electron transport) were identified. These three targets are either membrane or PSE localized, conserved in *Corynebacteria*, and targets for all species. Therefore, these three can be used for dual purpose.

PAI-related targets

Pathogenicity island targets are attractive in developing drug/vaccine and as per our selection criteria (9), as mentioned in method, we scanned 38 targets for PAI, and only dcd was identified. dcd is found to be a common target for all pathogens and has been reported as a target in *M. tuberculosis* (42).

Targets selected for 3D modeling

To design drug, we selected some important and common targets. A total of six targets were selected. As found in the analysis, the peptidoglycan biosynthesis pathway is the most attractive pathogens' unique metabolic pathway; murA and murE were selected from this pathway. From host-pathogen common metabolic pathways, folP (tetrahydrofolate biosynthesis pathway), nrdL (nucleotide metabolism), and the sole PAI-related target, dcd (pyrimidine biosynthesis pathways) were selected. Although nrdH is not present in *C. diphtheriae*, we considered it because of its importance in redox pathway that is essential for the survival of any pathogen inside the host. nrdH has also been reported as an attractive target in *M. tuberculosis* (43).

As the experimentally determined 3D structures are not available, protein models were built using comparative modeling techniques. Protein structures are more conserved among evolutionary-related homologs. Generally, medium to high-resolution models can be obtained if the sequence identity is >30%. The sequence identity in this work ranges from 41% to 82% as shown in Table S2, which assures the quality of the models. The similarity between the model and the crystal structure on the binding site is generally even higher, especially for dcd.

Virtual screening and docking

Compounds identified from virtual screening with most favorable binding energy were considered as hits. Hits with strongest binding energy were depicted in sticks binding on the surface of the pocket (Figure 2), while the chemical structures of the top five hits for each protein were listed in Figure 3. The physicochemical properties of top five hits based on glide scores for each target protein are represented in Table S3. Important amino acid residues that interact with docked compound are list in Table S4. The hits were named from c1(gene) to c5(gene) in the order of predicted binding affinity. The inspection on the docked conformation shows that the binding cavity on the protein was explored very effectively by this top hits. Although the hits are not validated *in vitro*, it is interesting to see that the top one hit to folP, c1(folP), is actually a substructure of an antibiotic drug cefmetazole. Among the top five hits to each protein, there is one compound shared by two proteins, c5(dcd)/c1(nrdL). Although structurally speaking, the two cavities are not quite similar, because small molecules are flexible and could adopt different conformations during binding. It may render more potent antibiotic activity by targeting two essential bacterial proteins simultaneously.

Discussion

Although subtractive genomics is frequently used to identify drug targets in human pathogenic bacteria, in this study, for the first time, the approach was applied to identify drug and vaccine targets

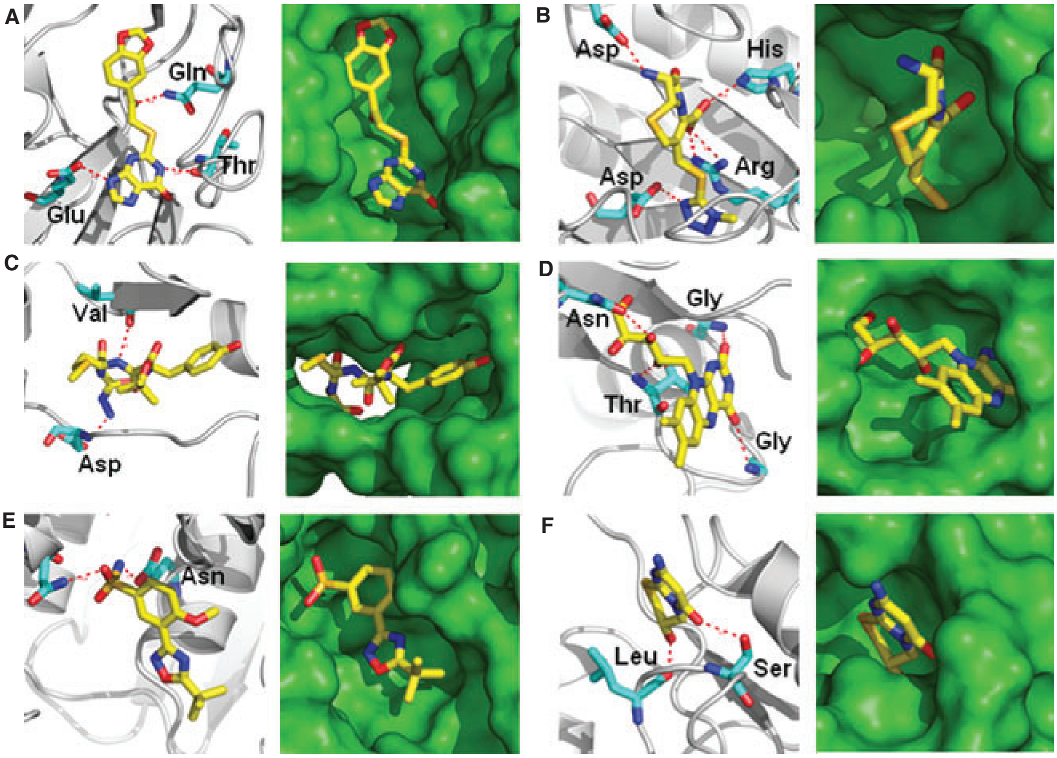


Figure 2: Ribbon and surface representation of the top compound bound to (A) dcd, (B) FolP, (C) nrdH, (D) nrdL, (E) murA, and (F) murE. The compounds are in stick representation with carbon, oxygen, and nitrogen atoms colored in yellow, red, and blue, respectively.

of a non-human pathogen. In DEG-based essential gene screening, most *Cp* hits were found with *M. tuberculosis*. During COG classification of essential *Cp* genes using two other *Corynebacterium* species, namely *C. diphtheriae* and *M. tuberculosis* proteomes available in NCBI, it was noted that *Cp* genes were shared by both species. A substantial number of *Cp* genes were conserved and present in *C. diphtheriae,* and a few genes that were not present were found in *M. tuberculosis* and *vice versa*. Essential genes for *C. diphtheriae* were not listed in DEG, but genes for *M. tuberculosis* were shown. It is interesting that while DEG listed 614 essential genes for *M. tuberculosis*, our analysis showed that the minimal genome of *Cp1002* consisted of approximately 724 genes. The higher number of essential genes in *Cp* relative to *M. tuberculosis* may be as a result of sharing and horizontal transfer of genes among CMN group of *Corynebacterium* species and other bacterial classes listed in DEG.

Polymorphic peptidoglycans are unique components that constitute the bacterial cell wall and play a vital role in bacterial defense, virulence, and survival. Therefore, the peptidoglycan biosynthesis pathways (I and II) that are unique to the bacteria are very crucial. Four cytoplasmic enzymes, murA, murD, murE, and murF, were identified as targets from this pathway. murA and murD were additionally involved in nucleotide sugar and glutamate metabolism pathways, respectively. murE and murF also were shown to play a vital role in lysine biosynthesis. While murD was conserved in *Corynebacterium*, murF was conserved in *Mycobacterium.* All four targets were previously reported in *Mycobacterium leprae* (44) and few other organisms (Table S1). In *Cp*, we found that this peptidoglycan biosynthesis pathway was the best targeting pathway as the above-mentioned four targets found here were essential non-host homologs with respect to all five hosts considered, and all targets were highly potential because of their additional involvement in multiple pathways. D-alanine is an essential component of the peptidoglycan layer in bacterial cell wall and D-alanine–D-alanine ligase (ddl) is a common target for various human pathogens in this pathway. But it is interesting that ddl was not found to be a target in *Cp.*

Bacterial transport system–related targets are attractive in developing antibiotics. Iron transport–related ABC transporters have been reported as essential genes and drug targets in *N. gonorrhoeae* (33) and *Clostridium perfringens* (45) among others. Such transporters were also predicted to be good vaccine targets because of their antigenic properties and exomembrane or PSE localization (46). We found membrane-localized ABC-type transporter (*Cp1002*_1094) and cytoplasmic iron-regulated ABC-type transporter (sufB) are broad-spectrum targets. Both of these targets have been identified in *C. perfringens* by Chhabra *et al.*, (45). *Cp1002*_1094, being a membrane protein, may be potential in developing drug as well as vaccine.

Putative lipoprotein signal peptidase (*Cp1002*_1377/lspA, EC: 3.4.23.36), which is conserved in *Corynebacterium*, was selected as an important target. It is a common target for all pathogens considered and also a non-homolog to all five hosts considered in this analysis. This target is involved in cell wall and membrane biogenesis, intracellular trafficking and secretion, membrane transport, protein export pathways, and an enzyme with the same EC number

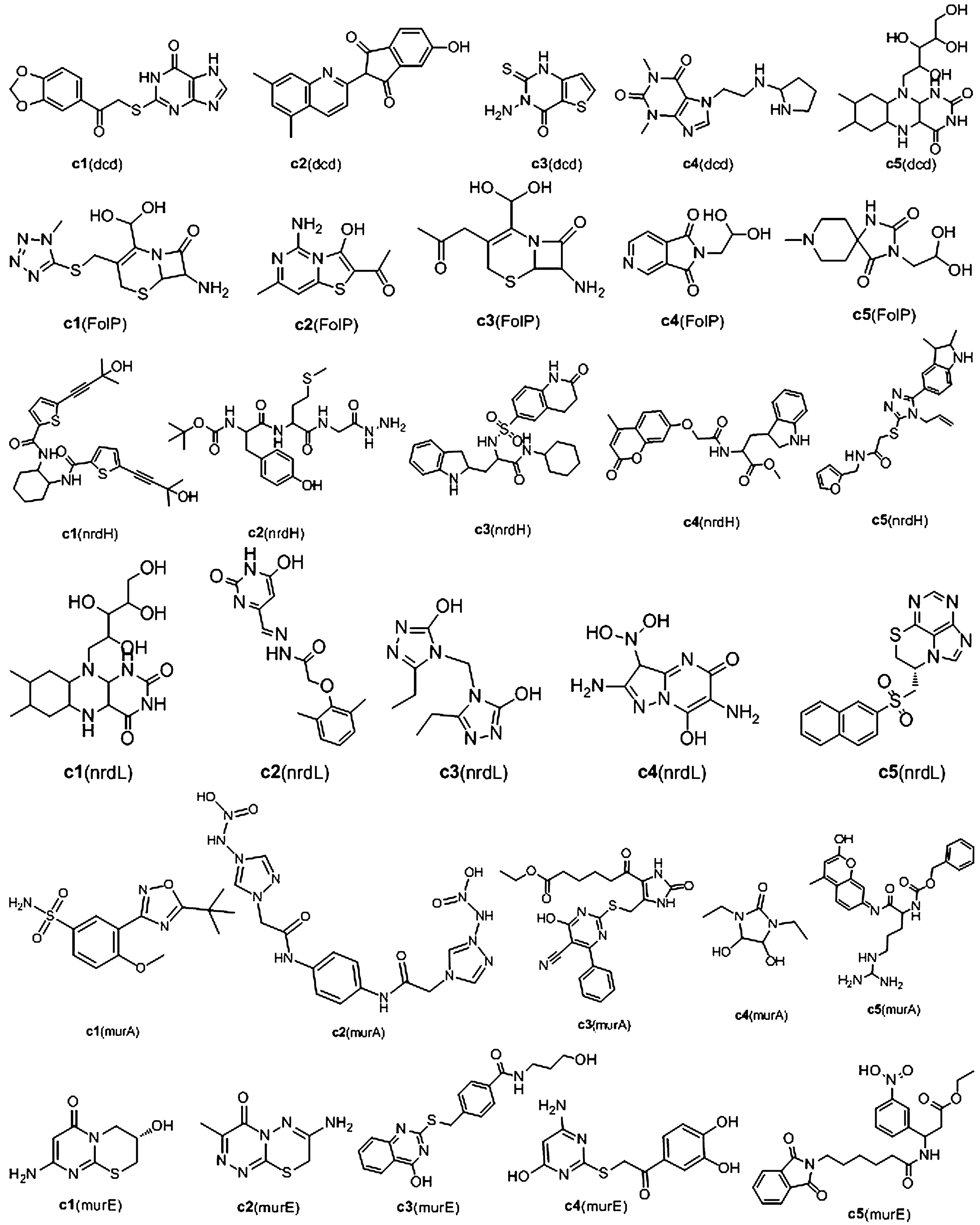


Figure 3: Chemical structures for the top five compounds predicted by GLIDE.

has been identified as a target in *A. hydrophila* (41). This protein is localized to the membrane and therefore it is also suitable for vaccine development.

Two other targets include putative amino acid carrier protein (*Cp1002*_1332, sodium transport) and large-conductance mechanosensitive channel protein (*Cp1002*_0665/mscL, transport, and membrane biogenesis), which are novel targets. These two targets are highly conserved in *Corynebacterium* and are membrane localized. Therefore, they were shown to be potential targets in developing both anti-*Cp* drugs and vaccines for all five hosts.

Cytochrome *C* biogenesis protein (resB) is non-homologous to all five hosts. Moreover, resB is an essential gene in *B. pseudomallei*, and it is involved in cytochrome C biogenesis. It also plays the role of an essential cofactor in oxidoreduction process (31). In this study, we also found resB as a potential novel target to inhibit the oxidoreduction process of *Cp*. resB is a PSE-localized protein; therefore, it may be potential to vaccine development. However, in *M. tuberculosis*, it is not found to be a potential target.

Bacterial two-component and secretion systems are unique pathways to bacteria and are critical for growth and survival of the organism in extreme conditions. Preprotein translocase subunit (secE) of the bacterial secretion system has been demonstrated as target in *Escherichia Coli* (47) and *N. gonorrhoeae* (33). We also found secE to be a potential target for all *Cp* strains but not in *M. tuberculosis* or *C. diphtheriae* (data not shown). Owing to its membrane localization, it may have potential for anti-*Cp* vaccine development.

From host-pathogen common pathways, several proteins were identified as potential drug and vaccine targets in *Cp*. Cytoplasmic enzymes, homoserine dehydrogenase (thrA) and homoserine kinase (thrB), which are involved in glycine, homoserine, threonine, and lysine metabolism pathways, were selected for *Cp*. Both these targets were non-homologous to all five hosts and have previously been identified as targets in *Mycobacterium* (42,48). But thrB is not found to be a target in *C. diphtheriae*.

From the histidine metabolism pathway, imidazole glycerol-phosphate dehydratase (hisB) and ATP phosphoribosyl transferase (hisG) were selected as targets in *Cp*. Both of these enzymes are identified as potential drug targets in many bacteria including *Mycobacterium* and *Pseudomonas* (42,44,49), and nitrobenzothiazole is used as an inhibitor for *M. tuberculosis* hisG (50); however, as *Cp* hisG was a partial horse homolog, it may not be a good target in developing anti-*Cp* drug for wide-ranging number of hosts.

Membrane enzyme phosphoribose diphosphate (ubiA) is involved in glycan biosynthesis and metabolism and, in *Cp*, we identified ubiA as potential target. ubiA is a reported target in *M. tuberculosis* (51), and the disruption of ubiA resulted in a complete loss of cell wall arabinan and death of *C. glutamicum* (52). We have also found that ubiA can also be targeted in *C. diphtheriae*. Being a membrane enzyme, it may also serve the dual purpose.

Cytoplasmic enzyme 1-deoxy-D-xylulose 5-phosphate reductoisomerase (dxr) is essential in the methylerythritol phosphate (MEP) pathway (53). The MEP pathway is extensively targeted in *M. tuberculosis* (53,54), and dxr is a promising target for *Mycobacterium* (55) and *Salmonella* (34). Fosmidomycin is an effective antibiotic that inhibits dxr (56). Our results also suggest that dxr is a potential target in *Cp* for goat, sheep, and bovine, but because of its partial sequence homology with horse, further analysis is required to explore its potentiality in multiple hosts.

Biotin is essential for the growth of various bacteria including *Sinorhizobium meliloti* (57); therefore, biotin biosynthesis pathways are important for bacterial survival and growth. Three cytoplasmic enzymes [biotin synthase family transferase (*Cp1002*_0903), biotin synthase (bioB), and dethiobiotin synthetase (bioD1)] were identified from biotin biosynthesis pathways for *Cp*. However, *Cp1002*_0903 is not found in both *Mycobacterium* and *C. diphtheriae*. bioB was previously reported as an essential gene as well as drug target in *S. typhi* (34) and *A. hydrophila* (41) that are human pathogens. Owing to a PAI-related protein, bioD1 is a good target against *Cp* considering sheep, goat, and bovine but not for horse as it has partial sequence to horse.

From thiamin biosynthesis pathway, thiamine monophosphate kinase (thiL), a cytoplasmic enzyme, is considered. Thus, thiL is a reported target in *M. leprae* (44) and as per our analysis, thiL is a broad-spectrum target (present in all six pathogen considered), which can also be used to develop drug for broad host range (all five hosts in this study).

Cell redox homeostasis is an essential survival mechanism for any intracellular pathogen like *Cp*. Among the several key cytoplasmic enzymes in this pathway, glutaredoxin-like protein (NrdH) was identified as an essential enzyme as well as drug target for *Cp*. NrdH is a novel redoxin in *E. coli* having thioredoxin-like activity (58) and was also found to be a good target in *Mycobacterium*(43). However, in our analysis, this was not found to be a target in *C. diphtheriae*.

FolP is an identified target in *A. hydrophila* (41) and *N. gonorrhoeae* (33). The enzyme catalyzes a condensation reaction yielding dihydropteroate, an intermediary metabolite, that is subsequently converted to tetrahydrofolic acid and is essential for the syntheses of purine, thymidylate, glycine, methionine, pantothenic acid, and *N*-formylmethionyl-tRNA. FolP was found to be a cytoplasmic enzyme and was conserved in *Corynebacterium*. Owing to the fact that the enzyme was found to be essential in the tetrahydrofolate biosynthesis pathway and to be a non-host homolog of *Cp* as applicable to other CMN species and all five hosts, it demonstrated high potentiality as an attractive target, possibly targeted by sulfonamide antibiotics.

Precorrin-4 C11-methyltransferase (cobM) is a crucial enzyme in adenosylcobalamin biosynthesis II, siroheme biosynthesis, and porphyrin and chlorophyll metabolism pathways for all bacteria. It is a potential target in *M. tuberculosis* (42) and *A. hydrophila* (41). In this analysis, we found that cobM is a universal target for all pathogens considered here.

Deoxycytidine triphosphate deaminase (dcd) is an important enzyme in the dUTP and pyrimidine deoxyribonucleotides de novo biosynthesis process. dcd was recently identified as a drug target in *Mycobacterium* (42). Here, we found dcd to be an essential gene in *Cp* and also for *C. diphtheriae* that is non-homologous to all five hosts and also associated with PAI, making dcd an attractive target in all studied pathogens.

Six novel targets consisting of three (rplA, rpmB, and H+ anti-porter subunit-c) from host-pathogen's common pathways and rest three (amino acid career protein, mscL, and resB) from pathogen's unique pathways have been identified. As a result, none except rpmB was found to be universal target because they are not present in *M. tuberculosis*. However, considering *Cp*, all unique pathway-related three targets may be used for developing anti-*Cp* drug as well as vaccine.

Five important broad-spectrum targets (murA, murE, folP, nrdL, and dcd), and *Cp*- and *M. tuberculosis*-specific nrdH were modeled and subjected to virtual screening to identify new molecular agents specific to these targets. We selected these targets because of their potentiality to be targets in other CMN group of human pathogens too, although they are not novel targets for *Cp*. Total 30 compounds, five for each target, have been identified, and one compound [c5(dcd)/c1(nrdL)] was found to be useful in targeting both dcd and nrdL. There is no specific drug available till date to treat *Cp* infection. Therefore, identified compounds can be tested for their efficacy to attain the corresponding targets toward the development of anti-*Cp* and anti-CMN drugs.

Conclusion

In this study, we identified several drug and vaccine targets that are common to four *Cp* strains (*Cp1002*, *Cp*C*231*, *CpFrc41,* and *CpI-19*). Twenty targets were found common to CMN group of pathogens including *Cp* with respect to a broad range of hosts (goat, sheep, bovine, horse, and human). It was also found that some targets can be used for all host ranges, and some are host specific. In general, the peptidoglycan biosynthesis pathway was most important for targeting, followed by ABC-type transport system. Glycan biosynthesis–related ubiA, biotin synthesis pathway enzymes bioB and thiL, cell redox homeostasis regulator NrdH, tetrahydrofolic acid biosynthesis–related folP, and dUTP- and pyrimidine deoxyribonucleotides biosynthesis–related dcd were found to be attractive targets in *Cp* with respect to all considered hosts. We also identified six novel targets that are not reported in any other bacteria, which can be used for broad host range. We also identified potential compounds for our six selected targets using virtual screening. All these targets and identified candidate lead compounds require experimental validation and consideration that the pathogen remains protected inside abscesses, thus proper delivery methods need to be developed. Several targets were found to be strain specific and some were specific to hosts. We have not considered most of the hypothetical proteins because of their strain specificity. These strain- and host-specific targets can be further explored. Currently, we are analyzing hypothetical proteins to enrich the target list. Also, we are adopting fold-level homology modeling and simulation methods for these identified targets and validating to develop broad-spectrum novel drugs and vaccines against CMN group of pathogens for a broad range of hosts.

Acknowledgments

D.B. conceived the method, designed the experiment, coordinated the entire work, and wrote the paper; D.B., N.J., and S.T. annotated and analyzed all genomes for target identification; V.D'A., A.A., A.R.S., L.C.G., S.C.S., A.M. A.S., and V.A. assembled and annotated newly sequenced genomes; D.B., A.B., and L.L. performed virtual screening; A.N.M., A.K., and V.A. monitored and provided technical guided throughout the work.

The work was mostly carried out at IIOAB; D.B., S.T., and N.J. thank IIOAB and its all collaborating Labs; S.T., N.J. had DBT's student fellowship; S.T., N.J., and A.K. acknowledge DBT's Bioinformatics facilities at DAVV; A.N.M. acknowledges the funding of DBT-BIF sanction No.BT/BI/25-001/2006 to ANM; For V.D'A., A.A., A.R.S., L.C.G., S.C.S., A.M. A.S., and V.A., this work was a part of the Rede Paraense de Genômica e Proteômica supported by Fundação de Amparo a Pesquisa do Estado do Pará (FAPESPA). V.A., A.S., and A.M., were supported by Conselho Nacional de Desenvolvimento Científico e Tecnolœgico (CNPq) and also had financial support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and Fundação de Amparo a Pesquisa do Estado de Minas Gerais (FAPEMIG).

References

1. Bayan N., Houssin C., Chami M., Leblon G. (2003) Mycomembrane and S-layer: two important structures of *Corynebacterium glutamicum* cell envelope with promising biotechnology applications. J Biotechnol;104:55–67.
2. Hard G.C. (1969) Electron microscopic examination of *Corynebacterium ovis*. J Bacteriol;97:1480–1485.
3. Songer J.G., Beckenbach K., Marshall M.M., Olson G.B., Kelley L. (1988) Biochemical and genetic characterization of Corynebacterium pseudotuberculosis. Am J Vet Res;49:223–226.
4. Hall V., Collins M.D., Hutson R.A., Lawson P.A., Falsen E., Duerden B.I. (2003) *Corynebacterium atypicum* sp. nov., from a human clinical source, does not contain corynomycolic acids. Int J Syst Evol Microbiol;53:1065–1068.
5. Hard G.C. (1975) Comparative toxic effect of the surface lipid of *Corynebacterium ovis* on peritoneal macrophages. Infect Immun;12:1439–1449.
6. Dorella F.A., Pacheco L.G.C., Oliveira S.C., Miyoshi A., Azevedo V. (2006) Corynebacterium pseudotuberculosis: microbiology, biochemical properties, pathogenesis and molecular studies of virulence. Vet Res;37:201–218.
7. Williamson L.H. (2001) Caseous lymphadenitis in small ruminants. Vet Clin North Am Food Anim Pract;17:359–371.
8. Merchant I.A., Packer R.A. (1967) The genus corynebacterium. In: Merchant I.A., Packer R.A., editors. Veterinary Bacteriology and Virology. Iowa: The Iowa State University Press; p. 425–440.

9. Brown C.C., Olander H.J., Alves S.F. (1987) Synergistic hemolysis-inhibition titers associated with caseous lymphadenitis in a slaughterhouse survey of goats and sheep in Northeastern Brazil. Can J Vet Res;51:46–49.
10. Collett M.G., Bath G.F., Cameron C.M. (1994) Corynebacterium pseudotuberculosis infections. In: Coetzer J.A.W., Thomson G.R., Tustin R.C., Kriek N.P.J., editors. Infectious Diseases of Livestock with Special Reference to Southern Africa. Cape Town: Oxford University Press; p. 1387–1395.
11. Paton M., Walker S., Rose I., Watt G. (2003) Prevalence of caseous lymphadenitis and usage of caseous lymphadenitis vaccines in sheep flocks. Aust Vet J;81:91–95.
12. Unanian M., Silva A.F., Pant K. (1985) Abscesses and caseous lymphadenitis in goats in tropical semi-arid north-east Brazil. Tropic Anim Health Prod, 17:57–62.
13. Yeruham I., Elad D., Van-Ham M., Shpigel N.Y., Perl S. (1997) *Corynebacterium pseudotuberculosis* infection in Israeli cattle: clinical and epidemiological studies. Vet Rec;140:423–427.
14. Ben Saïd M.S., Ben Maitigue H., Benzarti M. *et al.* (2002) Epidemiological and clinical studies of ovine caseous lymphadenitis. Arch Inst Pasteur Tunis;79:51–57.
15. Binns S.H., Bailey M., Green L.E. (2002) Postal survey of ovine caseous lymphadenitis in the United Kingdom between 1990 and 1999. Vet Rec;150:263–268.
16. Connor K.M., Quirie M.M., Baird G., Donachie W. (2000) Characterization of United Kingdom isolates of *Corynebacterium pseudotuberculosis* using pulsed-field gel electrophoresis. J Clin Microbiol;38:2633–2637.
17. Paton M., Rose I., Hart R. *et al.* (1994) New infection with *Corynebacterium pseudotuberculosis* reduces wool production. Aust Vet J;71:47–49.
18. Arsenault J., Girard C., Dubreuil P. *et al.* (2003) Prevalence of and carcass condemnation from maedi-visna, paratuberculosis and caseous lymphadenitis in culled sheep from Quebec, Canada. Prev Vet Med;59:67–81.
19. Ayers J.L. (1977) Caseous lymphadenitis in goats and sheep: a review of diagnosis, pathogenesis, and immunity. J Am Vet Med Assoc;171:1251–1254.
20. Peel M.M., Palmer G.G., Stacpoole A.M., Kerr T.G. (1997) Human lymphadenitis due to *Corynebacterium pseudotuberculosis*: report of ten cases from Australia and review. Clin Infect Dis;24:185–191.
21. Paton M. (1993) Control of cheesy gland in sheep. West Aust J Agric;34:31–37.
22. Menzies P.I., Hwang Y.T., Prescott J.F. (2004) Comparison of an interferon-gamma to a phospholipase D enzyme-linked immunosorbent assay for diagnosis of *Corynebacterium pseudotuberculosis* infection in experimentally infected goats. Vet Microbiol;100:129–137.
23. Fontaine M.C., Baird G., Connor K.M., Rudge K., Sales J. , Donachie W. (2006) Vaccination confers significant protection of sheep against infection with a virulent United Kingdom strain of *Corynebacterium pseudotuberculosis*. Vaccine;24:5986–5996.
24. Piontkowski M.D., Shivvers D.W. (1998) Evaluation of a commercially available vaccine against *Corynebacterium pseudotuberculosis* for use in sheep. J Am Vet Med Assoc;212:1765–1768.
25. Garg D.N., Nain S.P.S., Chandiramani N.K. (1985) Isolation and characterization of *Corynebacterium ovis* from sheep and goats. Indian Vet J;62:805–808.
26. Stanford K., Brogden K.A., McClelland L.A., Kozub G.C., Audibert F. (1998) The incidence of caseous lymphadenitis in Alberta sheep and assessment of impact by vaccination with commercial and experimental vaccines. Can J Vet Res;62:38–43.
27. D'Afonseca V., Prosdocimi F., Dorella F.A. *et al.* (2010) Survey of genome organization and gene content of *Corynebacterium pseudotuberculosis*. Microbiol Res;165:312–320.
28. Silva A., Schneider M.P., Cerdeira L. *et al.* (2010) Complete genome sequence of *Corynebacterium pseudotuberculosis* I-19, strain isolated from Israel Bovine mastitis. J Bacteriol; ;193:323–4.
29. Sakharkar K.R., Sakharkar M.K., Chow V.T.K. (2004) A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas Aeruginosa*. In silico Biol;4:355–360.
30. Dutta A., Singh S.K., Ghosh P. *et al.* (2006) In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. In silico Biol;6:43–47.
31. Chong C.E., Lim B.S., Nathan S., Mohamed R. (2006) In silico analysis of *Burkholderia pseudomallei* genome sequence for potential drug targets. In silico Biol;6:341–346.
32. Asif S.M., Asad A., Faizan A. *et al.* (2009) Dataset of potential targets for *Mycobacterium tuberculosis* H37Rv through comparative genome analysis. Bioinformation;4:245–248.
33. Barh D., Kumar A. (2009) In silico identification of candidate drug and vaccine targets from various pathways in *Neisseria gonorrhoeae*. In silico Biol;9:225–231.
34. Rathi B., Sarangi A.N., Trivedi N. (2009) Genome subtraction for novel target definition in *Salmonella typhi*. Bioinformation;4:143–150.
35. Gish W., States D.J. (1993) Identification of protein coding regions by database similarity search. Nat Genet;3:266–272.
36. Altschul S.F., Gish W., Miller W., Myers E.W., Lipman D.J. (1990) Basic local alignment search tool. J Mol Biol;215:403–410.
37. Zhang R., Lin Y. (2009) DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes. Nucleic Acids Res;37:D455–D458.
38. Kanehisa M., Goto S. (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res;28:27–30.
39. Jones D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol;292:195–202.
40. Irwin J.J., Shoichet B.K. (2005) ZINC – a free database of commercially available compounds for virtual screening. J Chem Inf Model;45:177–182.
41. Sharma V., Gupta P., Dixit A. (2008) In silico identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*. In silico Biol;8:331–338.
42. Anishetty S., Pulimi M., Pennathur G. (2005) Potential drug targets in *Mycobacterium tuberculosis* through metabolic pathway analysis. Comput Biol Chem;29:368–378.
43. Leiting W.U., Jianping X.I.E. (2010) Comparative genomics analysis of Mycobacterium NrdH-redoxins. Microb Pathog;48:97–102.

44. Shanmugam A., Natarajan J. (2010) Computational genome analyses of metabolic enzymes in *Mycobacterium leprae* for drug target identification. Bioinformation;4:392–395.
45. Chhabra G., Sharma P., Anant A. *et al.* (2010) Identification and modeling of a drug target for *Clostridium perfringens* SM101. Bioinformation;4:278–289.
46. Barh D., Misra A.N. (2009) Scientific commons: epitope design from transporter targets in *N. gonorrhoeae*.
47. Driessen A.J.M., Haril U.F., Wickner W. (2003) The enzymology of protein translocation across the *Escherichia coli* plasma membrane.
48. Hasan S., Daugelat S., Rao P.S.S., Schreiber M. (2006) Prioritizing genomic drug targets in pathogens: application to *Mycobacterium tuberculosis*. PLoS Comp Biol;2:e61.
49. Perumal D., Lim C.S., Sakharkar K.R., Sakharkar M.K. (2007) Differential genome analyses of metabolic enzymes in *Pseudomonas aeruginosa* for drug target identification. In silico Biol;7:453–465.
50. Cho Y., Ioerger T.R., Sacchettini J.C. (2008) Discovery of novel nitrobenzothiazole inhibitors for *Mycobacterium tuberculosis* ATP phosphoribosyl transferase (HisG) through virtual screening. J Med Chem;51:5984–5992.
51. Huang H., Berg S., Spencer J.S. *et al.* (2008) Identification of amino acids and domains required for catalytic activity of DPPR synthase, a cell wall biosynthetic enzyme of *Mycobacterium tuberculosis*. Microbiology (Reading, England);154:736–743.
52. Alderwick L.J., Radmacher E., Seidel M. *et al.* (2005) Deletion of Cg-emb in corynebacterianeae leads to a novel truncated cell wall arabinogalactan, whereas inactivation of Cg-ubiA results in an arabinan-deficient mutant with a cell wall galactan core. J Biol Chem;280:32362–32371.
53. Ershov I.V. (2007) 2-C-methylerythritol phosphate pathway of isoprenoid biosynthesis as a target in identifying of new antibiotics, herbicides, and immunomodulators (Review). Prikl Biokhim Mikrobiol;43:133–157.
54. Eoh H., Brennan P.J., Crick D.C. (2009) The *Mycobacterium tuberculosis* MEP (2C-methyl-d-erythritol 4-phosphate) pathway as a new drug target. Tuberculosis (Edinburgh, Scotland);89:1–11.
55. Brown A.C., Parish T. (2008) Dxr is essential in *Mycobacterium tuberculosis* and fosmidomycin resistance is due to a lack of uptake. BMC Microbiol;8:78.
56. Shigi Y. (1989) Inhibition of bacterial isoprenoid synthesis by fosmidomycin, a phosphonic acid-containing antibiotic. J Antimicrobial Chemother;24:131–145.
57. Watson R.J., Heys R., Martin T., Savard M. (2001) Sinorhizobium meliloti cells require biotin and either cobalt or methionine for growth. Appl Environ Microbiol;67:3767–3770.
58. Jordan A., Aslund F., Pontis E., Reichard P., Holmgren A. (1997) Characterization of *Escherichia coli* NrdH. A glutaredoxin-like protein with a thioredoxin-like activity profile. J Biol Chem;272:18044–18050.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Selected 38 common targets in CMN group of pathogens including *Cp*.

Table S2. Comparative 3D modeling data.

Table S3. Hits properties of top five compounds for each selected proteins.

Table S4. Lists the protein residue IDs that are in contact with at least one of the top five compounds.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

6.4 Anexos da seção 3.2.1, “*Análise in silico do panexoproteoma de cinco linhagens de C. pseudotuberculosis*“

As tabelas apresentadas a seguir são parte do artigo científico sobre o exoproteoma preditio *in silico* da *C. pseudotuberculosis* (Seção 3.2.1). Essas tabelas também apresentam resultados detalhados sobre o local subcelular predito para todas as proteínas apresentadas na aplicação da estatística MED em cinco genomas de *C. pseudotuberculosis* (Seção 3.2.5.2).

6.4.1 Pan secretoma predito *in silico*

The *Corynebacterium pseudotuberculosis in silico* predicted pan-exoproteome

Anderson Santos[1], Adriana Carneiro[2], Alfonso Gala-García[1], Anne Pinto[1], Debmalya Barh[3], Eudes Barbosa[1], Flávia Figueira[1], Fernanda Dorella[1], Flávia Souza[1], Louise Cerdeira[2], Luis Guimarães[1], Meritxell Turk[1], Rommel Ramos[2], Sintia Almeida[1], Siomar Soares[1], Ulisses Pereira[1], Vinícius Abreu[1], Artur Silva[2], Anderson Miyoshi[1], Vasco Azevedo[1§]

[1]Molecular and Celular Genetics Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2]DNA Polimorfism Laboratory, Universidade Federal do Pará, Campus do Guamá - Belém, PA, Brazil
[3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India
[§]Corresponding author: vasco@icb.ufmg.br

Additional file 1 – *C. pseudotb* predicted pan secretome

Set = gene coverage based on homology within five strains (1002, C231, I19, FRC41 and PAT10)

Locus tag	**Pan locus**	**Local subcellular**	**Predicted by**	**Set**
Cp1002_0024	plcpsec001	SECRETED	SurfG	5x
CpC231_0022	plcpsec001	SECRETED	SurfG	5x
cpfrc_00026	plcpsec001	SECRETED	SurfG	5x
CpI19_0024	plcpsec001	SECRETED	SurfG	5x
CpPAT10_0024	plcpsec001	SECRETED	SurfG	5x
Cp1002_0035	plcpsec002	SECRETED	SurfG	5x
CpC231_0033	plcpsec002	SECRETED	SurfG	5x
cpfrc_00037	plcpsec002	SECRETED	SurfG	5x
CpI19_0035	plcpsec002	SECRETED	SurfG	5x
CpPAT10_0035	plcpsec002	SECRETED	SurfG	5x
Cp1002_0038	plcpsec003	SECRETED	SurfG	5x
CpC231_0036	plcpsec003	SECRETED	SurfG	5x
cpfrc_00040	plcpsec003	SECRETED	SurfG	5x
CpI19_0038	plcpsec003	SECRETED	SurfG	5x
CpPAT10_0038	plcpsec003	SECRETED	SurfG	5x
Cp1002_0165	plcpsec004	SECRETED	SurfG	5x
CpC231_0168	plcpsec004	SECRETED	SurfG	5x
cpfrc_00167	plcpsec004	SECRETED	SurfG	5x
CpI19_0167	plcpsec004	SECRETED	SurfG	5x
CpPAT10_0168	plcpsec004	SECRETED	SurfG	5x
Cp1002_0185	plcpsec005	SECRETED	SurfG	5x
CpC231_0188	plcpsec005	SECRETED	SurfG	5x
cpfrc_00184	plcpsec005	SECRETED	SurfG	5x
CpI19_0187	plcpsec005	SECRETED	SurfG	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpPAT10_0188	plcpsec005	SECRETED	SurfG	5x
Cp1002_0193	plcpsec006	SECRETED	SurfG	5x
CpC231_0196	plcpsec006	SECRETED	SurfG	5x
cpfrc_00192	plcpsec006	SECRETED	SurfG	5x
CpI19_0195	plcpsec006	SECRETED	SurfG	5x
CpPAT10_0196	plcpsec006	SECRETED	SurfG	5x
Cp1002_0198	plcpsec007	SECRETED	SurfG	5x
CpC231_0201	plcpsec007	SECRETED	SurfG	5x
cpfrc_00198	plcpsec007	SECRETED	SurfG	5x
CpI19_0200	plcpsec007	SECRETED	SurfG	5x
CpPAT10_0204	plcpsec007	SECRETED	SurfG	5x
Cp1002_0200	plcpsec008	SECRETED	SurfG	5x
CpC231_0203	plcpsec008	SECRETED	SurfG	5x
cpfrc_00200	plcpsec008	SECRETED	SurfG	5x
CpI19_0202	plcpsec008	SECRETED	SurfG	5x
CpPAT10_0206	plcpsec008	SECRETED	SurfG	5x
Cp1002_0208	plcpsec009	SECRETED	SurfG	5x
CpC231_0211	plcpsec009	SECRETED	SurfG	5x
cpfrc_00208	plcpsec009	SECRETED	SurfG	5x
CpI19_0210	plcpsec009	SECRETED	SurfG	5x
CpPAT10_0214	plcpsec009	SECRETED	SurfG	5x
Cp1002_0221	plcpsec010	SECRETED	SurfG	5x
CpC231_0224	plcpsec010	SECRETED	SurfG	5x
cpfrc_00221	plcpsec010	SECRETED	SurfG	5x
CpI19_0223	plcpsec010	SECRETED	SurfG	5x
CpPAT10_0227	plcpsec010	SECRETED	SurfG	5x
Cp1002_0231	plcpsec011	SECRETED	SurfG	5x
CpC231_0234	plcpsec011	SECRETED	SurfG	5x
cpfrc_00231	plcpsec011	SECRETED	SurfG	5x
CpI19_0233	plcpsec011	SECRETED	SurfG	5x
CpPAT10_0237	plcpsec011	SECRETED	SurfG	5x
Cp1002_0237	plcpsec012	SECRETED	SurfG	5x
CpC231_0240	plcpsec012	SECRETED	SurfG	5x
cpfrc_00237	plcpsec012	SECRETED	SurfG	5x
CpI19_0239	plcpsec012	SECRETED	SurfG	5x
CpPAT10_0243	plcpsec012	SECRETED	SurfG	5x
Cp1002_0249	plcpsec013	SECRETED	SurfG	5x
CpC231_0252	plcpsec013	SECRETED	SurfG	5x
cpfrc_00248	plcpsec013	SECRETED	SurfG	5x
CpI19_0251	plcpsec013	SECRETED	SurfG	5x
CpPAT10_0254	plcpsec013	SECRETED	SurfG	5x
Cp1002_0269	plcpsec014	SECRETED	SurfG	5x
CpC231_0272	plcpsec014	SECRETED	SurfG	5x
cpfrc_00266	plcpsec014	SECRETED	SurfG	5x
CpI19_0271	plcpsec014	SECRETED	SurfG	5x
CpPAT10_0274	plcpsec014	SECRETED	SurfG	5x
Cp1002_0292	plcpsec015	SECRETED	SurfG	5x
CpC231_0295	plcpsec015	SECRETED	SurfG	5x
cpfrc_00289	plcpsec015	SECRETED	SurfG	5x
CpI19_0294	plcpsec015	SECRETED	SurfG	5x
CpPAT10_0297	plcpsec015	SECRETED	SurfG	5x
Cp1002_0368	plcpsec016	SECRETED	SurfG	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpC231_0371	plcpsec016	SECRETED	SurfG	5x
cpfrc_00366	plcpsec016	SECRETED	SurfG	5x
CpI19_0370	plcpsec016	SECRETED	SurfG	5x
CpPAT10_0372	plcpsec016	SECRETED	SurfG	5x
Cp1002_0376	plcpsec017	SECRETED	SurfG	5x
CpC231_0379	plcpsec017	SECRETED	SurfG	5x
cpfrc_00374	plcpsec017	SECRETED	SurfG	5x
CpI19_0378	plcpsec017	SECRETED	SurfG	5x
CpPAT10_0380	plcpsec017	SECRETED	SurfG	5x
Cp1002_0388	plcpsec018	SECRETED	SurfG	5x
CpC231_0391	plcpsec018	SECRETED	SurfG	5x
cpfrc_00387	plcpsec018	SECRETED	SurfG	5x
CpI19_0390	plcpsec018	SECRETED	SurfG	5x
CpPAT10_0392	plcpsec018	SECRETED	SurfG	5x
Cp1002_0415	plcpsec019	SECRETED	SurfG	5x
CpC231_0418	plcpsec019	SECRETED	SurfG	5x
cpfrc_00415	plcpsec019	SECRETED	SurfG	5x
CpI19_0416	plcpsec019	SECRETED	SurfG	5x
CpPAT10_0419	plcpsec019	SECRETED	SurfG	5x
Cp1002_0535	plcpsec020	SECRETED	SurfG	5x
CpC231_0538	plcpsec020	SECRETED	SurfG	5x
cpfrc_00536	plcpsec020	SECRETED	SurfG	5x
CpI19_0537	plcpsec020	SECRETED	SurfG	5x
CpPAT10_0537	plcpsec020	SECRETED	SurfG	5x
Cp1002_0536	plcpsec021	SECRETED	SurfG	5x
CpC231_0539	plcpsec021	SECRETED	SurfG	5x
cpfrc_00537	plcpsec021	SECRETED	SurfG	5x
CpI19_0538	plcpsec021	SECRETED	SurfG	5x
CpPAT10_0538	plcpsec021	SECRETED	SurfG	5x
Cp1002_0549	plcpsec022	SECRETED	SurfG	5x
CpC231_0552	plcpsec022	SECRETED	SurfG	5x
cpfrc_00550	plcpsec022	SECRETED	SurfG	5x
CpI19_0551	plcpsec022	SECRETED	SurfG	5x
CpPAT10_0551	plcpsec022	SECRETED	SurfG	5x
Cp1002_0567	plcpsec023	SECRETED	SurfG	5x
CpC231_0569	plcpsec023	SECRETED	SurfG	5x
cpfrc_00567	plcpsec023	SECRETED	SurfG	5x
CpI19_0568	plcpsec023	SECRETED	SurfG	5x
CpPAT10_0568	plcpsec023	SECRETED	SurfG	5x
Cp1002_0573	plcpsec024	SECRETED	SurfG	5x
CpC231_0575	plcpsec024	SECRETED	SurfG	5x
cpfrc_00574	plcpsec024	SECRETED	SurfG	5x
CpI19_0574	plcpsec024	SECRETED	SurfG	5x
CpPAT10_0574	plcpsec024	SECRETED	SurfG	5x
Cp1002_0594	plcpsec025	SECRETED	SurfG	5x
CpC231_0595	plcpsec025	SECRETED	SurfG	5x
cpfrc_00594	plcpsec025	SECRETED	SurfG	5x
CpI19_0594	plcpsec025	SECRETED	SurfG	5x
CpPAT10_0595	plcpsec025	SECRETED	SurfG	5x
Cp1002_0596	plcpsec026	SECRETED	SurfG	5x
CpC231_0597	plcpsec026	SECRETED	SurfG	5x
cpfrc_00597	plcpsec026	SECRETED	SurfG	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpI19_0596	plcpsec026	SECRETED	SurfG	5x
CpPAT10_0598	plcpsec026	SECRETED	SurfG	5x
Cp1002_0615	plcpsec027	SECRETED	SurfG	5x
CpC231_0615	plcpsec027	SECRETED	SurfG	5x
cpfrc_00617	plcpsec027	SECRETED	SurfG	5x
CpI19_0614	plcpsec027	SECRETED	SurfG	5x
CpPAT10_0616	plcpsec027	SECRETED	SurfG	5x
Cp1002_0666	plcpsec028	SECRETED	SurfG	5x
CpC231_0665	plcpsec028	SECRETED	SurfG	5x
cpfrc_00665	plcpsec028	SECRETED	SurfG	5x
CpI19_0665	plcpsec028	SECRETED	SurfG	5x
CpPAT10_0666	plcpsec028	SECRETED	SurfG	5x
Cp1002_0681	plcpsec029	SECRETED	SurfG	5x
CpC231_0680	plcpsec029	SECRETED	SurfG	5x
cpfrc_00679	plcpsec029	SECRETED	SurfG	5x
CpI19_0680	plcpsec029	SECRETED	SurfG	5x
CpPAT10_0681	plcpsec029	SECRETED	SurfG	5x
Cp1002_0686	plcpsec030	SECRETED	SurfG	5x
CpC231_0685	plcpsec030	SECRETED	SurfG	5x
cpfrc_00684	plcpsec030	SECRETED	SurfG	5x
CpI19_0685	plcpsec030	SECRETED	SurfG	5x
CpPAT10_0686	plcpsec030	SECRETED	SurfG	5x
Cp1002_0720	plcpsec031	SECRETED	SurfG	5x
CpC231_0719	plcpsec031	SECRETED	SurfG	5x
cpfrc_00720	plcpsec031	SECRETED	SurfG	5x
CpI19_0719	plcpsec031	SECRETED	SurfG	5x
CpPAT10_0718	plcpsec031	SECRETED	SurfG	5x
Cp1002_0763	plcpsec032	SECRETED	SurfG	5x
CpC231_0763	plcpsec032	SECRETED	SurfG	5x
cpfrc_00763	plcpsec032	SECRETED	SurfG	5x
CpI19_0763	plcpsec032	SECRETED	SurfG	5x
CpPAT10_0762	plcpsec032	SECRETED	SurfG	5x
Cp1002_0766	plcpsec033	SECRETED	SurfG	5x
CpC231_0766	plcpsec033	SECRETED	SurfG	5x
cpfrc_00766	plcpsec033	SECRETED	SurfG	5x
CpI19_0766	plcpsec033	SECRETED	SurfG	5x
CpPAT10_0765	plcpsec033	SECRETED	SurfG	5x
Cp1002_0783	plcpsec034	SECRETED	SurfG	5x
CpC231_0783	plcpsec034	SECRETED	SurfG	5x
cpfrc_00783	plcpsec034	SECRETED	SurfG	5x
CpI19_0783	plcpsec034	SECRETED	SurfG	5x
CpPAT10_0782	plcpsec034	SECRETED	SurfG	5x
Cp1002_0824	plcpsec035	SECRETED	SurfG	5x
CpC231_0826	plcpsec035	SECRETED	SurfG	5x
cpfrc_00826	plcpsec035	SECRETED	SurfG	5x
CpI19_0826	plcpsec035	SECRETED	SurfG	5x
CpPAT10_0824	plcpsec035	SECRETED	SurfG	5x
Cp1002_0882	plcpsec036	SECRETED	SurfG	5x
CpC231_0884	plcpsec036	SECRETED	SurfG	5x
cpfrc_00884	plcpsec036	SECRETED	SurfG	5x
CpI19_0885	plcpsec036	SECRETED	SurfG	5x
CpPAT10_0883	plcpsec036	SECRETED	SurfG	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
Cp1002_0893	plcpsec037	SECRETED	SurfG	5x
CpC231_0895	plcpsec037	SECRETED	SurfG	5x
cpfrc_00897	plcpsec037	SECRETED	SurfG	5x
CpI19_0896	plcpsec037	SECRETED	SurfG	5x
CpPAT10_0894	plcpsec037	SECRETED	SurfG	5x
Cp1002_0911	plcpsec038	SECRETED	SurfG	5x
CpC231_0915	plcpsec038	SECRETED	SurfG	5x
cpfrc_00916	plcpsec038	SECRETED	SurfG	5x
CpI19_0916	plcpsec038	SECRETED	SurfG	5x
CpPAT10_0912	plcpsec038	SECRETED	SurfG	5x
Cp1002_1000	plcpsec039	SECRETED	SurfG	5x
CpC231_0999	plcpsec039	SECRETED	SurfG	5x
cpfrc_01006	plcpsec039	SECRETED	SurfG	5x
CpI19_1005	plcpsec039	SECRETED	SurfG	5x
CpPAT10_0999	plcpsec039	SECRETED	SurfG	5x
Cp1002_1013	plcpsec040	SECRETED	SurfG	5x
CpC231_1012	plcpsec040	SECRETED	SurfG	5x
cpfrc_01018	plcpsec040	SECRETED	SurfG	5x
CpI19_1018	plcpsec040	SECRETED	SurfG	5x
CpPAT10_1012	plcpsec040	SECRETED	SurfG	5x
Cp1002_1068	plcpsec041	SECRETED	SurfG	5x
CpC231_1066	plcpsec041	SECRETED	SurfG	5x
cpfrc_01074	plcpsec041	SECRETED	SurfG	5x
CpI19_1073	plcpsec041	SECRETED	SurfG	5x
CpPAT10_1067	plcpsec041	SECRETED	SurfG	5x
Cp1002_1144	plcpsec042	SECRETED	SurfG	5x
CpC231_1143	plcpsec042	SECRETED	SurfG	5x
cpfrc_01148	plcpsec042	SECRETED	SurfG	5x
CpI19_1150	plcpsec042	SECRETED	SurfG	5x
CpPAT10_1143	plcpsec042	SECRETED	SurfG	5x
Cp1002_1234	plcpsec043	SECRETED	SurfG	5x
CpC231_1233	plcpsec043	SECRETED	SurfG	5x
cpfrc_01241	plcpsec043	SECRETED	SurfG	5x
CpI19_1240	plcpsec043	SECRETED	SurfG	5x
CpPAT10_1233	plcpsec043	SECRETED	SurfG	5x
Cp1002_1298	plcpsec044	SECRETED	SurfG	5x
CpC231_1297	plcpsec044	SECRETED	SurfG	5x
cpfrc_01302	plcpsec044	SECRETED	SurfG	5x
CpI19_1303	plcpsec044	SECRETED	SurfG	5x
CpPAT10_1296	plcpsec044	SECRETED	SurfG	5x
Cp1002_1345	plcpsec045	SECRETED	SurfG	5x
CpC231_1344	plcpsec045	SECRETED	SurfG	5x
cpfrc_01351	plcpsec045	SECRETED	SurfG	5x
CpI19_1350	plcpsec045	SECRETED	SurfG	5x
CpPAT10_1344	plcpsec045	SECRETED	SurfG	5x
Cp1002_1378	plcpsec046	SECRETED	SurfG	5x
CpC231_1377	plcpsec046	SECRETED	SurfG	5x
cpfrc_01384	plcpsec046	SECRETED	SurfG	5x
CpI19_1383	plcpsec046	SECRETED	SurfG	5x
CpPAT10_1377	plcpsec046	SECRETED	SurfG	5x
Cp1002_1389	plcpsec047	SECRETED	SurfG	5x
CpC231_1388	plcpsec047	SECRETED	SurfG	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
cpfrc_01395	plcpsec047	SECRETED	SurfG	5x
CpI19_1394	plcpsec047	SECRETED	SurfG	5x
CpPAT10_1388	plcpsec047	SECRETED	SurfG	5x
Cp1002_1416	plcpsec048	SECRETED	SurfG	5x
CpC231_1416	plcpsec048	SECRETED	SurfG	5x
cpfrc_01421	plcpsec048	SECRETED	SurfG	5x
CpI19_1423	plcpsec048	SECRETED	SurfG	5x
CpPAT10_1415	plcpsec048	SECRETED	SurfG	5x
Cp1002_1417	plcpsec049	SECRETED	SurfG	5x
CpC231_1417	plcpsec049	SECRETED	SurfG	5x
cpfrc_01422	plcpsec049	SECRETED	SurfG	5x
CpI19_1424	plcpsec049	SECRETED	SurfG	5x
CpPAT10_1416	plcpsec049	SECRETED	SurfG	5x
Cp1002_1476	plcpsec050	SECRETED	SurfG	5x
CpC231_1478	plcpsec050	SECRETED	SurfG	5x
cpfrc_01485	plcpsec050	SECRETED	SurfG	5x
CpI19_1485	plcpsec050	SECRETED	SurfG	5x
CpPAT10_1478	plcpsec050	SECRETED	SurfG	5x
Cp1002_1506	plcpsec051	SECRETED	SurfG	5x
CpC231_1509	plcpsec051	SECRETED	SurfG	5x
cpfrc_01516	plcpsec051	SECRETED	SurfG	5x
CpI19_1515	plcpsec051	SECRETED	SurfG	5x
CpPAT10_1509	plcpsec051	SECRETED	SurfG	5x
Cp1002_1631	plcpsec052	SECRETED	SurfG	5x
CpC231_1632	plcpsec052	SECRETED	SurfG	5x
cpfrc_01634	plcpsec052	SECRETED	SurfG	5x
CpI19_1639	plcpsec052	SECRETED	SurfG	5x
CpPAT10_1632	plcpsec052	SECRETED	SurfG	5x
Cp1002_1745	plcpsec053	SECRETED	SurfG	5x
CpC231_1737	plcpsec053	SECRETED	SurfG	5x
cpfrc_01745	plcpsec053	SECRETED	SurfG	5x
CpI19_1753	plcpsec053	SECRETED	SurfG	5x
CpPAT10_1746	plcpsec053	SECRETED	SurfG	5x
Cp1002_1765	plcpsec054	SECRETED	SurfG	5x
CpC231_1756	plcpsec054	SECRETED	SurfG	5x
cpfrc_01764	plcpsec054	SECRETED	SurfG	5x
CpI19_1773	plcpsec054	SECRETED	SurfG	5x
CpPAT10_1766	plcpsec054	SECRETED	SurfG	5x
Cp1002_1772	plcpsec055	SECRETED	SurfG	5x
CpC231_1762	plcpsec055	SECRETED	SurfG	5x
cpfrc_01770	plcpsec055	SECRETED	SurfG	5x
CpI19_1780	plcpsec055	SECRETED	SurfG	5x
CpPAT10_1772	plcpsec055	SECRETED	SurfG	5x
Cp1002_1802	plcpsec056	SECRETED	SurfG	5x
CpC231_1792	plcpsec056	SECRETED	SurfG	5x
cpfrc_01799	plcpsec056	SECRETED	SurfG	5x
CpI19_1810	plcpsec056	SECRETED	SurfG	5x
CpPAT10_1802	plcpsec056	SECRETED	SurfG	5x
Cp1002_1815	plcpsec057	SECRETED	SurfG	5x
CpC231_1807	plcpsec057	SECRETED	SurfG	5x
cpfrc_01813	plcpsec057	SECRETED	SurfG	5x
CpI19_1825	plcpsec057	SECRETED	SurfG	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpPAT10_1817	plcpsec057	SECRETED	SurfG	5x
Cp1002_1820	plcpsec058	SECRETED	SurfG	5x
CpC231_1812	plcpsec058	SECRETED	SurfG	5x
cpfrc_01818	plcpsec058	SECRETED	SurfG	5x
CpI19_1830	plcpsec058	SECRETED	SurfG	5x
CpPAT10_1822	plcpsec058	SECRETED	SurfG	5x
Cp1002_1843	plcpsec059	SECRETED	SurfG	5x
CpC231_1836	plcpsec059	SECRETED	SurfG	5x
cpfrc_01843	plcpsec059	SECRETED	SurfG	5x
CpI19_1854	plcpsec059	SECRETED	SurfG	5x
CpPAT10_1846	plcpsec059	SECRETED	SurfG	5x
Cp1002_1847	plcpsec060	SECRETED	SurfG	5x
CpC231_1840	plcpsec060	SECRETED	SurfG	5x
cpfrc_01847	plcpsec060	SECRETED	SurfG	5x
CpI19_1858	plcpsec060	SECRETED	SurfG	5x
CpPAT10_1850	plcpsec060	SECRETED	SurfG	5x
Cp1002_1852	plcpsec061	SECRETED	SurfG	5x
CpC231_1845	plcpsec061	SECRETED	SurfG	5x
cpfrc_01852	plcpsec061	SECRETED	SurfG	5x
CpI19_1863	plcpsec061	SECRETED	SurfG	5x
CpPAT10_1855	plcpsec061	SECRETED	SurfG	5x
Cp1002_1864	plcpsec062	SECRETED	SurfG	5x
CpC231_1857	plcpsec062	SECRETED	SurfG	5x
cpfrc_01866	plcpsec062	SECRETED	SurfG	5x
CpI19_1875	plcpsec062	SECRETED	SurfG	5x
CpPAT10_1868	plcpsec062	SECRETED	SurfG	5x
Cp1002_1893	plcpsec063	SECRETED	SurfG	5x
CpC231_1885	plcpsec063	SECRETED	SurfG	5x
cpfrc_01894	plcpsec063	SECRETED	SurfG	5x
CpI19_1905	plcpsec063	SECRETED	SurfG	5x
CpPAT10_1895	plcpsec063	SECRETED	SurfG	5x
Cp1002_1894	plcpsec064	SECRETED	SurfG	5x
CpC231_1886	plcpsec064	SECRETED	SurfG	5x
cpfrc_01895	plcpsec064	SECRETED	SurfG	5x
CpI19_1906	plcpsec064	SECRETED	SurfG	5x
CpPAT10_1896	plcpsec064	SECRETED	SurfG	5x
Cp1002_1948	plcpsec065	SECRETED	SurfG	5x
CpC231_1942	plcpsec065	SECRETED	SurfG	5x
cpfrc_01951	plcpsec065	SECRETED	SurfG	5x
CpI19_1963	plcpsec065	SECRETED	SurfG	5x
CpPAT10_1955	plcpsec065	SECRETED	SurfG	5x
Cp1002_1955	plcpsec066	SECRETED	SurfG	5x
CpC231_1949	plcpsec066	SECRETED	SurfG	5x
cpfrc_01958	plcpsec066	SECRETED	SurfG	5x
CpI19_1970	plcpsec066	SECRETED	SurfG	5x
CpPAT10_1962	plcpsec066	SECRETED	SurfG	5x
Cp1002_1957	plcpsec067	SECRETED	SurfG	5x
CpC231_1951	plcpsec067	SECRETED	SurfG	5x
cpfrc_01960	plcpsec067	SECRETED	SurfG	5x
CpI19_1972	plcpsec067	SECRETED	SurfG	5x
CpPAT10_1964	plcpsec067	SECRETED	SurfG	5x
Cp1002_1976	plcpsec068	SECRETED	SurfG	5x

Locus tag	**Pan locus**	**Local subcellular**	**Predicted by**	**Set**
CpC231_1970	plcpsec068	SECRETED	SurfG	5x
cpfrc_01980	plcpsec068	SECRETED	SurfG	5x
CpI19_1991	plcpsec068	SECRETED	SurfG	5x
CpPAT10_1983	plcpsec068	SECRETED	SurfG	5x
Cp1002_2055	plcpsec069	SECRETED	SurfG	5x
CpC231_2049	plcpsec069	SECRETED	SurfG	5x
cpfrc_02056	plcpsec069	SECRETED	SurfG	5x
CpI19_2070	plcpsec069	SECRETED	SurfG	5x
CpPAT10_2059	plcpsec069	SECRETED	SurfG	5x
Cp1002_2064	plcpsec070	SECRETED	SurfG	5x
CpC231_2058	plcpsec070	SECRETED	SurfG	5x
cpfrc_02065	plcpsec070	SECRETED	SurfG	5x
CpI19_2079	plcpsec070	SECRETED	SurfG	5x
CpPAT10_2068	plcpsec070	SECRETED	SurfG	5x
Cp1002_2069	plcpsec071	SECRETED	SurfG	5x
CpC231_2063	plcpsec071	SECRETED	SurfG	5x
cpfrc_02070	plcpsec071	SECRETED	SurfG	5x
CpI19_2084	plcpsec071	SECRETED	SurfG	5x
CpPAT10_2073	plcpsec071	SECRETED	SurfG	5x
Cp1002_2081	plcpsec072	SECRETED	SurfG	5x
CpC231_2074	plcpsec072	SECRETED	SurfG	5x
cpfrc_02081	plcpsec072	SECRETED	SurfG	5x
CpI19_2095	plcpsec072	SECRETED	SurfG	5x
CpPAT10_2084	plcpsec072	SECRETED	SurfG	5x
Cp1002_0126a	plcpsec074	SECRETED	SurfG	5x
CpC231_0129	plcpsec074	SECRETED	SurfG	5x
cpfrc_00129	plcpsec074	SECRETED	SurfG	5x
CpI19_0129	plcpsec074	SECRETED	SurfG	5x
CpPAT10_0128	plcpsec074	SECRETED	SurfG	5x
Cp1002_0279	plcpsec076	SECRETED	SurfG	5x
CpC231_0282	plcpsec076	SECRETED	SurfG	5x
cpfrc_00276	plcpsec076	SECRETED	SurfG	5x
CpI19_0281	plcpsec076	SECRETED	SurfG	5x
CpPAT10_0284	plcpsec076	SECRETED	SurfG	5x
Cp1002_0699	plcpsec078	SECRETED	SurfG	5x
CpC231_0698	plcpsec078	SECRETED	SurfG	5x
cpfrc_00699	plcpsec078	SECRETED	SurfG	5x
CpI19_0698	plcpsec078	SECRETED	SurfG	5x
CpPAT10_0699	plcpsec078	SECRETED	SurfG	5x
Cp1002_1657	plcpsec082	SECRETED	SurfG	5x
CpC231_1658	plcpsec082	SECRETED	SurfG	5x
cpfrc_01658	plcpsec082	SECRETED	SurfG	5x
CpI19_1666	plcpsec082	SECRETED	SurfG	5x
CpPAT10_1657	plcpsec082	SECRETED	SurfG	5x
Cp1002_1716	plcpsec084	SECRETED	SurfG	5x
CpC231_1708	plcpsec084	SECRETED	SurfG	5x
cpfrc_01715	plcpsec084	SECRETED	SurfG	5x
CpI19_1724	plcpsec084	SECRETED	SurfG	5x
CpPAT10_1716	plcpsec084	SECRETED	SurfG	5x
Cp1002_0027	plcpsec086	SECRETED	SurfG	5x
CpC231_0025	plcpsec086	SECRETED	SurfG	5x
cpfrc_00029	plcpsec086	SECRETED	SurfG	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpI19_0027	plcpsec086	SECRETED	SurfG	5x
CpPAT10_0027	plcpsec086	SECRETED	SurfG	5x
Cp1002_0113	plcpsec088	SECRETED	SurfG	5x
CpC231_0116	plcpsec088	SECRETED	SurfG	5x
cpfrc_00116	plcpsec088	SECRETED	SurfG	5x
CpI19_0116	plcpsec088	SECRETED	SurfG	5x
CpPAT10_0115	plcpsec088	SECRETED	SurfG	5x
Cp1002_0593	plcpsec089	SECRETED	SurfG	5x
CpC231_0593	plcpsec089	SECRETED	SurfG	5x
cpfrc_00593	plcpsec089	SECRETED	SurfG	5x
CpI19_0592	plcpsec089	SECRETED	SurfG	5x
CpPAT10_0593	plcpsec089	SECRETED	SurfG	5x
Cp1002_1143	plcpsec090	SECRETED	SurfG	5x
CpC231_1142	plcpsec090	SECRETED	SurfG	5x
cpfrc_01147	plcpsec090	SECRETED	SurfG	5x
CpI19_1149	plcpsec090	SECRETED	SurfG	5x
CpPAT10_1142	plcpsec090	SECRETED	SurfG	5x
Cp1002_1669	plcpsec093	SECRETED	SurfG	5x
CpC231_1669a	plcpsec093	SECRETED	SurfG	5x
cpfrc_01667a	plcpsec093	SECRETED	SurfG	5x
CpI19_1678	plcpsec093	SECRETED	SurfG	5x
CpPAT10_1669	plcpsec093	SECRETED	SurfG	5x
Cp1002_1888	plcpsec094	SECRETED	SurfG	5x
CpC231_1880	plcpsec094	SECRETED	SurfG	5x
cpfrc_01889a	plcpsec094	SECRETED	SurfG	5x
CpI19_1900	plcpsec094	SECRETED	SurfG	5x
CpPAT10_1890a	plcpsec094	SECRETED	SurfG	5x
Cp1002_0031	plcpsec073	SECRETED	SurfG	4x
CpC231_0029	plcpsec073	SECRETED	SurfG	4x
cpfrc_00033	plcpsec073	PSE E	SurfG	4x
CpI19_0031	plcpsec073	SECRETED	SurfG	4x
CpPAT10_0031	plcpsec073	SECRETED	SurfG	4x
Cp1002_0182	plcpsec075	SECRETED	SurfG	4x
CpC231_0185	plcpsec075	SECRETED	SurfG	4x
cpfrc_00181	plcpsec075	SECRETED	SurfG	4x
CpI19_0184	plcpsec075	PSEUDOGENE	SurfG	4x
CpPAT10_0185	plcpsec075	SECRETED	SurfG	4x
Cp1002_0387	plcpsec077	SECRETED	SurfG	4x
CpC231_0390	plcpsec077	SECRETED	SurfG	4x
cpfrc_00386	plcpsec077	SECRETED	SurfG	4x
CpI19_0389	plcpsec077	SECRETED	SurfG	4x
CpPAT10_0391	plcpsec077	CYTOPLASMIC	SurfG	4x
Cp1002_0713	plcpsec079	SECRETED	SurfG	4x
CpC231_0712	plcpsec079	SECRETED	SurfG	4x
cpfrc_00713	plcpsec079	PSE C	SurfG	4x
CpI19_0711	plcpsec079	SECRETED	SurfG	4x
CpPAT10_0711	plcpsec079	SECRETED	SurfG	4x
Cp1002_0813	plcpsec080	SECRETED	SurfG	4x
CpC231_0815	plcpsec080	SECRETED	SurfG	4x
cpfrc_00815	plcpsec080	SECRETED	SurfG	4x
CpI19_0815	plcpsec080	SECRETED	SurfG	4x
CpPAT10_0813	plcpsec080	PSE C	SurfG	4x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
Cp1002_1156	plcpsec081	SECRETED	SurfG	4x
CpC231_1155	plcpsec081	SECRETED	SurfG	4x
cpfrc_01159	plcpsec081	SECRETED	SurfG	4x
CpI19_1162	plcpsec081	SECRETED	SurfG	4x
CpPAT10_1154	plcpsec081	CYTOPLASMIC	SurfG	4x
Cp1002_1688	plcpsec083	SECRETED	SurfG	4x
CpC231	plcpsec083	NOTFOUND	SurfG	4x
cpfrc_01688	plcpsec083	SECRETED	SurfG	4x
CpI19_1696	plcpsec083	SECRETED	SurfG	4x
CpPAT10_1688	plcpsec083	SECRETED	SurfG	4x
Cp1002_1868	plcpsec085	SECRETED	SurfG	5x
CpC231_1862	plcpsec085	SECRETED	SurfG	5x
cpfrc_01871	plcpsec085	SECRETED	SurfG	5x
CpI19_1879	plcpsec085	SECRETED	SurfG	5x
CpPAT10_1873	plcpsec085	SECRETED	SurfG	5x
Cp1002_1811a	plcpsec091	SECRETED	SurfG	4x
CpC231_1803	plcpsec091	SECRETED	SurfG	4x
cpfrc_01809	plcpsec091	SECRETED	SurfG	4x
CpI19_1821	plcpsec091	SECRETED	SurfG	4x
CpPAT10_1813	plcpsec091	CYTOPLASMIC	SurfG	4x
Cp1002_0096	plcpsec087	SECRETED	SurfG	3x
CpC231_0097	plcpsec087	SECRETED	SurfG	3x
cpfrc_00098	plcpsec087	MEMBRANE	SurfG	3x
CpI19_0098	plcpsec087	PSE C	SurfG	3x
CpPAT10_0096	plcpsec087	SECRETED	SurfG	3x
Cp1002_1651	plcpsec092	SECRETED	SurfG	3x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpC231_1652	plcpsec092	SECRETED	SurfG	3x
cpfrc_01652	plcpsec092	CYTOPLASMIC	SurfG	3x
CpI19_1660	plcpsec092	SECRETED	SurfG	3x
CpPAT10_1651	plcpsec092	CYTOPLASMIC	SurfG	3x
Cp1002_1971	plcpsec095	SECRETED	SurfG	3x
CpC231_1965	plcpsec095	SECRETED	SurfG	3x
cpfrc_01975	plcpsec095	SECRETED	SurfG	3x
CpI19_1986	plcpsec095	CYTOPLASMIC	SurfG	3x
CpPAT10_1978	plcpsec095	CYTOPLASMIC	SurfG	3x
Cp1002_0014	plcpsec099	CYTOPLASMIC	SurfG	3x
CpC231_0012	plcpsec099	SECRETED	SurfG	3x
Cpfrc_00012	plcpsec099	CYTOPLASMIC	SurfG	3x
CpI19_0014	plcpsec099	SECRETED	SurfG	3x
CpPAT10_0014	plcpsec099	SECRETED	SurfG	3x
Cp1002_0510	plcpsec101	PSE C	SurfG	3x
CpC231_0514	plcpsec101	SECRETED	SurfG	3x
cpfrc_00513	plcpsec101	SECRETED	SurfG	3x
CpI19_0513	plcpsec101	PSE C	SurfG	3x
CpPAT10_0513	plcpsec101	SECRETED	SurfG	3x
Cp1002_0065	plcpsec096	SECRETED	SurfG	2x
CpC231_0064	plcpsec096	PSE C	SurfG	2x
cpfrc_00067	plcpsec096	SECRETED	SurfG	2x
CpI19_0065	plcpsec096	PSE C	SurfG	2x
CpPAT10_0066	plcpsec096	PSE C	SurfG	2x
Cp1002_1763	plcpsec097	SECRETED	SurfG	2x
CpC231_1754	plcpsec097	SECRETED	SurfG	2x

Locus tag	**Pan locus**	**Local subcellular**	**Predicted by**	**Set**
cpfrc_01762	plcpsec097	PSE C	SurfG	2x
CpI19_1771	plcpsec097	PSE C	SurfG	2x
CpPAT10_1764	plcpsec097	PSE E	SurfG	2x
Cp1002_1797	plcpsec098	SECRETED	SurfG	2x
CpC231_1787	plcpsec098	SECRETED	SurfG	2x
cpfrc_01795	plcpsec098	PSE C	SurfG	2x
CpI19_1805	plcpsec098	PSE C	SurfG	2x
CpPAT10_1797	plcpsec098	PSE C	SurfG	2x
Cp1002_0369	plcpsec100	PSEUDOGENE	SurfG	2x
CpC231_0372	plcpsec100	SECRETED	SurfG	2x
cpfrc_00367	plcpsec100	CYTOPLASMIC	SurfG	2x
CpI19_0371	plcpsec100	SECRETED	SurfG	2x
CpPAT10_0373	plcpsec100	PSE C	SurfG	2x
Cp1002_0903	plcpsec102	CYTOPLASMIC	SurfG	2x
CpC231_0905	plcpsec102	SECRETED	SurfG	2x
cpfrc_00907	plcpsec102	SECRETED	SurfG	2x
CpI19_0906	plcpsec102	PSEUDOGENE	SurfG	2x
CpPAT10_0904	plcpsec102	CYTOPLASMIC	SurfG	2x
Cp1002_1310	plcpsec104	PSE C	SurfG	1x
CpC231_1309	plcpsec104	SECRETED	SurfG	1x
cpfrc_01315	plcpsec104	PSE C	SurfG	1x
CpI19_1315	plcpsec104	PSE C	SurfG	1x
CpPAT10_1309	plcpsec104	PSE C	SurfG	1x
Cp1002_0102	plcpsec106	SECRETED	TatP	5x
CpC231_0103	plcpsec106	SECRETED	TatP	5x
cpfrc_00104	plcpsec106	SECRETED	TatP	5x

Locus tag	**Pan locus**	**Local subcellular**	**Predicted by**	**Set**
CpI19_0104	plcpsec106	SECRETED	TatP	5x
CpPAT10_0102	plcpsec106	SECRETED	TatP	5x
Cp1002_0172	plcpsec108	SECRETED	TatP	5x
CpC231_0175	plcpsec108	SECRETED	TatP	5x
cpfrc_00174	plcpsec108	SECRETED	TatP	5x
CpI19_0174	plcpsec108	SECRETED	TatP	5x
CpPAT10_0175	plcpsec108	SECRETED	TatP	5x
Cp1002_0502	plcpsec110	SECRETED	TatP	5x
CpC231_0506	plcpsec110	SECRETED	TatP	5x
cpfrc_00506	plcpsec110	SECRETED	TatP	5x
CpI19_0505	plcpsec110	SECRETED	TatP	5x
CpPAT10_0505	plcpsec110	SECRETED	TatP	5x
Cp1002_0505	plcpsec111	SECRETED	TatP	5x
CpC231_0509	plcpsec111	SECRETED	TatP	5x
cpfrc_00508	plcpsec111	SECRETED	TatP	5x
CpI19_0508	plcpsec111	SECRETED	TatP	5x
CpPAT10_0508	plcpsec111	SECRETED	TatP	5x
Cp1002_0664	plcpsec112	SECRETED	TatP	5x
CpC231_0636	plcpsec112	SECRETED	TatP	5x
CpC231_0663	plcpsec112	SECRETED	TatP	5x
CpI19_0663	plcpsec112	SECRETED	TatP	5x
CpPAT10_0664	plcpsec112	SECRETED	TatP	5x
Cp1002_0705	plcpsec113	SECRETED	TatP	5x
CpC231_0704	plcpsec113	SECRETED	TatP	5x
cpfrc_00705	plcpsec113	SECRETED	TatP	5x
CpI19_0704	plcpsec113	SECRETED	TatP	5x

Locus tag	**Pan locus**	**Local subcellular**	**Predicted by**	**Set**
CpPAT10_0704	plcpsec113	SECRETED	TatP	5x
Cp1002_0940	plcpsec115	SECRETED	TatP	5x
CpC231_0942	plcpsec115	SECRETED	TatP	5x
cpfrc_00945	plcpsec115	SECRETED	TatP	5x
CpI19_0945	plcpsec115	SECRETED	TatP	5x
CpPAT10_0941	plcpsec115	SECRETED	TatP	5x
Cp1002_0972	plcpsec116	SECRETED	TatP	5x
CpC231_0973	plcpsec116	SECRETED	TatP	5x
cpfrc_00978	plcpsec116	SECRETED	TatP	5x
CpI19_0977	plcpsec116	SECRETED	TatP	5x
CpPAT10_0972	plcpsec116	SECRETED	TatP	5x
Cp1002_1051	plcpsec118	SECRETED	TatP	5x
CpC231_1049	plcpsec118	SECRETED	TatP	5x
cpfrc_01056	plcpsec118	SECRETED	TatP	5x
CpI19_1056	plcpsec118	SECRETED	TatP	5x
CpPAT10_1050	plcpsec118	SECRETED	TatP	5x
Cp1002_1117	plcpsec119	SECRETED	TatP	5x
CpC231_1116	plcpsec119	SECRETED	TatP	5x
cpfrc_01121	plcpsec119	SECRETED	TatP	5x
CpI19_1123	plcpsec119	SECRETED	TatP	5x
CpPAT10_1116	plcpsec119	SECRETED	TatP	5x
Cp1002_1137	plcpsec120	SECRETED	TatP	5x
CpC231_1136	plcpsec120	SECRETED	TatP	5x
cpfrc_01141	plcpsec120	SECRETED	TatP	5x
CpI19_1143	plcpsec120	SECRETED	TatP	5x
CpPAT10_1136	plcpsec120	SECRETED	TatP	5x

Locus tag	**Pan locus**	**Local subcellular**	**Predicted by**	**Set**
Cp1002_1187	plcpsec121	SECRETED	TatP	5x
CpC231_1186	plcpsec121	SECRETED	TatP	5x
cpfrc_01191	plcpsec121	SECRETED	TatP	5x
CpI19_1193	plcpsec121	SECRETED	TatP	5x
CpPAT10_1185	plcpsec121	SECRETED	TatP	5x
Cp1002_1262	plcpsec122	SECRETED	TatP	5x
CpC231_1261	plcpsec122	SECRETED	TatP	5x
cpfrc_01267	plcpsec122	SECRETED	TatP	5x
CpI19_1268	plcpsec122	SECRETED	TatP	5x
CpPAT10_1260	plcpsec122	SECRETED	TatP	5x
Cp1002_1296	plcpsec123	SECRETED	TatP	5x
CpC231_1295	plcpsec123	SECRETED	TatP	5x
cpfrc_01300	plcpsec123	SECRETED	TatP	5x
CpI19_1301	plcpsec123	SECRETED	TatP	5x
CpPAT10_1294	plcpsec123	SECRETED	TatP	5x
Cp1002_1387	plcpsec124	SECRETED	TatP	5x
CpC231_1386	plcpsec124	SECRETED	TatP	5x
cpfrc_01393	plcpsec124	SECRETED	TatP	5x
CpI19_1392	plcpsec124	SECRETED	TatP	5x
CpPAT10_1386	plcpsec124	SECRETED	TatP	5x
Cp1002_1757	plcpsec128	SECRETED	TatP	5x
CpC231_1749	plcpsec128	SECRETED	TatP	5x
cpfrc_01757	plcpsec128	SECRETED	TatP	5x
CpI19_1765	plcpsec128	SECRETED	TatP	5x
CpPAT10_1758	plcpsec128	SECRETED	TatP	5x
Cp1002_1786	plcpsec129	SECRETED	TatP	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpC231_1776	plcpsec129	SECRETED	TatP	5x
cpfrc_01784	plcpsec129	SECRETED	TatP	5x
CpI19_1794	plcpsec129	SECRETED	TatP	5x
CpPAT10_1786	plcpsec129	SECRETED	TatP	5x
Cp1002_0132	plcpsec107	CYTOPLASMIC	TatP	4x
CpC231_0135	plcpsec107	SECRETED	TatP	4x
cpfrc_00135	plcpsec107	SECRETED	TatP	4x
CpI19_0135	plcpsec107	SECRETED	TatP	4x
CpPAT10_0137	plcpsec107	SECRETED	TatP	4x
Cp1002_0252	plcpsec109	SECRETED	TatP	4x
CpC231_0255	plcpsec109	SECRETED	TatP	4x
Cpfrc_00251	plcpsec109	CYTOPLASMIC	TatP	4x
CpI19_0254	plcpsec109	SECRETED	TatP	4x
CpPAT10_0257	plcpsec109	SECRETED	TatP	4x
Cp1002_1497	plcpsec125	PSEUDOGENE	TatP	4x
CpC231_1499	plcpsec125	SECRETED	TatP	4x
cpfrc_01506	plcpsec125	SECRETED	TatP	4x
CpI19_1505	plcpsec125	SECRETED	TatP	4x
CpPAT10_1499	plcpsec125	SECRETED	TatP	4x
Cp1002_1004	plcpsec117	SECRETED	TatP	3x
CpC231_1003	plcpsec117	SECRETED	TatP	3x
cpfrc_01010	plcpsec117	SECRETED	TatP	3x
CpI19_1009	plcpsec117	CYTOPLASMIC	TatP	3x
CpPAT10_1003	plcpsec117	CYTOPLASMIC	TatP	3x
Cp1002_1755	plcpsec127	CYTOPLASMIC	TatP	3x
CpC231_1747	plcpsec127	SECRETED	TatP	3x
cpfrc_01755	plcpsec127	CYTOPLASMIC	TatP	3x
CpI19_1763	plcpsec127	SECRETED	TatP	3x
CpPAT10_1756	plcpsec127	SECRETED	TatP	3x
Cp1002_0835	plcpsec114	CYTOPLASMIC	TatP	2x
CpC231_0837	plcpsec114	SECRETED	TatP	2x
cpfrc_00837	plcpsec114	SECRETED	TatP	2x
CpI19_0837	plcpsec114	CYTOPLASMIC	TatP	2x
CpPAT10_0835	plcpsec114	CYTOPLASMIC	TatP	2x
Cp1002_0884	plcpsec130	PSEUDOGENE	TatP	2x
CpC231_0886	plcpsec130	PSEUDOGENE	TatP	2x
cpfrc_00886	plcpsec130	PSEUDOGENE	TatP	2x
CpI19_0887	plcpsec130	SECRETED	TatP	2x
CpPAT10_0885	plcpsec130	SECRETED	TatP	2x
Cp1002_1527	plcpsec126	CYTOPLASMIC	TatP	1x
CpC231_1530	plcpsec126	PSEUDOGENE	TatP	1x
cpfrc_01536	plcpsec126	SECRETED	TatP	1x
CpI19_1536	plcpsec126	PSEUDOGENE	TatP	1x
CpPAT10_1530	plcpsec126	PSEUDOGENE	TatP	1x
Cp1002_0048	plcpsec127	SECRETED	SecP	5x
CpC231_0046	plcpsec127	SECRETED	SecP	5x
cpfrc_00050	plcpsec127	SECRETED	SecP	5x
CpI19_0048	plcpsec127	SECRETED	SecP	5x
CpPAT10_0048	plcpsec127	SECRETED	SecP	5x
Cp1002_0058	plcpsec128	SECRETED	SecP	5x
CpC231_0057	plcpsec128	SECRETED	SecP	5x
cpfrc_00060	plcpsec128	SECRETED	SecP	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpI19_0058	plcpsec128	SECRETED	SecP	5x
CpPAT10_0059	plcpsec128	SECRETED	SecP	5x
Cp1002_0485	plcpsec129	SECRETED	SecP	5x
CpC231_0489	plcpsec129	SECRETED	SecP	5x
cpfrc_00490	plcpsec129	SECRETED	SecP	5x
CpI19_0488	plcpsec129	SECRETED	SecP	5x
CpPAT10_0490	plcpsec129	SECRETED	SecP	5x
Cp1002_0630	plcpsec130	SECRETED	SecP	5x
CpC231_0630	plcpsec130	SECRETED	SecP	5x
cpfrc_00631	plcpsec130	SECRETED	SecP	5x
CpI19_0629	plcpsec130	SECRETED	SecP	5x
CpPAT10_0631	plcpsec130	SECRETED	SecP	5x
Cp1002_0708	plcpsec131	SECRETED	SecP	5x
CpC231_0707	plcpsec131	SECRETED	SecP	5x
Cpfrc_00707a	plcpsec131	SECRETED	SecP	5x
CpI19_0706a	plcpsec131	SECRETED	SecP	5x
CpPAT10_0706a	plcpsec131	SECRETED	SecP	5x
Cp1002_0988	plcpsec132	SECRETED	SurfG	5x
CpC231_0989	plcpsec132	SECRETED	SurfG	5x
cpfrc_00995	plcpsec132	SECRETED	SurfG	5x
CpI19_0993	plcpsec132	SECRETED	SurfG	5x
CpPAT10_0988	plcpsec132	SECRETED	SurfG	5x
Cp1002_0988a	plcpsec133	SECRETED	SecP	5x
CpC231_0989a	plcpsec133	SECRETED	SecP	5x
cpfrc_00996	plcpsec133	SECRETED	SecP	5x
CpI19_0993a	plcpsec133	SECRETED	SecP	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
CpPAT10_0988a	plcpsec133	SECRETED	SecP	5x
Cp1002_1034	plcpsec134	SECRETED	SecP	5x
CpC231_1033	plcpsec134	SECRETED	SecP	5x
cpfrc_01038	plcpsec134	SECRETED	SecP	5x
CpI19_1039	plcpsec134	SECRETED	SecP	5x
CpPAT10_1033	plcpsec134	SECRETED	SecP	5x
Cp1002_1082	plcpsec135	SECRETED	SecP	5x
CpC231_1081	plcpsec135	SECRETED	SecP	5x
cpfrc_01086a	plcpsec135	SECRETED	SecP	5x
CpI19_1088	plcpsec135	SECRETED	SecP	5x
CpPAT10_1081	plcpsec135	SECRETED	SecP	5x
Cp1002_1146	plcpsec136	SECRETED	SecP	5x
CpC231_1145	plcpsec136	SECRETED	SecP	5x
cpfrc_01149a	plcpsec136	SECRETED	SecP	5x
CpI19_1152	plcpsec136	SECRETED	SecP	5x
CpPAT10_1144a	plcpsec136	SECRETED	SecP	5x
Cp1002_1159	plcpsec137	SECRETED	SecP	5x
CpC231_1158	plcpsec137	SECRETED	SecP	5x
cpfrc_01162	plcpsec137	SECRETED	SecP	5x
CpI19_1165	plcpsec137	SECRETED	SecP	5x
CpPAT10_1157	plcpsec137	SECRETED	SecP	5x
Cp1002_1208	plcpsec138	SECRETED	SecP	5x
CpC231_1207	plcpsec138	SECRETED	SecP	5x
cpfrc_01214	plcpsec138	SECRETED	SecP	5x
CpI19_1214	plcpsec138	SECRETED	SecP	5x
CpPAT10_1207	plcpsec138	SECRETED	SecP	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
Cp1002_1401	plcpsec139	SECRETED	SecP	5x
CpC231_1400	plcpsec139	SECRETED	SecP	5x
cpfrc_01406a	plcpsec139	SECRETED	SecP	5x
CpI19_1407	plcpsec139	SECRETED	SecP	5x
CpPAT10_1400	plcpsec139	SECRETED	SecP	5x
Cp1002_1462	plcpsec140	SECRETED	SecP	5x
CpC231_1464	plcpsec140	SECRETED	SecP	5x
cpfrc_01472	plcpsec140	SECRETED	SecP	5x
CpI19_1471	plcpsec140	SECRETED	SecP	5x
CpPAT10_1465	plcpsec140	SECRETED	SecP	5x
Cp1002_1483	plcpsec141	SECRETED	SecP	5x
CpC231_1485	plcpsec141	SECRETED	SecP	5x
cpfrc_01492	plcpsec141	SECRETED	SecP	5x
CpI19_1492	plcpsec141	SECRETED	SecP	5x
CpPAT10_1485	plcpsec141	SECRETED	SecP	5x
Cp1002_1667	plcpsec142	SECRETED	SecP	5x
CpC231_1668	plcpsec142	SECRETED	SecP	5x
cpfrc_01666b	plcpsec142	SECRETED	SecP	5x
CpI19_1676	plcpsec142	SECRETED	SecP	5x
CpPAT10_1667	plcpsec142	SECRETED	SecP	5x
Cp1002_1668	plcpsec143	SECRETED	SecP	5x
CpC231_1669	plcpsec143	SECRETED	SecP	5x
cpfrc_01667	plcpsec143	SECRETED	SecP	5x
CpI19_1677	plcpsec143	SECRETED	SecP	5x
CpPAT10_1668	plcpsec143	SECRETED	SecP	5x

Locus tag	Pan locus	Local subcellular	Predicted by	Set
Cp1002_1721	plcpsec144	SECRETED	SecP	5x
CpC231_1713	plcpsec144	SECRETED	SecP	5x
cpfrc_01720	plcpsec144	SECRETED	SecP	5x
CpI19_1729	plcpsec144	SECRETED	SecP	5x
CpPAT10_1721	plcpsec144	SECRETED	SecP	5x
Cp1002_1751	plcpsec145	SECRETED	SecP	5x
CpC231_1743	plcpsec145	SECRETED	SecP	5x
cpfrc_01751	plcpsec145	SECRETED	SecP	5x
CpI19_1759	plcpsec145	SECRETED	SecP	5x
CpPAT10_1752	plcpsec145	SECRETED	SecP	5x
Cp1002_1923	plcpsec146	SECRETED	SecP	5x
CpC231_1917	plcpsec146	SECRETED	SecP	5x
cpfrc_01928a	plcpsec146	SECRETED	SecP	5x
CpI19_1938	plcpsec146	SECRETED	SecP	5x
CpPAT10_1930	plcpsec146	SECRETED	SecP	5x
Cp1002_2014a	plcpsec148	SECRETED	SecP	5x
CpC231_2009	plcpsec148	SECRETED	SecP	5x
cpfrc_002020a	plcpsec148	SECRETED	SecP	5x
CpI19_2030a	plcpsec148	SECRETED	SecP	5x
CpPAT10_2023	plcpsec148	SECRETED	SecP	5x
Cp1002_1867	plcpsec147	PSE RN	SecP	1x
CpC231_1861	plcpsec147	SECRETED	SecP	1x
cpfrc_01870	plcpsec147	PSE RN	SecP	1x
CpI19_1878	plcpsec147	PSE RN	SecP	1x
CpPAT10_1872	plcpsec147	PSE RN	SecP	1x

6.4.2 Pan superficioma predito *in silico*

The *Corynebacterium pseudotuberculosis in silico* predicted pan-exoproteome

Anderson Santos[1], Adriana Carneiro[2], Alfonso Gala-García[1], Anne Pinto[1], Debmalya Barh[3], Eudes Barbosa[1], Flávia Figueira[1], Fernanda Dorella[1], Flávia Souza[1], Louise Cerdeira[2], Luis Guimarães[1], Meritxell Turk[1], Rommel Ramos[2], Sintia Almeida[1], Siomar Soares[1], Ulisses Pereira[1], Vinícius Abreu[1], Artur Silva[2], Anderson Miyoshi[1], Vasco Azevedo[1§]

[1]Molecular and Celular Genetics Laboratory, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil
[2]DNA Polimorfism Laboratory, Universidade Federal do Pará, Campus do Guamá - Belém, PA, Brazil
[3]Centre for Genomics and Applied Gene Technology, Institute of Integrative Omics and Applied Biotechnology, Nonakuri, Purba Medinipur, West Bengal, India
[§]Corresponding author: vasco@icb.ufmg.br

Additional file 2 – *C. pseudotb* predicted pan surfaceome

Set = gene coverage based on homology within five strains (1002, C231, I19, FRC41 and PAT10)

Locus tag	Pan locus	Set	Local subcellular
Cp1002_0016	plcppse001	5x	PSE E
CpC231_0014	plcppse001	5x	PSE E
cpfrc_00014	plcppse001	5x	PSE E
CpI19_0016	plcppse001	5x	PSE E
CpPAT10_0016	plcppse001	5x	PSE E
Cp1002_0033	plcppse002	5x	PSE C
CpC231_0031	plcppse002	5x	PSE C
cpfrc_00035	plcppse002	5x	PSE C
CpI19_0033	plcppse002	5x	PSE C
CpPAT10_0033	plcppse002	5x	PSE C
Cp1002_0037	plcppse003	5x	PSE N
CpC231_0035	plcppse003	5x	PSE N
cpfrc_00039	plcppse003	5x	PSE N
CpI19_0037	plcppse003	5x	PSE N
CpPAT10_0037	plcppse003	5x	PSE N
Cp1002_0043	plcppse004	5x	PSE E
CpC231_0041	plcppse004	5x	PSE E
cpfrc_00045	plcppse004	5x	PSE E
CpI19_0043	plcppse004	5x	PSE E
CpPAT10_0043	plcppse004	5x	PSE E
Cp1002_0046	plcppse005	5x	PSE N
CpC231_0044	plcppse005	5x	PSE N
cpfrc_00048	plcppse005	5x	PSE N
CpI19_0046	plcppse005	5x	PSE N
CpPAT10_0046	plcppse005	5x	PSE N
Cp1002_0053	plcppse006	5x	PSE C
CpC231_0052	plcppse006	5x	PSE C
cpfrc_00055	plcppse006	5x	PSE C
CpI19_0053	plcppse006	5x	PSE C
CpPAT10_0054	plcppse006	5x	PSE C
Cp1002_0064	plcppse007	5x	PSE E
CpC231_0063	plcppse007	5x	PSE E
cpfrc_00066	plcppse007	5x	PSE E
CpI19_0064	plcppse007	5x	PSE E
CpPAT10_0065	plcppse007	5x	PSE E
Cp1002_0072	plcppse008	5x	PSE L
CpC231_0072	plcppse008	5x	PSE L
cpfrc_00074	plcppse008	5x	PSE L
CpI19_0073	plcppse008	5x	PSE L
CpPAT10_0073	plcppse008	5x	PSE L
Cp1002_0077	plcppse009	5x	PSE N
CpC231_0077	plcppse009	5x	PSE N
cpfrc_00079	plcppse009	5x	PSE N
CpI19_0078	plcppse009	5x	PSE N
CpPAT10_0078	plcppse009	5x	PSE N
Cp1002_0079	plcppse010	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_0079	plcppse010	5x	PSE E
cpfrc_00081	plcppse010	5x	PSE E
CpI19_0080	plcppse010	5x	PSE E
CpPAT10_0080	plcppse010	5x	PSE E
Cp1002_0099	plcppse011	5x	PSE C
CpC231_0100	plcppse011	5x	PSE C
cpfrc_00101	plcppse011	5x	PSE C
CpI19_0101	plcppse011	5x	PSE C
CpPAT10_0099	plcppse011	5x	PSE C
Cp1002_0125	plcppse012	5x	PSE L
CpC231_0127	plcppse012	5x	PSE L
cpfrc_00128	plcppse012	5x	PSE L
CpI19_0128	plcppse012	5x	PSE L
CpPAT10_0127	plcppse012	5x	PSE L
Cp1002_0134	plcppse013	5x	PSE N
CpC231_0137	plcppse013	5x	PSE N
cpfrc_00136	plcppse013	5x	PSE N
CpI19_0136	plcppse013	5x	PSE N
CpPAT10_0135	plcppse013	5x	PSE N
Cp1002_0164	plcppse014	5x	PSE RN
CpC231_0167	plcppse014	5x	PSE RN
cpfrc_00166	plcppse014	5x	PSE RN
CpI19_0166	plcppse014	5x	PSE RN
CpPAT10_0167	plcppse014	5x	PSE RN
Cp1002_0166	plcppse015	5x	PSE RN
CpC231_0169	plcppse015	5x	PSE RN
cpfrc_00168	plcppse015	5x	PSE RN
CpI19_0168	plcppse015	5x	PSE RN
CpPAT10_0169	plcppse015	5x	PSE RN
Cp1002_0170	plcppse016	5x	PSE RL
CpC231_0173	plcppse016	5x	PSE RN
cpfrc_00172	plcppse016	5x	PSE RN
CpI19_0172	plcppse016	5x	PSE RL
CpPAT10_0173	plcppse016	5x	PSE RL
Cp1002_0192	plcppse017	5x	PSE N
CpC231_0195	plcppse017	5x	PSE N
cpfrc_00191	plcppse017	5x	PSE N
CpI19_0194	plcppse017	5x	PSE N
CpPAT10_0195	plcppse017	5x	PSE N
Cp1002_0210	plcppse018	5x	PSE C
CpC231_0213	plcppse018	5x	PSE C
cpfrc_00210	plcppse018	5x	PSE C
CpI19_0212	plcppse018	5x	PSE C
CpPAT10_0216	plcppse018	5x	PSE C
Cp1002_0214	plcppse019	5x	PSE N
CpC231_0217	plcppse019	5x	PSE N
cpfrc_00214	plcppse019	5x	PSE N
CpI19_0216	plcppse019	5x	PSE N
CpPAT10_0220	plcppse019	5x	PSE N
Cp1002_0227	plcppse020	5x	PSE C
CpC231_0230	plcppse020	5x	PSE C
cpfrc_00227	plcppse020	5x	PSE C
CpI19_0229	plcppse020	5x	PSE C
CpPAT10_0233	plcppse020	5x	PSE C
Cp1002_0259	plcppse021	5x	PSE C
CpC231_0262	plcppse021	5x	PSE C
cpfrc_00258	plcppse021	5x	PSE C
CpI19_0261	plcppse021	5x	PSE C
CpPAT10_0264	plcppse021	5x	PSE C
Cp1002_0284	plcppse022	5x	PSE E
CpC231_0287	plcppse022	5x	PSE E
cpfrc_00281	plcppse022	5x	PSE E
CpI19_0286	plcppse022	5x	PSE E
CpPAT10_0289	plcppse022	5x	PSE E
Cp1002_0286	plcppse023	5x	PSE L
CpC231_0289	plcppse023	5x	PSE L
cpfrc_00283	plcppse023	5x	PSE L
CpI19_0288	plcppse023	5x	PSE L
CpPAT10_0291	plcppse023	5x	PSE L
Cp1002_0289	plcppse024	5x	PSE E
CpC231_0292	plcppse024	5x	PSE E
cpfrc_00286	plcppse024	5x	PSE E
CpI19_0291	plcppse024	5x	PSE E
CpPAT10_0294	plcppse024	5x	PSE E
Cp1002_0315	plcppse025	5x	PSE RN
CpC231_0319	plcppse025	5x	PSE RN
cpfrc_00313	plcppse025	5x	PSE RN
CpI19_0318	plcppse025	5x	PSE RN
CpPAT10_0320	plcppse025	5x	PSE RN
Cp1002_0317	plcppse026	5x	PSE N

Locus tag	Pan locus	Set	Local subcellular
CpC231_0321	plcppse026	5x	PSE N
cpfrc_00315	plcppse026	5x	PSE N
CpI19_0320	plcppse026	5x	PSE N
CpPAT10_0322	plcppse026	5x	PSE N
Cp1002_0320	plcppse027	5x	PSE N
CpC231_0324	plcppse027	5x	PSE N
cpfrc_00318	plcppse027	5x	PSE N
CpI19_0323	plcppse027	5x	PSE N
CpPAT10_0325	plcppse027	5x	PSE N
Cp1002_0321	plcppse028	5x	PSE L
CpC231_0325	plcppse028	5x	PSE L
cpfrc_00319	plcppse028	5x	PSE L
CpI19_0324	plcppse028	5x	PSE L
CpPAT10_0326	plcppse028	5x	PSE L
Cp1002_0325	plcppse029	5x	PSE C
CpC231_0328	plcppse029	5x	PSE C
cpfrc_00322	plcppse029	5x	PSE C
CpI19_0327	plcppse029	5x	PSE C
CpPAT10_0329	plcppse029	5x	PSE C
Cp1002_0357	plcppse030	5x	PSE E
CpC231_0360	plcppse030	5x	PSE E
cpfrc_00355	plcppse030	5x	PSE E
CpI19_0359	plcppse030	5x	PSE E
CpPAT10_0362	plcppse030	5x	PSE E
Cp1002_0377	plcppse031	5x	PSE E
CpC231_0380	plcppse031	5x	PSE E
cpfrc_00375	plcppse031	5x	PSE E
CpI19_0379	plcppse031	5x	PSE E
CpPAT10_0381	plcppse031	5x	PSE E
Cp1002_0396	plcppse032	5x	PSE C
CpC231_0399	plcppse032	5x	PSE C
cpfrc_00395	plcppse032	5x	PSE C
CpI19_0398	plcppse032	5x	PSE C
CpPAT10_0400	plcppse032	5x	PSE C
Cp1002_0398	plcppse033	5x	PSE N
CpC231_0401	plcppse033	5x	PSE N
cpfrc_00397	plcppse033	5x	PSE N
CpI19_0400	plcppse033	5x	PSE N
CpPAT10_0402	plcppse033	5x	PSE N
Cp1002_0401	plcppse034	5x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_0404	plcppse034	5x	PSE C
cpfrc_00399	plcppse034	5x	PSE C
CpI19_0402	plcppse034	5x	PSE C
CpPAT10_0404	plcppse034	5x	PSE C
Cp1002_0402	plcppse035	5x	PSE C
CpC231_0405	plcppse035	5x	PSE C
cpfrc_00400	plcppse035	5x	PSE C
CpI19_0403	plcppse035	5x	PSE C
CpPAT10_0405	plcppse035	5x	PSE C
Cp1002_0422	plcppse036	5x	PSE C
CpC231_0425	plcppse036	5x	PSE C
cpfrc_00424	plcppse036	5x	PSE C
CpI19_0423	plcppse036	5x	PSE C
CpPAT10_0426	plcppse036	5x	PSE C
Cp1002_0429	plcppse037	5x	PSE C
CpC231_0432	plcppse037	5x	PSE C
cpfrc_00432	plcppse037	5x	PSE C
CpI19_0430	plcppse037	5x	PSE C
CpPAT10_0434	plcppse037	5x	PSE C
Cp1002_0432	plcppse038	5x	PSE E
CpC231_0435	plcppse038	5x	PSE E
cpfrc_00435	plcppse038	5x	PSE E
CpI19_0433	plcppse038	5x	PSE E
CpPAT10_0437	plcppse038	5x	PSE E
Cp1002_0436	plcppse039	5x	PSE E
CpC231_0439	plcppse039	5x	PSE E
cpfrc_00439	plcppse039	5x	PSE E
CpI19_0437	plcppse039	5x	PSE E
CpPAT10_0441	plcppse039	5x	PSE E
Cp1002_0439	plcppse040	5x	PSE E
CpC231_0443	plcppse040	5x	PSE E
cpfrc_00443	plcppse040	5x	PSE E
CpI19_0441	plcppse040	5x	PSE E
CpPAT10_0444	plcppse040	5x	PSE E
Cp1002_0450	plcppse041	5x	PSE N
CpC231_0454	plcppse041	5x	PSE N
cpfrc_00454	plcppse041	5x	PSE N
CpI19_0452	plcppse041	5x	PSE N
CpPAT10_0455	plcppse041	5x	PSE N
Cp1002_0451	plcppse042	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_0455	plcppse042	5x	PSE E
cpfrc_00455	plcppse042	5x	PSE E
CpI19_0454	plcppse042	5x	PSE E
CpPAT10_0456	plcppse042	5x	PSE E
Cp1002_0463	plcppse043	5x	PSE N
CpC231_0467	plcppse043	5x	PSE N
cpfrc_00468	plcppse043	5x	PSE N
CpI19_0466	plcppse043	5x	PSE N
CpPAT10_0468	plcppse043	5x	PSE N
Cp1002_0480	plcppse044	5x	PSE C
CpC231_0484	plcppse044	5x	PSE C
cpfrc_00485	plcppse044	5x	PSE C
CpI19_0483	plcppse044	5x	PSE C
CpPAT10_0485	plcppse044	5x	PSE C
Cp1002_0486	plcppse045	5x	PSE N
CpC231_0490	plcppse045	5x	PSE N
cpfrc_00491	plcppse045	5x	PSE N
CpI19_0489	plcppse045	5x	PSE N
CpPAT10_0491	plcppse045	5x	PSE N
Cp1002_0497	plcppse046	5x	PSE E
CpC231_0501	plcppse046	5x	PSE E
cpfrc_00502	plcppse046	5x	PSE E
CpI19_0500	plcppse046	5x	PSE E
CpPAT10_0501	plcppse046	5x	PSE E
Cp1002_0499	plcppse047	5x	PSE C
CpC231_0503	plcppse047	5x	PSE C
cpfrc_00504	plcppse047	5x	PSE C
CpI19_0502	plcppse047	5x	PSE C
CpPAT10_0503	plcppse047	5x	PSE C
Cp1002_0516	plcppse048	5x	PSE C
CpC231_0520	plcppse048	5x	PSE C
cpfrc_00519	plcppse048	5x	PSE C
CpI19_0519	plcppse048	5x	PSE C
CpPAT10_0519	plcppse048	5x	PSE C
Cp1002_0539	plcppse049	5x	PSE C
CpC231_0542	plcppse049	5x	PSE C
cpfrc_00540	plcppse049	5x	PSE C
CpI19_0541	plcppse049	5x	PSE C
CpPAT10_0541	plcppse049	5x	PSE C
Cp1002_0550	plcppse050	5x	PSE E
CpC231_0553	plcppse050	5x	PSE E
cpfrc_00551	plcppse050	5x	PSE E
CpI19_0552	plcppse050	5x	PSE E
CpPAT10_0552	plcppse050	5x	PSE E
Cp1002_0552	plcppse051	5x	PSE C
CpC231_0555	plcppse051	5x	PSE C
cpfrc_00553	plcppse051	5x	PSE C
CpI19_0554	plcppse051	5x	PSE C
CpPAT10_0554	plcppse051	5x	PSE C
Cp1002_0560	plcppse052	5x	PSE C
CpC231_0563	plcppse052	5x	PSE C
cpfrc_00561	plcppse052	5x	PSE C
CpI19_0562	plcppse052	5x	PSE C
CpPAT10_0562	plcppse052	5x	PSE C
Cp1002_0562	plcppse053	5x	PSE C
CpC231_0564	plcppse053	5x	PSE C
cpfrc_00562	plcppse053	5x	PSE C
CpI19_0563	plcppse053	5x	PSE C
CpPAT10_0563	plcppse053	5x	PSE C
Cp1002_0565	plcppse054	5x	PSE N
CpC231_0567	plcppse054	5x	PSE N
cpfrc_00565	plcppse054	5x	PSE N
CpI19_0566	plcppse054	5x	PSE N
CpPAT10_0566	plcppse054	5x	PSE N
Cp1002_0581	plcppse055	5x	PSE L
CpC231_0582	plcppse055	5x	PSE L
cpfrc_00581	plcppse055	5x	PSE L
CpI19_0582	plcppse055	5x	PSE L
CpPAT10_0581	plcppse055	5x	PSE L
Cp1002_0584	plcppse056	5x	PSE E
CpC231_0585	plcppse056	5x	PSE E
cpfrc_00583	plcppse056	5x	PSE E
CpI19_0584	plcppse056	5x	PSE E
CpPAT10_0584	plcppse056	5x	PSE E
Cp1002_0585	plcppse057	5x	PSE E
CpC231_0586	plcppse057	5x	PSE E
cpfrc_00585	plcppse057	5x	PSE E
CpI19_0585	plcppse057	5x	PSE E
CpPAT10_0585	plcppse057	5x	PSE E
Cp1002_0607	plcppse058	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_0608	plcppse058	5x	PSE E
cpfrc_00610	plcppse058	5x	PSE E
CpI19_0607	plcppse058	5x	PSE E
CpPAT10_0609	plcppse058	5x	PSE E
Cp1002_0623	plcppse059	5x	PSE E
CpC231_0623	plcppse059	5x	PSE E
cpfrc_00625	plcppse059	5x	PSE E
CpI19_0622	plcppse059	5x	PSE E
CpPAT10_0624	plcppse059	5x	PSE E
Cp1002_0643	plcppse060	5x	PSE C
CpC231_0642	plcppse060	5x	PSE C
cpfrc_00643	plcppse060	5x	PSE C
CpI19_0642	plcppse060	5x	PSE C
CpPAT10_0643	plcppse060	5x	PSE C
Cp1002_0648	plcppse061	5x	PSE E
CpC231_0647	plcppse061	5x	PSE E
cpfrc_00648	plcppse061	5x	PSE E
CpI19_0647	plcppse061	5x	PSE E
CpPAT10_0648	plcppse061	5x	PSE E
Cp1002_0661	plcppse062	5x	PSE C
CpC231_0660	plcppse062	5x	PSE C
cpfrc_00661	plcppse062	5x	PSE C
CpI19_0660	plcppse062	5x	PSE C
CpPAT10_0661	plcppse062	5x	PSE C
Cp1002_0706	plcppse063	5x	PSE E
CpC231_0705	plcppse063	5x	PSE C
cpfrc_00706	plcppse063	5x	PSE C
CpI19_0705	plcppse063	5x	PSE C
CpPAT10_0705	plcppse063	5x	PSE C
Cp1002_0715	plcppse064	5x	PSE C
CpC231_0714	plcppse064	5x	PSE C
cpfrc_00715	plcppse064	5x	PSE C
CpI19_0714	plcppse064	5x	PSE C
CpPAT10_0713	plcppse064	5x	PSE C
Cp1002_0734	plcppse065	5x	PSE E
CpC231_0733	plcppse065	5x	PSE E
cpfrc_00733	plcppse065	5x	PSE E
CpI19_0733	plcppse065	5x	PSE E
CpPAT10_0731	plcppse065	5x	PSE E
Cp1002_0744	plcppse066	5x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_0744	plcppse066	5x	PSE C
cpfrc_00744	plcppse066	5x	PSE C
CpI19_0744	plcppse066	5x	PSE C
CpPAT10_0742	plcppse066	5x	PSE C
Cp1002_0749	plcppse067	5x	PSE C
CpC231_0749	plcppse067	5x	PSE C
cpfrc_00748	plcppse067	5x	PSE C
CpI19_0749	plcppse067	5x	PSE C
CpPAT10_0747	plcppse067	5x	PSE C
Cp1002_0759	plcppse068	5x	PSE E
CpC231_0759	plcppse068	5x	PSE E
cpfrc_00759	plcppse068	5x	PSE E
CpI19_0759	plcppse068	5x	PSE E
CpPAT10_0758	plcppse068	5x	PSE E
Cp1002_0797	plcppse069	5x	PSE E
CpC231_0797	plcppse069	5x	PSE E
cpfrc_00797	plcppse069	5x	PSE E
CpI19_0797	plcppse069	5x	PSE E
CpPAT10_0795	plcppse069	5x	PSE E
Cp1002_0849	plcppse070	5x	PSE C
CpC231_0851	plcppse070	5x	PSE C
cpfrc_00851	plcppse070	5x	PSE C
CpI19_0851	plcppse070	5x	PSE C
CpPAT10_0849	plcppse070	5x	PSE C
Cp1002_0876	plcppse071	5x	PSE E
CpC231_0878	plcppse071	5x	PSE E
cpfrc_00878	plcppse071	5x	PSE E
CpI19_0878	plcppse071	5x	PSE E
CpPAT10_0876	plcppse071	5x	PSE E
Cp1002_0923	plcppse072	5x	PSE E
CpC231_0927	plcppse072	5x	PSE E
cpfrc_00928	plcppse072	5x	PSE E
CpI19_0928	plcppse072	5x	PSE E
CpPAT10_0924	plcppse072	5x	PSE E
Cp1002_0930	plcppse073	5x	PSE C
CpC231_0932	plcppse073	5x	PSE C
cpfrc_00934	plcppse073	5x	PSE C
CpI19_0935	plcppse073	5x	PSE C
CpPAT10_0931	plcppse073	5x	PSE C
Cp1002_0942	plcppse074	5x	PSE N

Locus tag	Pan locus	Set	Local subcellular
CpC231_0944	plcppse074	5x	PSE N
cpfrc_00947	plcppse074	5x	PSE N
CpI19_0947	plcppse074	5x	PSE N
CpPAT10_0943	plcppse074	5x	PSE N
Cp1002_0962	plcppse075	5x	PSE N
CpC231_0964	plcppse075	5x	PSE N
cpfrc_00968	plcppse075	5x	PSE N
CpI19_0967	plcppse075	5x	PSE N
CpPAT10_0962	plcppse075	5x	PSE N
Cp1002_0963	plcppse076	5x	PSE C
CpC231_0965	plcppse076	5x	PSE C
cpfrc_00969	plcppse076	5x	PSE C
CpI19_0968	plcppse076	5x	PSE C
CpPAT10_0963	plcppse076	5x	PSE C
Cp1002_0979	plcppse077	5x	PSE C
CpC231_0980	plcppse077	5x	PSE C
cpfrc_00985	plcppse077	5x	PSE C
CpI19_0984	plcppse077	5x	PSE C
CpPAT10_0979	plcppse077	5x	PSE C
Cp1002_0989	plcppse078	5x	PSE RN
CpC231_0990	plcppse078	5x	PSE RN
cpfrc_00997	plcppse078	5x	PSE RN
CpI19_0994	plcppse078	5x	PSE RN
CpPAT10_0989	plcppse078	5x	PSE RN
Cp1002_1002	plcppse079	5x	PSE C
CpC231_1001	plcppse079	5x	PSE C
cpfrc_01008	plcppse079	5x	PSE C
CpI19_1007	plcppse079	5x	PSE C
CpPAT10_1001	plcppse079	5x	PSE C
Cp1002_1009	plcppse080	5x	PSE N
CpC231_1008	plcppse080	5x	PSE N
cpfrc_01015	plcppse080	5x	PSE N
CpI19_1014	plcppse080	5x	PSE N
CpPAT10_1008	plcppse080	5x	PSE N
Cp1002_1017	plcppse081	5x	PSE L
CpC231_1016	plcppse081	5x	PSE L
cpfrc_01021	plcppse081	5x	PSE L
CpI19_1022	plcppse081	5x	PSE L
CpPAT10_1016	plcppse081	5x	PSE L
Cp1002_1052	plcppse082	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_1050	plcppse082	5x	PSE E
cpfrc_01057	plcppse082	5x	PSE E
CpI19_1057	plcppse082	5x	PSE E
CpPAT10_1051	plcppse082	5x	PSE E
Cp1002_1074	plcppse083	5x	PSE N
CpC231_1073	plcppse083	5x	PSE N
cpfrc_01080	plcppse083	5x	PSE N
CpI19_1080	plcppse083	5x	PSE N
CpPAT10_1073	plcppse083	5x	PSE N
Cp1002_1083	plcppse084	5x	PSE N
CpC231_1082	plcppse084	5x	PSE RN
cpfrc_01087	plcppse084	5x	PSE RN
CpI19_1089	plcppse084	5x	PSE RN
CpPAT10_1082	plcppse084	5x	PSE RN
Cp1002_1122	plcppse085	5x	PSE E
CpC231_1121	plcppse085	5x	PSE E
cpfrc_01126	plcppse085	5x	PSE E
CpI19_1128	plcppse085	5x	PSE E
CpPAT10_1121	plcppse085	5x	PSE E
Cp1002_1151	plcppse086	5x	PSE C
CpC231_1150	plcppse086	5x	PSE C
cpfrc_01154	plcppse086	5x	PSE C
CpI19_1157	plcppse086	5x	PSE C
CpPAT10_1149	plcppse086	5x	PSE C
Cp1002_1153	plcppse087	5x	PSE L
CpC231_1152	plcppse087	5x	PSE L
cpfrc_01156	plcppse087	5x	PSE L
CpI19_1159	plcppse087	5x	PSE L
CpPAT10_1151	plcppse087	5x	PSE L
Cp1002_1164	plcppse088	5x	PSE C
CpC231_1163	plcppse088	5x	PSE C
cpfrc_01168	plcppse088	5x	PSE C
CpI19_1170	plcppse088	5x	PSE C
CpPAT10_1162	plcppse088	5x	PSE C
Cp1002_1168	plcppse089	5x	PSE E
CpC231_1167	plcppse089	5x	PSE E
cpfrc_01172	plcppse089	5x	PSE E
CpI19_1174	plcppse089	5x	PSE E
CpPAT10_1166	plcppse089	5x	PSE E
Cp1002_1169	plcppse090	5x	PSE L

Locus tag	Pan locus	Set	Local subcellular
CpC231_1168	plcppse090	5x	PSE L
cpfrc_01173	plcppse090	5x	PSE L
CpI19_1175	plcppse090	5x	PSE L
CpPAT10_1167	plcppse090	5x	PSE L
Cp1002_1170	plcppse091	5x	PSE L
CpC231_1169	plcppse091	5x	PSE L
cpfrc_01174	plcppse091	5x	PSE L
CpI19_1176	plcppse091	5x	PSE L
CpPAT10_1168	plcppse091	5x	PSE L
Cp1002_1173	plcppse092	5x	PSE E
CpC231_1172	plcppse092	5x	PSE E
cpfrc_01177	plcppse092	5x	PSE E
CpI19_1179	plcppse092	5x	PSE E
CpPAT10_1171	plcppse092	5x	PSE E
Cp1002_1188	plcppse093	5x	PSE E
CpC231_1187	plcppse093	5x	PSE E
cpfrc_01192	plcppse093	5x	PSE E
CpI19_1194	plcppse093	5x	PSE E
CpPAT10_1186	plcppse093	5x	PSE E
Cp1002_1189	plcppse094	5x	PSE N
CpC231_1188	plcppse094	5x	PSE N
cpfrc_01193	plcppse094	5x	PSE N
CpI19_1195	plcppse094	5x	PSE N
CpPAT10_1187	plcppse094	5x	PSE N
Cp1002_1230	plcppse095	5x	PSE N
CpC231_1229	plcppse095	5x	PSE N
cpfrc_01238	plcppse095	5x	PSE N
CpI19_1236	plcppse095	5x	PSE N
CpPAT10_1229	plcppse095	5x	PSE N
Cp1002_1260	plcppse096	5x	PSE C
CpC231_1259	plcppse096	5x	PSE C
cpfrc_01265	plcppse096	5x	PSE C
CpI19_1266	plcppse096	5x	PSE C
CpPAT10_1258	plcppse096	5x	PSE C
Cp1002_1281	plcppse097	5x	PSE E
CpC231_1280	plcppse097	5x	PSE E
cpfrc_01285	plcppse097	5x	PSE E
CpI19_1287	plcppse097	5x	PSE E
CpPAT10_1279	plcppse097	5x	PSE E
Cp1002_1319	plcppse098	5x	PSE C
CpC231_1318	plcppse098	5x	PSE C
cpfrc_01324	plcppse098	5x	PSE C
CpI19_1324	plcppse098	5x	PSE C
CpPAT10_1318	plcppse098	5x	PSE C
Cp1002_1328	plcppse099	5x	PSE C
CpC231_1327	plcppse099	5x	PSE C
cpfrc_01333	plcppse099	5x	PSE C
CpI19_1333	plcppse099	5x	PSE C
CpPAT10_1327	plcppse099	5x	PSE C
Cp1002_1362	plcppse100	5x	PSE E
CpC231_1361	plcppse100	5x	PSE E
cpfrc_01368	plcppse100	5x	PSE E
CpI19_1367	plcppse100	5x	PSE E
CpPAT10_1361	plcppse100	5x	PSE E
Cp1002_1379	plcppse101	5x	PSE E
CpC231_1378	plcppse101	5x	PSE E
cpfrc_01385	plcppse101	5x	PSE E
CpI19_1384	plcppse101	5x	PSE E
CpPAT10_1378	plcppse101	5x	PSE E
Cp1002_1397	plcppse102	5x	PSE C
CpC231_1396	plcppse102	5x	PSE C
cpfrc_01403	plcppse102	5x	PSE C
CpI19_1402	plcppse102	5x	PSE C
CpPAT10_1396	plcppse102	5x	PSE C
Cp1002_1409	plcppse103	5x	PSE N
CpC231_1409	plcppse103	5x	PSE N
cpfrc_01414	plcppse103	5x	PSE N
CpI19_1416	plcppse103	5x	PSE N
CpPAT10_1408	plcppse103	5x	PSE N
Cp1002_1421	plcppse104	5x	PSE C
CpC231_1420	plcppse104	5x	PSE C
cpfrc_01424	plcppse104	5x	PSE C
CpI19_1427	plcppse104	5x	PSE C
CpPAT10_1418	plcppse104	5x	PSE C
Cp1002_1422	plcppse105	5x	PSE L
CpC231_1421	plcppse105	5x	PSE L
cpfrc_01425	plcppse105	5x	PSE L
CpI19_1428	plcppse105	5x	PSE L
CpPAT10_1419	plcppse105	5x	PSE L
Cp1002_1425	plcppse106	5x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_1424	plcppse106	5x	PSE C
cpfrc_01428	plcppse106	5x	PSE C
CpI19_1431	plcppse106	5x	PSE C
CpPAT10_1422	plcppse106	5x	PSE C
Cp1002_1466	plcppse107	5x	PSE N
CpC231_1468	plcppse107	5x	PSE N
cpfrc_01476	plcppse107	5x	PSE N
CpI19_1475	plcppse107	5x	PSE N
CpPAT10_1469	plcppse107	5x	PSE N
Cp1002_1467	plcppse108	5x	PSE N
CpC231_1469	plcppse108	5x	PSE N
cpfrc_01477	plcppse108	5x	PSE N
CpI19_1476	plcppse108	5x	PSE N
CpPAT10_1470	plcppse108	5x	PSE N
Cp1002_1492	plcppse109	5x	PSE L
CpC231_1494	plcppse109	5x	PSE L
cpfrc_01502	plcppse109	5x	PSE L
CpI19_1501	plcppse109	5x	PSE L
CpPAT10_1494	plcppse109	5x	PSE L
Cp1002_1493	plcppse110	5x	PSE C
CpC231_1495	plcppse110	5x	PSE C
cpfrc_01503	plcppse110	5x	PSE C
CpI19_1502	plcppse110	5x	PSE C
CpPAT10_1495	plcppse110	5x	PSE C
Cp1002_1503	plcppse111	5x	PSE E
CpC231_1506	plcppse111	5x	PSE E
cpfrc_01513	plcppse111	5x	PSE E
CpI19_1512	plcppse111	5x	PSE E
CpPAT10_1506	plcppse111	5x	PSE E
Cp1002_1510	plcppse112	5x	PSE RC
CpC231_1513	plcppse112	5x	PSE RC
cpfrc_01520	plcppse112	5x	PSE RC
CpI19_1519	plcppse112	5x	PSE RC
CpPAT10_1513	plcppse112	5x	PSE RC
Cp1002_1517	plcppse113	5x	PSE L
CpC231_1520	plcppse113	5x	PSE L
cpfrc_01526	plcppse113	5x	PSE L
CpI19_1526	plcppse113	5x	PSE L
CpPAT10_1520	plcppse113	5x	PSE L
Cp1002_1540	plcppse114	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_1543	plcppse114	5x	PSE E
cpfrc_01549	plcppse114	5x	PSE E
CpI19_1548	plcppse114	5x	PSE E
CpPAT10_1543	plcppse114	5x	PSE E
Cp1002_1549	plcppse115	5x	PSE E
CpC231_1551	plcppse115	5x	PSE E
cpfrc_01558	plcppse115	5x	PSE E
CpI19_1556	plcppse115	5x	PSE E
CpPAT10_1551	plcppse115	5x	PSE E
Cp1002_1573	plcppse116	5x	PSE C
CpC231_1575	plcppse116	5x	PSE C
cpfrc_01580	plcppse116	5x	PSE C
CpI19_1580	plcppse116	5x	PSE C
CpPAT10_1575	plcppse116	5x	PSE C
Cp1002_1604	plcppse117	5x	PSE C
CpC231_1606	plcppse117	5x	PSE C
cpfrc_01608	plcppse117	5x	PSE C
CpI19_1612	plcppse117	5x	PSE C
CpPAT10_1605	plcppse117	5x	PSE C
Cp1002_1610	plcppse118	5x	PSE C
CpC231_1611	plcppse118	5x	PSE C
cpfrc_01615	plcppse118	5x	PSE C
CpI19_1617	plcppse118	5x	PSE C
CpPAT10_1610	plcppse118	5x	PSE C
Cp1002_1647	plcppse119	5x	PSE C
CpC231_1648	plcppse119	5x	PSE C
cpfrc_01649	plcppse119	5x	PSE C
CpI19_1656	plcppse119	5x	PSE C
CpPAT10_1648	plcppse119	5x	PSE C
Cp1002_1705	plcppse121	5x	PSE C
CpC231_1697	plcppse121	5x	PSE C
cpfrc_01704	plcppse121	5x	PSE C
CpI19_1713	plcppse121	5x	PSE C
CpPAT10_1705	plcppse121	5x	PSE C
Cp1002_1706	plcppse122	5x	PSE C
CpC231_1698	plcppse122	5x	PSE C
cpfrc_01705	plcppse122	5x	PSE C
CpI19_1714	plcppse122	5x	PSE C
CpPAT10_1706	plcppse122	5x	PSE C
Cp1002_1714	plcppse123	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_1706	plcppse123	5x	PSE E
cpfrc_01713	plcppse123	5x	PSE E
CpI19_1722	plcppse123	5x	PSE E
CpPAT10_1714	plcppse123	5x	PSE E
Cp1002_1741	plcppse124	5x	PSE C
CpC231_1733	plcppse124	5x	PSE C
cpfrc_01740	plcppse124	5x	PSE C
CpI19_1749	plcppse124	5x	PSE C
CpPAT10_1741	plcppse124	5x	PSE C
Cp1002_1749	plcppse125	5x	PSE N
CpC231_1741	plcppse125	5x	PSE N
cpfrc_01749	plcppse125	5x	PSE N
CpI19_1757	plcppse125	5x	PSE N
CpPAT10_1750	plcppse125	5x	PSE N
Cp1002_1753	plcppse126	5x	PSE E
CpC231_1745	plcppse126	5x	PSE E
cpfrc_01753	plcppse126	5x	PSE E
CpI19_1761	plcppse126	5x	PSE E
CpPAT10_1754	plcppse126	5x	PSE E
Cp1002_1764	plcppse127	5x	PSE N
CpC231_1755	plcppse127	5x	PSE N
cpfrc_01763	plcppse127	5x	PSE N
CpI19_1772	plcppse127	5x	PSE N
CpPAT10_1765	plcppse127	5x	PSE N
Cp1002_1768	plcppse128	5x	PSE C
CpC231_1758	plcppse128	5x	PSE C
cpfrc_01766	plcppse128	5x	PSE C
CpI19_1776	plcppse128	5x	PSE C
CpPAT10_1768	plcppse128	5x	PSE C
Cp1002_1780	plcppse129	5x	PSE C
CpC231_1770	plcppse129	5x	PSE C
cpfrc_01778	plcppse129	5x	PSE C
CpI19_1788	plcppse129	5x	PSE C
CpPAT10_1780	plcppse129	5x	PSE C
Cp1002_1794	plcppse130	5x	PSE C
CpC231_1784	plcppse130	5x	PSE C
cpfrc_01792	plcppse130	5x	PSE C
CpI19_1802	plcppse130	5x	PSE C
CpPAT10_1794	plcppse130	5x	PSE C
Cp1002_1800	plcppse131	5x	PSE C
CpC231_1790	plcppse131	5x	PSE C
cpfrc_01798	plcppse131	5x	PSE C
CpI19_1808	plcppse131	5x	PSE C
CpPAT10_1800	plcppse131	5x	PSE C
Cp1002_1804	plcppse132	5x	PSE N
CpC231_1794	plcppse132	5x	PSE N
cpfrc_01801	plcppse132	5x	PSE N
CpI19_1812	plcppse132	5x	PSE N
CpPAT10_1804	plcppse132	5x	PSE N
Cp1002_1811	plcppse133	5x	PSE RN
CpC231_1802	plcppse133	5x	PSE RN
cpfrc_01808	plcppse133	5x	PSE RN
CpI19_1820	plcppse133	5x	PSE RN
CpPAT10_1812	plcppse133	5x	PSE RN
Cp1002_1825	plcppse134	5x	PSE C
CpC231_1817	plcppse134	5x	PSE C
cpfrc_01823	plcppse134	5x	PSE C
CpI19_1835	plcppse134	5x	PSE C
CpPAT10_1827	plcppse134	5x	PSE C
Cp1002_1845	plcppse135	5x	PSE L
CpC231_1838	plcppse135	5x	PSE L
cpfrc_01845	plcppse135	5x	PSE L
CpI19_1856	plcppse135	5x	PSE L
CpPAT10_1848	plcppse135	5x	PSE L
Cp1002_1846	plcppse136	5x	PSE C
CpC231_1839	plcppse136	5x	PSE C
cpfrc_01846	plcppse136	5x	PSE C
CpI19_1857	plcppse136	5x	PSE C
CpPAT10_1849	plcppse136	5x	PSE C
Cp1002_1869	plcppse137	5x	PSE RL
CpC231_1863	plcppse137	5x	PSE RL
cpfrc_01872	plcppse137	5x	PSE RL
CpI19_1880	plcppse137	5x	PSE RL
CpPAT10_1874	plcppse137	5x	PSE RL
Cp1002_1870	plcppse138	5x	PSE L
CpC231_1864	plcppse138	5x	PSE L
cpfrc_01873	plcppse138	5x	PSE L
CpI19_1881	plcppse138	5x	PSE L
CpPAT10_1875	plcppse138	5x	PSE L
Cp1002_1872	plcppse139	5x	PSE RN

Locus tag	Pan locus	Set	Local subcellular
CpC231_1865	plcppse139	5x	PSE RN
cpfrc_01874	plcppse139	5x	PSE RL
CpI19_1882	plcppse139	5x	PSE RN
CpPAT10_1876	plcppse139	5x	PSE RN
Cp1002_1874	plcppse140	5x	PSE L
CpC231_1867	plcppse140	5x	PSE L
cpfrc_01875	plcppse140	5x	PSE L
CpI19_1884	plcppse140	5x	PSE L
CpPAT10_1877	plcppse140	5x	PSE L
Cp1002_1878	plcppse141	5x	PSE N
CpC231_1871	plcppse141	5x	PSE N
cpfrc_01879	plcppse141	5x	PSE N
CpI19_1888	plcppse141	5x	PSE N
CpPAT10_1881	plcppse141	5x	PSE N
Cp1002_1885	plcppse142	5x	PSE N
CpC231_1877	plcppse142	5x	PSE N
cpfrc_01887	plcppse142	5x	PSE N
CpI19_1897	plcppse142	5x	PSE N
CpPAT10_1888	plcppse142	5x	PSE N
Cp1002_1887	plcppse143	5x	PSE C
CpC231_1879	plcppse143	5x	PSE C
cpfrc_01889	plcppse143	5x	PSE C
CpI19_1899	plcppse143	5x	PSE C
CpPAT10_1890	plcppse143	5x	PSE C
Cp1002_1901	plcppse144	5x	PSE N
CpC231_1893	plcppse144	5x	PSE N
cpfrc_01905	plcppse144	5x	PSE N
CpI19_1914	plcppse144	5x	PSE N
CpPAT10_1906	plcppse144	5x	PSE N
Cp1002_1909	plcppse145	5x	PSE C
CpC231_1903	plcppse145	5x	PSE C
cpfrc_01915	plcppse145	5x	PSE C
CpI19_1924	plcppse145	5x	PSE C
CpPAT10_1916	plcppse145	5x	PSE C
Cp1002_1914	plcppse146	5x	PSE E
CpC231_1908	plcppse146	5x	PSE E
cpfrc_01920	plcppse146	5x	PSE E
CpI19_1929	plcppse146	5x	PSE E
CpPAT10_1921	plcppse146	5x	PSE E
Cp1002_1925	plcppse147	5x	PSE L
CpC231_1919	plcppse147	5x	PSE L
cpfrc_01930	plcppse147	5x	PSE L
CpI19_1940	plcppse147	5x	PSE L
CpPAT10_1932	plcppse147	5x	PSE L
Cp1002_1926	plcppse148	5x	PSE C
CpC231_1920	plcppse148	5x	PSE C
cpfrc_01931	plcppse148	5x	PSE C
CpI19_1941	plcppse148	5x	PSE C
CpPAT10_1933	plcppse148	5x	PSE C
Cp1002_1933	plcppse149	5x	PSE C
CpC231_1927	plcppse149	5x	PSE C
cpfrc_01937	plcppse149	5x	PSE C
CpI19_1948	plcppse149	5x	PSE C
CpPAT10_1940	plcppse149	5x	PSE C
Cp1002_1936	plcppse150	5x	PSE L
CpC231_1930	plcppse150	5x	PSE L
cpfrc_01939	plcppse150	5x	PSE L
CpI19_1951	plcppse150	5x	PSE L
CpPAT10_1942	plcppse150	5x	PSE L
Cp1002_1938	plcppse151	5x	PSE L
CpC231_1932	plcppse151	5x	PSE L
cpfrc_01941	plcppse151	5x	PSE L
CpI19_1953	plcppse151	5x	PSE L
CpPAT10_1944	plcppse151	5x	PSE L
Cp1002_1939	plcppse152	5x	PSE L
CpC231_1933	plcppse152	5x	PSE L
cpfrc_01942	plcppse152	5x	PSE L
CpI19_1954	plcppse152	5x	PSE L
CpPAT10_1945	plcppse152	5x	PSE L
Cp1002_1945	plcppse153	5x	PSE L
CpC231_1939	plcppse153	5x	PSE L
cpfrc_01948	plcppse153	5x	PSE L
CpI19_1960	plcppse153	5x	PSE L
CpPAT10_1952	plcppse153	5x	PSE L
Cp1002_1954	plcppse154	5x	PSE E
CpC231_1948	plcppse154	5x	PSE E
cpfrc_01957	plcppse154	5x	PSE E
CpI19_1969	plcppse154	5x	PSE E
CpPAT10_1961	plcppse154	5x	PSE E
Cp1002_1958	plcppse155	5x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_1952	plcppse155	5x	PSE C
cpfrc_01961	plcppse155	5x	PSE C
CpI19_1973	plcppse155	5x	PSE C
CpPAT10_1965	plcppse155	5x	PSE C
Cp1002_1962	plcppse156	5x	PSE N
CpC231_1956	plcppse156	5x	PSE N
cpfrc_01965	plcppse156	5x	PSE N
CpI19_1977	plcppse156	5x	PSE N
CpPAT10_1969	plcppse156	5x	PSE N
Cp1002_1964	plcppse157	5x	PSE RN
CpC231_1958	plcppse157	5x	PSE RN
cpfrc_01967	plcppse157	5x	PSE RN
CpI19_1979	plcppse157	5x	PSE RN
CpPAT10_1971	plcppse157	5x	PSE RN
Cp1002_1965	plcppse158	5x	PSE E
CpC231_1959	plcppse158	5x	PSE E
cpfrc_01968	plcppse158	5x	PSE E
CpI19_1980	plcppse158	5x	PSE E
CpPAT10_1972	plcppse158	5x	PSE E
Cp1002_1970	plcppse159	5x	PSE E
CpC231_1964	plcppse159	5x	PSE E
cpfrc_01973	plcppse159	5x	PSE E
CpI19_1985	plcppse159	5x	PSE E
CpPAT10_1977	plcppse159	5x	PSE E
Cp1002_1982	plcppse160	5x	PSE C
CpC231_1976	plcppse160	5x	PSE C
cpfrc_01986	plcppse160	5x	PSE C
CpI19_1997	plcppse160	5x	PSE C
CpPAT10_1989	plcppse160	5x	PSE C
Cp1002_1983	plcppse161	5x	PSE N
CpC231_1977	plcppse161	5x	PSE N
cpfrc_01987	plcppse161	5x	PSE N
CpI19_1998	plcppse161	5x	PSE N
CpPAT10_1990	plcppse161	5x	PSE N
Cp1002_1984	plcppse162	5x	PSE N
CpC231_1978	plcppse162	5x	PSE N
cpfrc_01988	plcppse162	5x	PSE N
CpI19_1999	plcppse162	5x	PSE N
CpPAT10_1991	plcppse162	5x	PSE N
Cp1002_2005	plcppse163	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_1999	plcppse163	5x	PSE E
cpfrc_02012	plcppse163	5x	PSE E
CpI19_2021	plcppse163	5x	PSE E
CpPAT10_2014	plcppse163	5x	PSE E
Cp1002_2008	plcppse164	5x	PSE C
CpC231_2002	plcppse164	5x	PSE C
cpfrc_02015	plcppse164	5x	PSE C
CpI19_2024	plcppse164	5x	PSE C
CpPAT10_2017	plcppse164	5x	PSE C
Cp1002_2034	plcppse165	5x	PSE C
CpC231_2028	plcppse165	5x	PSE C
cpfrc_02038	plcppse165	5x	PSE C
CpI19_2050	plcppse165	5x	PSE C
CpPAT10_2041	plcppse165	5x	PSE C
Cp1002_2047	plcppse166	5x	PSE L
CpC231_2041	plcppse166	5x	PSE L
cpfrc_02050	plcppse166	5x	PSE L
CpI19_2062	plcppse166	5x	PSE L
CpPAT10_2054	plcppse166	5x	PSE L
Cp1002_2053	plcppse167	5x	PSE L
CpC231_2047	plcppse167	5x	PSE L
cpfrc_02054	plcppse167	5x	PSE L
CpI19_2068	plcppse167	5x	PSE L
CpPAT10_2057	plcppse167	5x	PSE L
Cp1002_2054	plcppse168	5x	PSE R
CpC231_2048	plcppse168	5x	PSE R
cpfrc_02055	plcppse168	5x	PSE R
CpI19_2069	plcppse168	5x	PSE R
CpPAT10_2058	plcppse168	5x	PSE R
Cp1002_2056	plcppse169	5x	PSE C
CpC231_2050	plcppse169	5x	PSE C
cpfrc_02057	plcppse169	5x	PSE C
CpI19_2071	plcppse169	5x	PSE C
CpPAT10_2060	plcppse169	5x	PSE C
Cp1002_2066	plcppse170	5x	PSE C
CpC231_2060	plcppse170	5x	PSE C
cpfrc_02067	plcppse170	5x	PSE C
CpI19_2081	plcppse170	5x	PSE C
CpPAT10_2070	plcppse170	5x	PSE C
Cp1002_2089	plcppse171	5x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_2082	plcppse171	5x	PSE E
cpfrc_02089	plcppse171	5x	PSE E
CpI19_2103	plcppse171	5x	PSE E
CpPAT10_2092	plcppse171	5x	PSE E
Cp1002_2097	plcppse172	5x	PSE N
CpC231_2090	plcppse172	5x	PSE N
cpfrc_02097	plcppse172	5x	PSE N
CpI19_2111	plcppse172	5x	PSE N
CpPAT10_2100	plcppse172	5x	PSE N
Cp1002_2098	plcppse173	5x	PSE C
CpC231_2091	plcppse173	5x	PSE C
cpfrc_02098	plcppse173	5x	PSE C
CpI19_2112	plcppse173	5x	PSE C
CpPAT10_2101	plcppse173	5x	PSE C
Cp1002_2102	plcppse174	5x	PSE R
CpC231_2095	plcppse174	5x	PSE R
cpfrc_02102	plcppse174	5x	PSE R
CpI19_2116	plcppse174	5x	PSE R
CpPAT10_2105	plcppse174	5x	PSE R
Cp1002_0050a	plcppse176	5x	PSE C
CpC231_0049	plcppse176	5x	PSE C
cpfrc_00052	plcppse176	5x	PSE C
CpI19_0051a	plcppse176	5x	PSE C
CpPAT10_0051	plcppse176	5x	PSE C
Cp1002_1829a	plcppse187	5x	PSE E
CpC231_1822	plcppse187	5x	PSE E
cpfrc_01828	plcppse187	5x	PSE E
CpI19_1840	plcppse187	5x	PSE E
CpPAT10_1832	plcppse187	5x	PSE E
Cp1002_1880	plcppse195	5x	PSE L
CpC231_1873	plcppse195	5x	PSE L
cpfrc_01881	plcppse195	5x	PSE N
CpI19_1890	plcppse195	5x	PSE L
CpPAT10_1883	plcppse195	5x	PSE L
Cp1002_2058	plcppse196	5x	PSE RN
CpC231_2052	plcppse196	5x	PSE RN
cpfrc_02059	plcppse196	5x	PSE RN
CpI19_2073	plcppse196	5x	PSE RN
CpPAT10_2062	plcppse196	5x	PSE RN
Cp1002_0300	plcppse198	5x	PSE E

Locus tag	Pan locus	Set	Local subcellular
CpC231_0304	plcppse198	5x	PSE E
cpfrc_00297	plcppse198	5x	PSE E
CpI19_0302	plcppse198	5x	PSE E
CpPAT10_0305	plcppse198	5x	PSE E
Cp1002_0438a	plcppse200	5x	PSE C
CpC231_0441a	plcppse200	5x	PSE C
cpfrc_00442	plcppse200	5x	PSE C
CpI19_0440	plcppse200	5x	PSE C
CpPAT10_0443a	plcppse200	5x	PSE C
Cp1002_0981	plcppse201	5x	PSE E
CpC231_0982	plcppse201	5x	PSE E
cpfrc_00987	plcppse201	5x	PSE E
CpI19_0986	plcppse201	5x	PSE E
CpPAT10_0981	plcppse201	5x	PSE E
Cp1002_1398	plcppse202	5x	PSE C
CpC231_1397	plcppse202	5x	PSE C
cpfrc_01404	plcppse202	5x	PSE C
CpI19_1403	plcppse202	5x	PSE C
CpPAT10_1397	plcppse202	5x	PSE C
Cp1002_0887	plcppse207	5x	PSE C
CpC231_0888	plcppse207	5x	PSE C
cpfrc_00888a	plcppse207	5x	PSE C
CpI19_0890	plcppse207	5x	PSE C
CpPAT10_0888	plcppse207	5x	PSE C
Cp1002_0902	plcppse221	5x	PSE C
CpC231_0904	plcppse221	5x	PSE C
cpfrc_00906	plcppse221	5x	PSE C
CpI19_0905	plcppse221	5x	PSE C
CpPAT10_0903	plcppse221	5x	PSE C
Cp1002_1637	plcppse222	5x	PSE C
CpC231_1638	plcppse222	5x	PSE E
cpfrc_01639	plcppse222	5x	PSE C
CpI19_1646	plcppse222	5x	PSE E
CpPAT10_1638	plcppse222	5x	PSE E
Cp1002_1687	plcppse120	4x	PSE RL
CpC231_1686	plcppse120	4x	PSE C
cpfrc_01685	plcppse120	4x	PSEUDOGENE
CpI19_1695	plcppse120	4x	PSE RL
CpPAT10_1687	plcppse120	4x	PSE RL
Cp1002_0019	plcppse175	4x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_0017	plcppse175	4x	PSE C
cpfrc_00017	plcppse175	4x	PSE C
CpI19_0019	plcppse175	4x	MEMBRANE
CpPAT10_0019	plcppse175	4x	PSE C
Cp1002_0052	plcppse177	4x	PSE E
CpC231_0051	plcppse177	4x	PSE E
cpfrc_00054	plcppse177	4x	PSE E
CpI19_0052	plcppse177	4x	MEMBRANE
CpPAT10_0053	plcppse177	4x	PSE E
Cp1002_0098	plcppse178	4x	PSE E
CpC231_0099	plcppse178	4x	PSE E
cpfrc_00100	plcppse178	4x	PSE E
CpI19_0100	plcppse178	4x	PSE E
CpPAT10_0098	plcppse178	4x	CYTOPLASMIC
Cp1002_0430	plcppse179	4x	PSE C
CpC231_0433	plcppse179	4x	PSE C
cpfrc_00433	plcppse179	4x	PSE C
CpI19_0431	plcppse179	4x	PSE C
CpPAT10_0435	plcppse179	4x	MEMBRANE
Cp1002_0454	plcppse180	4x	PSE N
CpC231_0458	plcppse180	4x	PSE N
cpfrc_00458	plcppse180	4x	PSE N
CpI19_0457	plcppse180	4x	PSE N
CpPAT10_0459	plcppse180	4x	PSEUDOGENE
Cp1002_0517	plcppse181	4x	PSE E
CpC231_0521	plcppse181	4x	CYTOPLASMIC
cpfrc_00520	plcppse181	4x	PSE E
CpI19_0520	plcppse181	4x	PSE E
CpPAT10_0520	plcppse181	4x	PSE E
Cp1002_0799	plcppse182	4x	PSE C
CpC231_0799	plcppse182	4x	PSE C
cpfrc_00799	plcppse182	4x	PSE C
CpI19_0799	plcppse182	4x	PSE C
CpPAT10_0797	plcppse182	4x	PSEUDOGENE
Cp1002_0810	plcppse183	4x	PSE C
CpC231_0812	plcppse183	4x	PSE C
cpfrc_00812	plcppse183	4x	PSE C
CpI19_0812	plcppse183	4x	PSE C
CpPAT10_0810	plcppse183	4x	MEMBRANE
Cp1002_1018	plcppse184	4x	PSE C
CpC231_1017	plcppse184	4x	PSE C
cpfrc_01022	plcppse184	4x	PSE C
CpI19_1023	plcppse184	4x	MEMBRANE
CpPAT10_1017	plcppse184	4x	PSE C
Cp1002_1310	plcppse185	4x	PSE C
CpC231_1309	plcppse185	4x	SECRETED
cpfrc_01315	plcppse185	4x	PSE C
CpI19_1315	plcppse185	4x	PSE C
CpPAT10_1309	plcppse185	4x	PSE C
ABSENT	plcppse186	4x	ABSENT
Cp1002_1693	plcppse186	4x	PSE L
cpfrc_01693	plcppse186	4x	PSE L
CpI19_1701	plcppse186	4x	PSE L
CpPAT10_1693	plcppse186	4x	PSE L
Cp1002_1851	plcppse188	4x	PSE C
CpC231_1844	plcppse188	4x	PSE C
cpfrc_01851	plcppse188	4x	PSE C
CpI19_1862	plcppse188	4x	PSEUDOGENE
CpPAT10_1854	plcppse188	4x	PSE C
Cp1002_1867	plcppse189	4x	PSE RN
CpC231_1861	plcppse189	4x	SECRETED
cpfrc_01870	plcppse189	4x	PSE RN
CpI19_1878	plcppse189	4x	PSE RN
CpPAT10_1872	plcppse189	4x	PSE RN
Cp1002_1910	plcppse190	4x	PSE C
CpC231_1904	plcppse190	4x	PSE C
cpfrc_01916	plcppse190	4x	PSE C
CpI19_1925	plcppse190	4x	MEMBRANE
CpPAT10_1917	plcppse190	4x	PSE C
Cp1002_1953	plcppse191	4x	PSE C
CpC231_1947	plcppse191	4x	PSE C
cpfrc_01956	plcppse191	4x	PSE C
CpI19_1968	plcppse191	4x	PSE C
CpPAT10_1960	plcppse191	4x	PSEUDOGENE
Cp1002_0316	plcppse204	4x	PSE E
CpC231_0320	plcppse204	4x	PSE E
cpfrc_00314	plcppse204	4x	CYTOPLASMIC
CpI19_0319	plcppse204	4x	PSE E
CpPAT10_0321	plcppse204	4x	PSE E
Cp1002_1684	plcppse211	4x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_1684	plcppse211	4x	PSE C
cpfrc_01681	plcppse211	4x	PSE C
CpI19_1693	plcppse211	4x	CYTOPLASMIC
CpPAT10_1684	plcppse211	4x	PSE C
Cp1002_1905	plcppse215	4x	PSE N
CpC231_1897	plcppse215	4x	PSE N
cpfrc_01910	plcppse215	4x	PSE N
CpI19_1918	plcppse215	4x	MEMBRANE
CpPAT10_1910	plcppse215	4x	PSE N
Cp1002_2040	plcppse217	4x	PSE C
CpC231_2034	plcppse217	4x	PSE C
cpfrc_02044	plcppse217	4x	CYTOPLASMIC
CpI19_2056	plcppse217	4x	PSE C
CpPAT10_2047	plcppse217	4x	PSE C
Cp1002_1684a	plcppse228	4x	PSE C
CpC231_1684a	plcppse228	4x	PSE C
cpfrc_01682	plcppse228	4x	PSE C
CpI19_1693a	plcppse228	4x	MEMBRANE
CpPAT10_1685	plcppse228	4x	PSE C
Cp1002_0662	plcppse192	3x	CYTOPLASMIC
CpC231_0661	plcppse192	3x	PSE C
cpfrc_00662	plcppse192	3x	PSE C
CpI19_0661	plcppse192	3x	PSE C
CpPAT10_0662	plcppse192	3x	CYTOPLASMIC
Cp1002_1763	plcppse193	3x	SECRETED
CpC231_1754	plcppse193	3x	SECRETED
cpfrc_01762	plcppse193	3x	PSE C
CpI19_1771	plcppse193	3x	PSE C
CpPAT10_1764	plcppse193	3x	PSE E
Cp1002_1797	plcppse194	3x	SECRETED
CpC231_1787	plcppse194	3x	SECRETED
cpfrc_01795	plcppse194	3x	PSE C
CpI19_1805	plcppse194	3x	PSE C
CpPAT10_1797	plcppse194	3x	PSE C
Cp1002_0219	plcppse197	3x	MEMBRANE
CpC231_0222	plcppse197	3x	PSE N
cpfrc_00219	plcppse197	3x	PSE N
CpI19_0221	plcppse197	3x	MEMBRANE
CpPAT10_0225	plcppse197	3x	PSE N
Cp1002_0065	plcppse218	3x	SECRETED

Locus tag	Pan locus	Set	Local subcellular
CpC231_0064	plcppse218	3x	PSE C
cpfrc_00067	plcppse218	3x	SECRETED
CpI19_0065	plcppse218	3x	PSE C
CpPAT10_0066	plcppse218	3x	PSE C
Cp1002_1838	plcppse223	3x	PSEUDOGENE
CpC231_1831	plcppse223	3x	PSE C
cpfrc_01837	plcppse223	3x	PSEUDOGENE
CpI19_1849	plcppse223	3x	PSE C
CpPAT10_1841	plcppse223	3x	PSE C
Cp1002_0437	plcppse199	2x	PSE E
CpC231_0440	plcppse199	2x	CYTOPLASMIC
cpfrc_00440	plcppse199	2x	PSE E
CpI19_0438	plcppse199	2x	CYTOPLASMIC
CpPAT10_0442	plcppse199	2x	CYTOPLASMIC
Cp1002_1900	plcppse203	2x	PSE RN
CpC231_1892	plcppse203	2x	PSEUDOGENE
cpfrc_01904	plcppse203	2x	PSE RN
CpI19_1913	plcppse203	2x	PSEUDOGENE
CpPAT10_1905	plcppse203	2x	PSEUDOGENE
Cp1002_0510	plcppse205	2x	PSE C
CpC231_0514	plcppse205	2x	SECRETED
cpfrc_00513	plcppse205	2x	SECRETED
CpI19_0513	plcppse205	2x	PSE C
CpPAT10_0513	plcppse205	2x	SECRETED
Cp1002_1859	plcppse212	2x	PSE RN
CpC231_1852	plcppse212	2x	PSE RN
cpfrc_01860	plcppse212	2x	PSEUDOGENE
CpI19_1870	plcppse212	2x	PSEUDOGENE
CpPAT10_1863	plcppse212	2x	PSEUDOGENE
Cp1002_1904	plcppse214	2x	PSE C
CpC231_1896	plcppse214	2x	PSE C
cpfrc_01908	plcppse214	2x	PSEUDOGENE
CpI19_1917	plcppse214	2x	PSEUDOGENE
CpPAT10_1909	plcppse214	2x	PSEUDOGENE
Cp1002_0559	plcppse219	2x	MEMBRANE
CpC231_0562	plcppse219	2x	PSE N
cpfrc_00560	plcppse219	2x	MEMBRANE
CpI19_0561	plcppse219	2x	MEMBRANE
CpPAT10_0561	plcppse219	2x	PSE N
Cp1002_0624	plcppse220	2x	MEMBRANE

Locus tag	Pan locus	Set	Local subcellular
CpC231_0624	plcppse220	2x	PSE N
cpfrc_00626	plcppse220	2x	MEMBRANE
CpI19_0623	plcppse220	2x	MEMBRANE
CpPAT10_0625	plcppse220	2x	PSE N
Cp1002_1204	plcppse206	1x	CYTOPLASMIC
CpC231_1203	plcppse206	1x	CYTOPLASMIC
cpfrc_01210	plcppse206	1x	PSE C
CpI19_1210	plcppse206	1x	CYTOPLASMIC
CpPAT10_1203	plcppse206	1x	CYTOPLASMIC
Cp1002_1562	plcppse208	1x	MEMBRANE
CpC231_1564	plcppse208	1x	MEMBRANE
cpfrc_01569	plcppse208	1x	PSE C
CpI19_1569	plcppse208	1x	MEMBRANE
CpPAT10_1564	plcppse208	1x	MEMBRANE
Cp1002_2065	plcppse209	1x	MEMBRANE
CpC231_2059	plcppse209	1x	MEMBRANE
cpfrc_02066	plcppse209	1x	PSE N
CpI19_2080	plcppse209	1x	MEMBRANE
CpPAT10_2069	plcppse209	1x	MEMBRANE
Cp1002_1058	plcppse210	1x	PSE N
CpC231_1056	plcppse210	1x	CYTOPLASMIC
cpfrc_01063	plcppse210	1x	CYTOPLASMIC
CpI19_1063	plcppse210	1x	CYTOPLASMIC
CpPAT10_1057	plcppse210	1x	CYTOPLASMIC
Cp1002_1883	plcppse213	1x	PSE C

Locus tag	Pan locus	Set	Local subcellular
CpC231_1875	plcppse213	1x	MEMBRANE
cpfrc_01883	plcppse213	1x	MEMBRANE
CpI19_1893	plcppse213	1x	MEMBRANE
CpPAT10_1885	plcppse213	1x	MEMBRANE
Cp1002_0096	plcppse224	1x	SECRETED
CpC231_0097	plcppse224	1x	SECRETED
cpfrc_00098	plcppse224	1x	MEMBRANE
CpI19_0098	plcppse224	1x	PSE C
CpPAT10_0096	plcppse224	1x	SECRETED
Cp1002_0369	plcppse226	1x	PSEUDOGENE
CpC231_0372	plcppse226	1x	SECRETED
cpfrc_00367	plcppse226	1x	CYTOPLASMIC
CpI19_0371	plcppse226	1x	SECRETED
CpPAT10_0373	plcppse226	1x	PSE C
Cp1002_0813	plcppse227	1x	SECRETED
CpC231_0815	plcppse227	1x	SECRETED
cpfrc_00815	plcppse227	1x	SECRETED
CpI19_0815	plcppse227	1x	SECRETED
CpPAT10_0813	plcppse227	1x	PSE C
CpPAT10_0670	plcppse228	1x	PSE N
cpfrc_00669	plcppse228	1x	CYTOPLASMIC
CpI19_0669	plcppse228	1x	CYTOPLASMIC
Cp1002_0670	plcppse228	1x	CYTOPLASMIC
CpC231_0669	plcppse228	1x	CYTOPLASMIC

6.5 *Curriculum vitae*

Dados Pessoais

Nome Anderson Rodrigues dos Santos
Nascimento 18/11/1971 - Belo Horizonte/MG - Brasil

Formação Acadêmica/Titulação

2008 Doutorado em Bioinformática.

Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil

Título: A GENÔMICA COMO FERRAMENTA PARA SELEÇÃO DE ALVOS CONTRA A LINFADENITE CASEOSA

Orientador: Dr. Vasco Ariston de Carvalho Azevedo

Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico

2007 - 2007 Aperfeiçoamento em Bioinformática.

Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil

Título: .

1997 - 1999 Mestrado em Ciências da Computação.

Universidade Federal de Minas Gerais, UFMG, Belo Horizonte, Brasil

Título: Construção de uma base de conhecimento para projetos de redes telefônicas utilizando KADS, Ano de obtenção: 1999

Orientador: José Lopes de Siqueira Neto

1991 - 1995 Graduação em Ciência da Computação.

Pontifícia Universidade Católica de Minas Gerais, PUC Minas, Belo Horizonte, Brasil

Formação complementar

2010 - 2010 Curso de curta duração em Clonagem e expressão de antígenos recombinantes.

Universidade Federal de Pelotas, UFPEL, Pelotas, Brasil

Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico

2010 - 2010 Curso de curta duração em Plataforma de Sequenciamento de Nova Geração.

Universidade Federal do Pará, UFPA, Belém, Brasil

Atuação profissional

1. Oi/TELEMAR NORTE LESTE SA - OI

Vínculo institucional

2000 - 2007 Vínculo: Colaborador , Enquadramento funcional: ANALISTA DE SISTEMAS , Carga horária: 40, Regime: Dedicação Exclusiva

2. Faculdade de Informática do Oeste de Minas Gerais - FIOM

Vínculo institucional

2000 - 2000 Vínculo: Celetista formal , Enquadramento funcional: Professor temporário , Carga horária: 16, Regime: Parcial

Atividades

02/2000 - 06/2000 Graduação, Informática

Disciplinas Ministradas:

Programação em lógica com Prolog

3. Faculdade Metropolitana de Belo Horizonte - FAME

Vínculo institucional

2000 - 2000 Vínculo: Professor convidado , Enquadramento funcional: Professor , Carga horária: 16, Regime: Parcial

4. Pontifícia Universidade Católica de Minas Gerais - PUC Minas

Vínculo institucional

2007 - 2007 Vínculo: Celetista formal , Enquadramento funcional: Professor temporário , Carga horária: 6, Regime: Parcial

2000 - 2000 Vínculo: Celetista formal , Enquadramento funcional: Professor temporário , Carga horária: 16, Regime: Parcial

Atividades

08/2007 - 12/2007 Graduação, Ciência da Informação

Disciplinas Ministradas:

17091 – INTERFACE HOMEM-MÁQUINA

08/2007 - 12/2007 Graduação, Jogos Digitais

Disciplinas Ministradas:

20637 - PROGRAMAÇÃO COM BIBLIOTECA GRÁFICA II (DirectX 9 e MS Visual Studio C++)

02/2000 - 11/2000 Graduação, Engenharia Elétrica

Disciplinas Ministradas:

Algoritmos e programação em linguagem C

02/2000 - 02/2000 Graduação, Ciências Contábeis

Disciplinas Ministradas:

Gerenciamento de dados por meio de Sistemas Gerenciadores de Bancos de Dados, Linguagem SQL e Recursos avançados de planilhas eletrônicas

5. Universidade Federal de Minas Gerais - UFMG

Vínculo institucional

1995 - 2000 Vínculo: Celetista formal , Enquadramento funcional: Analista de Sistemas , Carga horária: 40, Regime: Integral

Atividades

08/1995 - 05/2000 Serviço Técnico Especializado, Instituto de Ciências Exatas, Departamento de Ciência da Computação

Especificação:

Desenvolvedor de Software

Áreas de atuação

1. Ciência da Computação

2. Bioinformática

3. Inteligência Artificial

4. Banco de Dados

5. Linguagens de Programação

6. Genética Molecular e de Microrganismos

Idiomas

Inglês	Compreende Bem , Fala Razoavelmente, Escreve Razoavelmente, Lê Bem
Português	Compreende Bem , Fala Bem, Escreve Bem, Lê Bem

Produção em C, T& A

Produção bibliográfica

Artigos completos publicados em periódicos

1. Pacheco, Luis GC, Slade, Susan E, Seyffert, Núbia, SANTOS, A. R., Castro, TLP, Silva, Wanderson M, Santos, Agenor V, Santos, Simone G, Farias, Luiz M, Carvalho, Maria AR, Pimenta,

Adriano MC, Meyer, Roberto, Silva, Artur, Scrivens, James H, Oliveira, Sergio C, Miyoshi, Anderson, Dowson, Christopher G, Azevedo, Vasco. A combined approach for comparative exoproteome analysis of *Corynebacterium pseudotuberculosis*. BMC Microbiology (Online). , v.11, p.12 - , 2011.

2. Barh, Debmalya, Jain, Neha, Tiwari, Sandeep, D'AFONSECA, V., Li, Liwei, Ali, A., Santos, Anderson Rodrigues, Guimarães, Luís Carlos, SOARES, S. C., Miyoshi, Anderson, Bhattacharjee, Atanu, Misra, Amarendra Narayan, Silva, Artur, Kumar, Anil, Azevedo, Vasco. A novel comparative genomics analysis for common drug and vaccine targets in *Corynebacterium pseudotuberculosis* and other CMN group of human pathogens. Chemical Biology & Drug Design (Print). , p.no - no, 2011.

3. Santos, Anderson R, Santos, Marcos A, Baumbach, Jan, McCulloch, John A, Oliveira, Guilherme C, Silva, Artur, Miyoshi, Anderson, Azevedo, Vasco. A singular value decomposition approach for improved taxonomic classification of biological sequences. BMC Genomics. , v.12, p.S11 - , 2011.

4. CERDEIRA, L. T., Schneider, M. P. C., PINTO, A. C., de Almeida, S. S., dos Santos, A. R., Barbosa, E. G. V., Ali, A., Aburjaile, F. F., de Abreu, V. A. C., Guimaraes, L. C., Soares, S. d. C., Dorella, F. A., Rocha, F. S., BOL, E., Gomes de Sa, P. H. C., LOPES, T. S., Barbosa, M. S., Carneiro, A. R., Juca Ramos, R. T., Coimbra, N. A. d. R., LIMA, A. R. J., Barh, D., Jain, N., Tiwari, S., RAJA, R., ZAMBARE, V., Ghosh, P., Trost, E., Tauch, A., MIYOSHI, A., AZEVEDO, V., SILVA, A. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* Strain CIP 52.97, Isolated from a Horse in Kenya. Journal of Bacteriology (Print). , v.193, p.7025 - 7026, 2011.

5. STYNEN, A. P. R., LAGE, A. P., Moore, R. J., REZENDE, A. M., de Resende, V. D. d. S., Ruy, P. d. C., Daher, N., Resende, D. d. M., de Almeida, S. S., Soares, S. d. C., de Abreu, V. A. C., Rocha, A. A. C. M., dos Santos, A. R., Barbosa, E. G. V., Costa, D. F., Dorella, F. A., MIYOSHI, A., de Lima, A. R. J., Campos, F. D. d. S., de Sa, P. G., LOPES, T. S., Rodrigues, R. M. A., Carneiro, A. R., LEAO, T., CERDEIRA, L. T., RAMOS, R. T. J., SILVA, A., AZEVEDO, V., Ruiz, J. C. Complete Genome Sequence of Type Strain Campylobacter fetus subsp. venerealis NCTC 10354T. Journal of Bacteriology (Print). , v.193, p.5871 - 5872, 2011.

6. Resende, B.C., Rebelato, A.B., D'AFONSECA, V., Santos, A.R., Stutzman, T., AZEVEDO, V., MIYOSHI, A., Lopes, Débora O. DNA repair in *Corynebacterium* model. Gene (Amsterdam). , p.21497183 - , 2011.

7. Ruiz, Jerônimo C., D'AFONSECA, V., Silva, Artur, Ali, A., Pinto, Anne C., Santos, Anderson R., Rocha, Aryanne A. M. C., Lopes, Débora O., Dorella, F. A., Pacheco, Luis G. C., Costa, Marcília P., Turk, Meritxell Z., Seyffert, Núbia, Moraes, Pablo M. R. O., SOARES, S. C., ALMEIDA, S. S., Castro, TLP, ABREU, V. A. C., Trost, Eva, Baumbach, Jan, Tauch, Andreas, Schneider, M. P. C., McCulloch, John, CERDEIRA, L. T., RAMOS, R. T. J., Zerlotini, Adhemar, Dominitini, Anderson, Resende, Daniela M., Coser, Elisângela M., Oliveira, Luciana M., Pedrosa, André L., Vieira, Carlos U., Guimarães, Cláudia T., Bartholomeu, Daniela C., Oliveira, Diana M., Santos, Fabrício R., Rabelo, Élida Mara, Lobo, Francisco P., Franco, Glória R., Costa, Ana Flávia. Evidence for Reductive Genome

Evolution and Lateral Acquisition of Virulence Functions in Two *Corynebacterium pseudotuberculosis* Strains. Plos One. , v.6, p.e18551 - , 2011.

8. Seyffert, Núbia, Pacheco, Luis G. C., Silva, Wanderson M, Castro, TLP, Santos, Agenor V, SANTOS, A. R., McCulloch, John, Rodrigues, M. R., Santos, Simone G, Farias, Luiz M, Carvalho, Maria AR, Pimenta, Adriano MC, SILVA, A., Meyer, Roberto, Miyoshi, Anderson, AZEVEDO, V. Serological secretome analysis of *Corynebacterium pseudotuberculosis*. Journal of Integrated Omics. , v.1, p.54 - , 2011.

9. SANTOS, A. R., Ali, A., BARBOSA, E., SILVA, A., MIYOSHI, A., AZEVEDO, V. The reverse vaccinology - A contextual overview. The IIOAB Journal. , v.2, p.8 - 15, 2011.

10. CERDEIRA, L. T., PINTO, A. C., Schneider, M. P. C., de Almeida, S. S., dos Santos, A. R., Barbosa, E. G. V., Ali, A., Barbosa, M. S., Carneiro, A. R., RAMOS, R. T. J., de Oliveira, R. S., Barh, D., BARVE, N., ZAMBARE, V., Belchior, S. E., Guimaraes, L. C., de Castro Soares, S., Dorella, F. A., Rocha, F. S., de Abreu, V. A. C., Tauch, A., Trost, E., MIYOSHI, A., AZEVEDO, V., SILVA, A. Whole-Genome Sequence of *Corynebacterium pseudotuberculosis* PAT10 Strain Isolated from Sheep in Patagonia, Argentina. Journal of Bacteriology (Print). , v.193, p.6420 - 6421, 2011.

11. SILVA, A., Schneider, M. P. C., CERDEIRA, L. T., Barbosa, M. S., RAMOS, R. T. J., Carneiro, A. R., Santos, R., Lima, M., D'AFONSECA, V., ALMEIDA, S. S., SANTOS, A. R., SOARES, S. C., Pinto, A. C., Ali, A., Dorella, F. A., Rocha, F., de Abreu, V. A. C., Trost, Eva, Tauch, Andreas, Shpigel, N., MIYOSHI, A., AZEVEDO, V. Complete Genome Sequence of *Corynebacterium pseudotuberculosis* I19, a Strain Isolated from a Cow in Israel with Bovine Mastitis. Journal of Bacteriology (Print). , v.193, p.323 - 324, 2010.

12. Barh, Debmalya, Tiwari, Sandeep, Jain, Neha, Ali, A., Santos, Anderson Rodrigues, Misra, Amarendra Narayan, Azevedo, Vasco, Kumar, Anil. *in silico* subtractive genomics for target identification in human bacterial pathogens. Drug Development Research (Print). , p.n/a - n/a, 2010.

Capítulos de livros publicados

1. AZEVEDO, V., SILVA, A., MIYOSHI, A., BOREM, A., ABREU, V. A. C., ALMEIDA, S. S., AMÁDIO, A, BARBOSA, M S R, Carneiro, A. R., CERDEIRA, L. T., D'AFONSECA, V., GUEDES, R. L. M., MELO, H. V. F., OLIVEIRA, R. S., ORTEGA, J. M., RAMOS, R. T. J., SANTOS, A. R., SCHNEIDER, M. P. Anotação Funcional de Genomas Realizada Computacionalmente In: Manual Prático – Teórico: Sequenciamento, Montagem e Anotação de Genomas Bacterianos ed.Viçosa : Universidade Federal de, 2011, p. 91-109.

2. Carneiro, A. R., Ali, A., SANTOS, A. R., PINTO, A. C., Rocha, Aryanne A. M. C., BARBOSA, E., CERDEIRA, L. T., RAMOS, R. T. J., ALMEIDA, S. S., SOARES, S. C., ABREU, V. A. C.,

SCHNEIDER, M. P., SILVA, A., MIYOSHI, A., AZEVEDO, V. Whole genome annotation: *in silico* analysis In: Bioinformatics: Trends and Methodologies ed. : Intech, 2011, p. 679-703.

Artigos em revistas (Magazine)

1. SOARES, S. C., SILVA, A., RAMOS, R. T. J., CERDEIRA, L. T., Ali, A., SANTOS, A. R., PINTO, A. C., CASSIANO, A.A.M., FIGUEIRA, F., Carneiro, A. R., Guimarães, Luís Carlos, BARBOSA, E., ALMEIDA, S. S., ABREU, V. A. C., MIYOSHI, A., AZEVEDO, V. Plasticidade Genômica e Evolução Bacteriana. Informativo SBM (0102-8189). 26 CBM - Foz do Iguaçu, p.8 - 31, 2011.

2. RUIZ, J., SANTOS, A. R., PINTO, A. C., RESENDE, D. M., CERDEIRA, L. T., RAMOS, R. T. J., ORELLANA, S. C., ALMEIDA, S. S., SOARES, S. C., D'AFONSECA, V., AZEVEDO, V., SILVA, A. Segunda Revolução Genômica: Utilização de Sequenciadores de Nova Geração. Informativo SBM (0102-8189). Informática, p.15/11 - 18, 2009.

Apresentação de Trabalho

1. ABREU, V. A. C., SANTOS, A. R., ALMEIDA, S. S., FIGUEIRA, F., BARBOSA, E., FIAUX, K., SILVA, A., MIYOSHI, A., AZEVEDO, V. **CpDB: A relational database schema and tools for bacterial genomes annotation and posgenome research**, 2011. (Congresso,Apresentação de Trabalho)

2. BARBOSA, E., SANTOS, A. R., Guimarães, Luís Carlos, ABREU, V. A. C., PINTO, A. C., FIAUX, K., ALMEIDA, S. S., AZEVEDO, V. **PSEUDOGENE ANALYSIS IN FIVE STRAINS OF *Corynebacterium pseudotuberculosis***, 2011. (Outra,Apresentação de Trabalho)

3. SANTOS, A. R., Carneiro, A. R., GALA-GARCIA, A., PINTO, A. C., Barh, Debmalya, BARBOSA, E., Dorella, F. A., AZEVEDO, V. **The Corynecabcterium *pseudotuberculosis* pan genomics reverse vaccinology**, 2011. (Congresso,Apresentação de Trabalho)

4. SANTOS, A. R., SANTOS, M. A., MCCULLOCH, J. A., BAUMBACH, J., OLIVEIRA, G. C., SILVA, A., MIYOSHI, A., AZEVEDO, V. **A singular value decomposition approach for improved taxonomy classification of biological sequences**, 2010. (Comunicação,Apresentação de Trabalho)

5. SANTOS, A. R. **Bancos de dados relacionais para Montagem e Anotação de Genomas**, 2010. (Conferência ou palestra,Apresentação de Trabalho)

6. PINTO, A. C., SANTOS, A. R., ALMEIDA, S. S., SOARES, S. C., FARIA, C.J., MAGALHÃES, A., RUIZ, J., MIYOSHI, A., AZEVEDO, V. **Estudo da arquitetura genômica de duas linhagens do patogêno *Corynebacterium pseudotuberculosis* e seu estilo de vida**, 2010. (Outra,Apresentação de Trabalho)

7. ALMEIDA, S. S., Seyffert, Núbia, PRUDÊNCIO, C.R., SANTOS, F.A.A., SOARES, S. C., D'AFONSECA, V., PINTO, A. C., SANTOS, A. R., CASSIANO, A.A.M., FARIA, C.L., MIYOSHI, A.,

MOORE, R., GOULART, L.R., AZEVEDO, V. **Identificação de desordem proteica em proteoma de *Corynebacterium pseudotuberculosis* usando dados de Phage Display**, 2010. (Outra,Apresentação de Trabalho)

8. SOARES, S. C., ABREU, V. A. C., MCCULLOCH, J. A., D'AFONSECA, V., RAMOS, R. T. J., Ali, A., SANTOS, A. R., PINTO, A. C., ALMEIDA, S. S., SILVA, A., MIYOSHI, A., AZEVEDO, V. **PIPS: pathogenicity island prediction software**, 2010. (Outra,Apresentação de Trabalho)

9. SANTOS, A. R., SANTOS, M. A., AZEVEDO, V., OLIVEIRA, G. C. **Predição de epitopos lineares por meio da álgebra linear**, 2009. (Conferência ou palestra,Apresentação de Trabalho)

10. SANTOS, A. R., SANTOS, M. A., RUIZ, J., MCCULLOCH, J. A., OLIVEIRA, G. C., MIYOSHI, A., AZEVEDO, V. **Uma abordagem para produzir matrizes de distâncias por meio da decomposição de valores singulares**, 2009. (Conferência ou palestra,Apresentação de Trabalho)

Orientações e Supervisões

Orientações e Supervisões concluídas

Trabalhos de conclusão de curso de graduação

1. Eudes Guilherme Vieira Barbosa. **ANÁLISE DE PSEUDOGENES EM DIVERSAS LINHAGENS DE *Corynebacterium pseudotuberculosis***. 2011. Curso (Ciências Biológicas) - Universidade Federal de Minas Gerais

Iniciação científica

1. Eudes Guilherme Vieira Barbosa. **Anotação automática de genomas bacterianos**. 2011. Iniciação científica (Ciências Biológicas) - Universidade Federal de Minas Gerais

2. Flávia Figueira Aburjaile. **Curadoria de montagem e anotação do pan genoma de *Corynebacterium pseudotuberculosis***. 2011. Iniciação científica (BIOMEDICINA) - Universidade FUMEC

Citações em bases bibliográficas

Web of Science Número total de citações:2; Número de trabalhos:7; Data : 29/10/2011; Fator H:1;

Nome(s) do autor utilizado(s) na consulta para obter o total de citações:

Year Published=(2011) AND Author=(Santos A*), Affiliation: Universidade Federal de Minas Gerais

SCOPUS Número total de citações: 4; Número de trabalhos: 5; Data : 29/10/2011

Nome(s) do autor utilizado(s) na consulta para obter o total de citações:

Last name: Santos, First name: Anderson, Affiliation: Universidade Federal de Minas Gerais