UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

EDGAR LACERDA DE AGUIAR

Sequenciamento, montagem e anotação do genoma de *Streptococcus Agalactiae* GBS85147: uma abordagem comparativa

Belo Horizonte
2015

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
PROGRAMA INTERUNIDADES DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

EDGAR LACERDA DE AGUIAR

Sequenciamento, montagem e anotação do genoma de *Streptococcus Agalactiae* GBS85147: uma abordagem comparativa

Dissertação apresentada ao Programa de Pós-Graduação em Bioinformática da UFMG como requisito para a obtenção do grau de Mestre em Bioinformática.

Orientador: Prof. Dr. Vasco Ariston Carvalho de Azevedo
Coorientadora: Dra. Anne Cybele Pinto

Belo Horizonte,
2015

“Não importa o que mundo diz de mim,
o que importa é que eu nunca fiz nada
que contrariasse os meus princípios
e nunca farei”

“A lâmina mais forte é aquela que pode
proteger o que você quer proteger e
cortar aquilo que você quer cortar”

“Não pense naquilo que você perdeu
pense naquilo que você ainda tem”

Eiichiro Oda

Dedicatória

Dedico essa dissertação aos meus pais Alexandre e Margarida por acreditarem nos meus sonhos, respeitarem minhas escolhas pelo caminho que tomei, e claro, por todo amor e dedicação e a Deus por tudo.

Agradecimentos

Agradeço:

A Deus pela graça, saúde e misericórdia comigo.

Aos meus familiares pelo suporte, confiança, amor e toda educação depositada em mim. Agradecimento em especial aos meus pais Alexandre e Margarida, as minhas irmãs Anita e Ester, aos meus avós Carlos e Wanda e à minha bisavó Maria (*in memoriam).* Ao meu filho felino, Lord John, o pior gato do mundo e a Bernadette (*in memoriam)* por mudar minha opinião sobre os felinos e despertar meu amor pela espécie.

Aos meus orientadores Prof. Dr. Vasco Azevedo e Dra. Anne Cybele Pinto pela paciência, confiança e auxílio.

Aos meus Professores de Graduação que muito me ensinaram: Sandro, Ernani, Hélio, Linderberg, Juliana e Zilma.

Aos meus companheiros do Laboratório de Genética Celular Molecular da UFMG, em especial Karina Santana, Marcus Canário, Paulo Daltron, Flavia Rocha, Leandro Benevides, Lucas Amorim, Nat Tartaglia, Thiago Sousa, Sandeep Tiwari, Lucas Amorim, Flavia Figueira, Cassiana Jones, Carlos Diniz, Mariana Parise, Doglas Parise, Kátia Morais, pelo suporte, respeito e carinho.

Aos parceiros do Aquacen pelo auxílio e atenção, em especial ao Felipe Luiz, Fernanda Dorella e Prof. Dr. Henrique Figueiredo.

À colaboração do Laboratório de Biologia e Fisiologia de Estreptococos da UFRJ, em especial à Profa. Dra. Prescilla Emy Nagao e Camila Antunes.

As melhores secretarias da UFMG, Sheila Magalhães, Ana Paula, Marcia Natália e Fernanda Magalhães por serem tão prestativas e solicitas comigo.

À Profa. Dra. Raquel Minardi pela influência positiva, modificando minha visão sobre meus resultados e como demonstrá-los; aos amigos do LBS: Diego Mariano, Pedro e Alexandre Fassio.

Ao Prof. Dr. Miguel Ortega por suas ideias e visão única sobre o projeto, aos amigos do Biodados sempre solícitos: Carlos, Vêronica e Tetsu.

A Profa. Dra. Gloria Franco pela insistência em equilibrar os conhecimentos biológicos e os computacionais, as amigas do LGB: Mainá Bittar e Jéssica Hickson.

Aos amigos do Cebio, Francislon, Juliana, Laura e Izinara.

Aos Mahou's friends e agregados, Carlos, Denise, Emanuelle, Iasminne, Izack, Julia, Koshiro, Leticia, Lorena, Natalia, Raquel, Reginaldo, Renata, Renato.

Ao RTTEE melhor grupo de role play que já tive prazer de participar, aos amigos Diogo, Luciano, Renato, Rubens e Rogério.

As Nutri Girls da UFV, Alice e Millena em especial a Karina Ramos por ser uma amiga tão gentil e única.

As amigas essenciais: Andressa Coelho, Carla Batista, Estephany Hendi, Laura Pereira, Lívia Oliveira, Tarcila Araújo e Tatiana Pedra por todo suporte, paciência, respeito e carinho nos últimos anos.

Aos incríveis companheiros e funcionários que tornaram a nossa Pós-Graduação a melhor de Bioinformática do Brasil.

Muito obrigados a todos que me ajudaram nessa longa caminhada.

Sumário

Lista de abreviaturas 8
Colaborações 13
1. Revisão de Literatura 15
1.1 Microrganismos como alvos de diferentes estudos 15
1.2 *Streptococcus agalactiae* 18
1.2.1 Doenças causadas por *S. agalactiae* em humanos 20
1.2.2 Doenças causadas por *S. agalactiae* em bovideos 23
1.2.3 Doenças causadas por *S. agalactiae* em peixes 24
1.3 Fatores de Virulência 26
1.4 Estudos genômicos em *S. agalactiae* 29
1.4.1 Métodos de sequenciamento genômico 31
1.4.2 Plataforma Ion Torrent 32
1.4.3 Montagem e finalização de genomas 35
1.5 Análises comparativas com ferramentas de Bioinformática 36
1.5.1 Análises de sintenia 36
1.5.2 Análises Funcional Genômica 36
1.5.3 Análises de MLST 36
1.5.4 Filogenia molecular 37
2. Justificativa 39
3. Objetivos 40
3.1 - Objetivo Geral 40
3.2 - Objetivos Específicos 40
4. Capítulos 41
4.1 Capítulo 1 - Artigo submetido à revista “Standards in Genomic Sciences” 41
4.2 Capítulo 2 - Análises comparativas 61
4.2.1 Materiais e métodos 61
4.2.2 Resultados e discussões 67
5. Conclusões finais 87
5.1. Perspectivas para o futuro 88
6. Referências bibliográficas 89
Anexos 106

Lista de abreviaturas

ALP	*Alpha-like proteins*
AP	*Allelic Profile*
BAM	*Binary Alignment/Map*
BD	Banco de dados
BLAST	*Basic local alignment search tool*
CC	Complexos Clonais
CDS	*Coding DNA Sequence*
CLI	*Command-line interface*
CPS	*Capsular polysaccharides*
dATP	*Deoxyadenosine triphosphate*
dCTP	*Deoxycytidine triphosphate*
dGTP	*Deoxyguanosine triphosphate*
DNA	*Deoxyribonucleic acid*
dNTP	*Deoxyribonucleotide triphosphate*
dTTP	*Deoxythymidine triphosphate*
EOD	*Early-Onset Disease*
Gb	Gigabase
GB	Gigabyte
GBS	Group B *Streptococcus*
HD	*Hard disk*
INDEL	*Insertion or deletion*
ISFET	*Ion-sensitive field-effect transistor*
Kb	Kilobase
LOD	*Late-Onset Diseases*
Mb	Megabase
MB	Megabyte
NCBI	*National Center for Biotechnology Information*
NGS	*Next-Generation Sequencer*
NT	*Not Type*
OLC	*Overlap-Layout-Consensus*
Pb	Pares de base
PCR	*Polymerase Chain Reaction*
PGM	*Personal Genome Machine*
PGH	Projeto Genoma Humano
RN	Recém Nascido
RNA	*Ribonucleic acid*
rRNA	*Ribosomal Ribonucleic Acid*
SAM	*Sequence Alignment/Map*
SFF	*Standard flowgram format*
SNP	*Single nucleotide polymorphism*
SOLiD	*Sequencing by Oligonucleotide Ligation and Detection*
ST	*Sequence Type*
WGS	*Whole Genome Shotgun*

Lista de figuras

Figura 1 - Gráfico dos projetos separados por Domínios. 17
Figura 2 - Gráfico com Quantidade de Artigos por Ano relacionados a bactérias. 17
Figura 3 - Quantidade de Genomas de Bactérias por Filo. 18
Figura 4 - *Workflow* de passos resumidos do processo de sequenciamento do Ion PGM. 33
Figura 5 - *Template* da Reação DNA Polimerase 33
Figura 6 - Relatório dos sinais elétricos gerados pelo transistor ISFET. 34
Figura 7 - Processo Biológicos Nível 2 com Linhagem GBS85147. 68
Figura 8 - Processo Biológicos Nível 3 com Linhagem GBS85147. 68
Figura 9 – Análise comparativa entre genomas de *S. agalactiae* usando a ferramenta Mauve. 70
Figura 10 – Análise comparativa entre genomas de *S. agalactiae* GBS85147 (abaixo) com a linhagem *ILRI005* (acima) usando a ferramenta CONTIGuator. 71
Figura 11 – Análise comparativa entre genomas de *S. agalactiae* GBS85147 (acima) com a linhagem 138spar (abaixo) usando a ferramenta CONTIGuator. Para uma melhor visualização foi necessário o uso da fita reversa de GBS85147. 72
Figura 12 - Análise sintenia entre genomas de *S. agalactiae* usando a ferramenta BRIG. 73
Figura 13 – Visualização de ilhas genômicas obtidas pela ferramenta Gipsy. 75
Figura 14 – Visualização de ilhas patogênicas obtidas pela ferramenta Gipsy. 76
Figura 15 – Visualização de ilhas resistência obtidas pela ferramenta Gipsy. 77
Figura 16 – Alinhamento dos genes 16s das 13 linhagens visualizado pelo *software* Jalview. 79
Figura 17 – Árvore filogenética com genes 16s das 13 linhagens. 80
Figura 18 – Árvore filogenética com genes 16s das 17 linhagens, mais grupo externo. 82
Figura 19 - Árvore filogenética com gene *rpoB* das 17 linhagens, mais grupo externo. 83
Figura 20 – Alinhamento dos genes *16s* e *rpoB* das 20 linhagens visualizado pelo Mega6. 84
Figura 21 - Árvore filogenética com os gene *rpoB* e 16s das 17 Linhagens, mais grupo externo. 84

Lista de tabelas

Tabela 1 - Alguns dos principais fatores de virulência de *S. agalactiae* com hospedeiro humano, funcionamento e genes envolvidos. 27
Tabela 2 - Informações de genomas completos da *S. agalactiae* no NCBI. 30
Tabela 3 - Probabilidade de erro por base pelo algoritmo Phred. 35
Tabela 4 - Lista de Parâmetros utilizados no SIMBA. 62
Tabela 5 - Resultados dos Montadores com SIMBA. 67
Tabela 6 - Resultado MLST. 85

Resumo

Streptococcus agalactiae (*Lancefield group* B, GBS) é uma bactéria patogênica Gram-positiva, causadora de doenças em humanos, bovinos e peixes. Em humanos, ela está associada com sepse neonatal e meningite. Ela também pode afetar adultos imunocomprometidos, além de ser colonizadora comum do trato gastrointestinal e geniturinário. Para melhor conhecimento biológico desse microorganismo, o genoma de *S. agalactiae* GBS85147 foi sequenciado usando a plataforma *Ion Torrent* PGM, por biblioteca de fragmentos e uma cobertura de aproximadamente 246x foi obtida. Alcançou-se uma qualidade Phred maior ou igual a 20 em 91,25% das bases e os dados foram montados com o *software* Mira versão 3.9.18 (valor de N50 de 104.996 pb). Sobreposições entre contigs foram removidas usando *in-house scripts* e os gaps foram curados manualmente usando a extração da consenso do mapeamento dos dados brutos sobre o genoma das referências *S. agalactiae* GD201008 e *S. agalactiae* 09mas018883. Para a predição gênica foi utilizado o software FgenesB, utilizando-se a *S. agalactiae* 09mas018883 como referência, e os software RNAmmer e tRNAscan foram utilizados para predição de rRNA e tRNA, respectivamente. A curadoria de *frameshifts* foi realizada através do *software* CLC Genomics Workbench 7. O genoma foi manualmente anotado usando o banco de dados Uniprot e a ferramenta Blastp. O Interproscan 5 foi utilizado para avaliar as proteínas hipotéticas encontradas no genoma. *S. agalactiae* GBS85147 apresenta um genoma circular com 1.996.163 pb, um conteúdo GC de 35,98%, 1.925 sequências codificantes, 18 rRNA, 63 tRNA e 2 pseudogenes. Foram realizadas análises comparativas da linhagem com outras da mesma espécie para avaliar a similaridade, pois esta espécie bacteriana apresenta alta diversidade genética e um considerável número de hospedeiros, que destacam a importância do sequenciamento de novas linhagens, pois é essencial melhor conhecimento do organismo para o desenvolvimento de estratégias terapêuticas, determinar novos fatores de virulência, que podem contribuir para desenvolvimento de novas drogas, vacinas que podem ser úteis para minimizar os impactos socioeconômicos da bactéria na sociedade.

Palavras-chave: *Streptococcus agalactiae;* bactéria*;* montagem de genomas

Abstract

Streptococcus agalactiae (Lancefield group B, GBS) is a gram-positive and cocci, bacterial pathogen. This species can causes diseases in humans, cattles and fishes. In humans, it is associated with neonatal sepsis and meningitis. It can also affect immunocompromised adults, although also being a common colonizer of the gastrointestinal and genitourinary tracts. In this work, the genome was sequenced with PGM Ion Torrent, using fragment library approaches, and 200 bp sequencing kit, according to manufacturer's recommendations. A coverage approximately of 246x was obtained, with 578,082,183 bp in 2,973,022 reads with mean length of 203 bp and Phred quality greater than or equal to 20 in 91.25% of bases. The data was assembled with Mira Assembler version 3.9.18 (N50 length of 104.996 bp) using the recommended parameters. One hundred and four contigs were mapped over reference (S. agalactiae GD201008) using CONTiguator 2.0, and 34 contigs had similarity. There were micro contigs remaining, less than 600bp. The overlaps contigs were removed using "in-house scripts" and the last 8 gaps were filled manualy using extract consensus of reads map over S. agalactiae GD201008 and S. agalactiae 09mas018883. Through CLC Genomics Workbench 7 version it was possible to curate the frameshifts. For prediction of rRNA and tRNA, it was used RNAmer and tRNA Scan, respectly, through prokaryote parameters. For gene prediction, the data was performed in FgenesB, where using Streptococcus agalactiae 09mas018883 as reference. The genome was in Artemis, using the protein blast of Uniprot database. The Interproscan 5 was used for re-analysis of hypothetical proteins that were found in the genome. S. agalactiae strain GBS85147 is comprised by a circular chromosome with 1996163 bp. There are 1925 coding sequences, 18 rRNA genes, 63 tRNA genes and G+C content of 35,98%. The bacterium has a high genetic diversity and considerable amount of hosts, which highlights the importance of the sequencing of new samples for classification as to the pathogenesis, characteristics, etiology and target host. The elucidation of these aspects is essential for the development of improved therapeutic strategies, and new genomic comparative analysis between samples, realized in this work, can minimize the socio economic impact of bacteria in the society.

***Keywords**: Streptococcus agalactiae; bacteria; genome assembly*

Colaborações

Este trabalho foi desenvolvido por meio de uma rede de colaborações que envolve três laboratórios: LGCM (Laboratório de Genética Celular e Molecular), AQUACEN (Laboratório Oficial do Ministério da Pesca e Aquicultura), ambos sediados na UFMG (Universidade Federal de Minas Gerais) e o Laboratório de Biologia e Fisiologia de Estreptococos sediado na UERJ (Universidade do Estado do Rio de Janeiro), e tem como objetivo o desenvolvimento de novas pesquisas e trabalhos com microrganismos patogênicos de interesse médico ou veterinário. A orientação desse trabalho foi de responsabilidade do Prof. Dr. Vasco Azevedo (LGCM, UFMG) e a coorientação da Dra. Anne Cybele Pinto (LGCM, UFMG). Este trabalho contou com o auxílio financeiro da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), do CNPq (Conselbho Nacional de Desenvolvimento Científico e Tecnológico) e da FAPEMIG (Fundação de Amparo à Pesquisa de Minas Gerais).

Estrutura do Manuscrito

Este manuscrito está divido em seções de acordo com os temas a serem abordados, como segue: (i) a primeira parte conta com uma revisão de literatura sobre as principais características referentes ao gênero *Streptococcus* e à espécie *Streptococcus agalactiae*; serão abordados, ainda, temas sobre sequenciamento, montagem e finalização de genoma; (ii) a segunda parte será apresentada em capítulos: (a) no capítulo 1 será abordado o artigo científico com título "*Complete Genome Sequence of Streptococcus agalactiae Strain GBS85147 Sorotype Ia Isolated From Human Oropharynx*", submetido para a publicação no periódico SIGS; (b) no capítulo seguinte serão apresentadas análises comparativas entre genomas de diferentes linhagens de *Streptococcus agalactiae*, material e métodos, resultados e discussões complementares ao artigo relatado no capitulo 1; e (iii) por último, será abordada uma conclusão geral dos trabalhos realizados e futuras perspectivas. A produção acadêmica e as análises adicionais foram inseridas em anexo.

1. Revisão de Literatura

1.1 Microrganismos como alvos de diferentes estudos

Os microrganismos podem ser classificados e observados em diferentes reinos biológicos, sendo amplamente distribuídos pela natureza. A alta diversidade de microrganismos contribui para uma grande quantidade de informações que impulsionam estudos em diversas áreas como: agrícola, medicinal e industrial (Tortora, 2010).

Na agricultura, estudos relacionados com a utilização dos microrganismos são aplicados no desenvolvimento de defensivos agrícolas para o controle biológico como: pesticidas, praguicidas, e desinfetantes de solo, tendo como objetivo o aumento da produção e a melhoria na qualidade dos alimentos (Tortora, 2010). Nas áreas medicinal e industrial, é possível, em virtude da tecnologia do DNA recombinante, o desenvolvimento de microrganismos geneticamente modificados, que podem ser utilizados como ferramentas biotecnológicas para produção de vacinas e antibióticos (Madigan, 2004).

Os microrganismos podem ser detectados nos lugares mais “hostis”, afetando diretamente o padrão ambiental e outros organismos habitantes do mesmo. Levando-se em consideração essas informações, e ao extrapolar para o convívio humano, é possível apontar que tais características possam refletir em impactos significativos na economia, saúde e ambiente de forma positiva ou negativa (Madigan, 2004). Por isso, estudos com os microrganismos são de extrema importância para ciência e para o futuro da humanidade.

Dentre os microrganismos mais estudados estão as bactérias. Elas são de grande importância medicinal, biotecnológica, veterinária, ambiental e um dos mais antigos organismos da Terra, com amostras localizadas em rochas de 3,8 milhões de anos (Brandão, 2001). Bactérias são encontradas em todos os tipos de meio ambiente terrestres e aquáticos. Elas possuem uma alta diversidade e resistência a diferentes temperaturas, podendo ser encontradas entre 90ºC a 120ºC (hipertermófilas) e de 0ºC a 18ºC (psicrófilas) (Madigan, 2004).

As bactérias, geralmente, possuem uma parede celular rígida composta por uma camada de peptideoglicano, que fornece rigidez à parede, protege dos meios externos e determina o formato das bactérias (Quinn, 2005). Elas possuem também

uma alta diversidade de forma, como bacilos, cocos, helicoidais, filamentos ramificados, e com dimensões variando de 0,5µm a 5µm de comprimento (Madigan, 2004).

Existem vários fatores que tornam as bactérias bons modelos para diversos estudos genômicos e metabolômicos, como: conservação de funções celulares em comum com organismos mais complexos, a possibilidade de se obter rapidamente em meio de cultura adequado uma elevada quantidade de material biológico, e possui a vantagem de ter custo de desenvolvimento significativamente menor quando comparado a organismos eucariotos (Pelczar, 1996; Madigan, 2004).

De todas as características das bactérias, uma das mais importantes é a reprodução, pois ela permite a geração de mutações, transferências de genes entre espécies, trocas e ganhos de fatores de virulência e aquisição de novas resistências (Madigan, 2004). Os novos fatores de virulência e resistência aumentam significativamente o risco para os hospedeiros, pois a cada nova resistência adquirida, aumenta-se a necessidade de se desenvolver novos antibacterianos e bactericidas para combatê-los, sendo que a cada dia, antigos fármacos se tornam menos eficazes e, assim, aumentam o custo de pesquisas para se combater as novas resistências obtidas (Quinn, 2005; Tortora, 2010).

Tendo em vista as informações que são categoricamente discutidas pela literatura, nota-se que os microrganismos, em especial as bactérias, despertam grande interesse em pesquisas direcionadas as áreas abordadas neste trabalho. Tal observação pode ser visualizada na figura 1, que descreve o total de projetos de genoma depositados na base de dados do GOLD (Gold, 2015). Desde a sua criação, em 2007, o número de bactérias com genomas depositados sempre foi maior que dos demais domínios. Em 2011, o banco de dados armazenava aproximadamente 10 mil genomas bacterianos, e no ano de 2014, o valor chegou a aproximadamente 40 mil genomas.

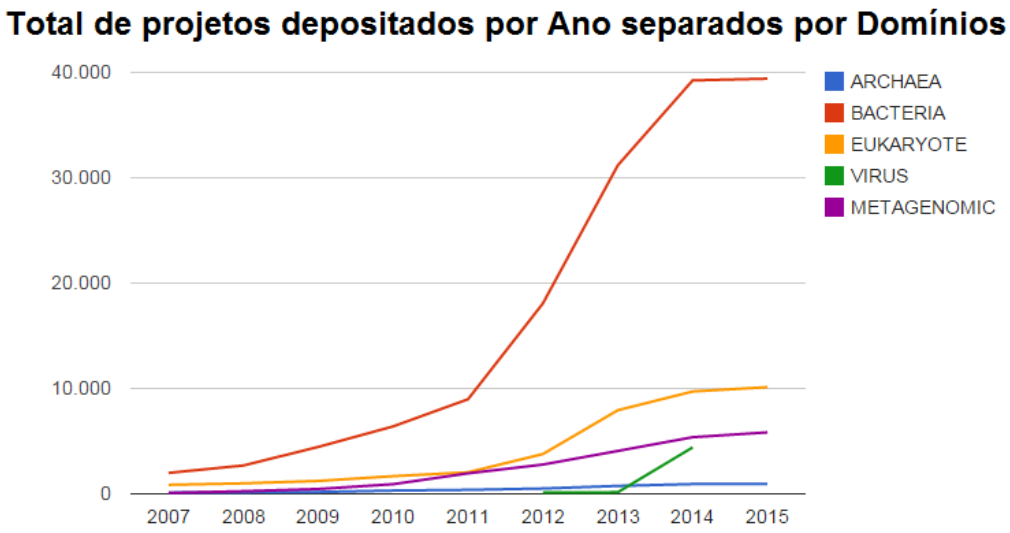


Figura 1 - Gráfico dos projetos separados por Domínios.
Fonte: Adaptado de https://gold.jgi-psf.org/statistics.

A quantidade de artigos e trabalhos envolvendo bactérias acompanha o crescimento do número de projetos relacionados aos genomas bacterianos, como é possível visualizar no figura 2. Na última década, a quantidade de artigos relacionados a bactérias registrados no PubMeb subiu de 53 mil para 82 mil, sendo que resultados prévios de 2015 já apresentam cerca de 71 mil artigos.

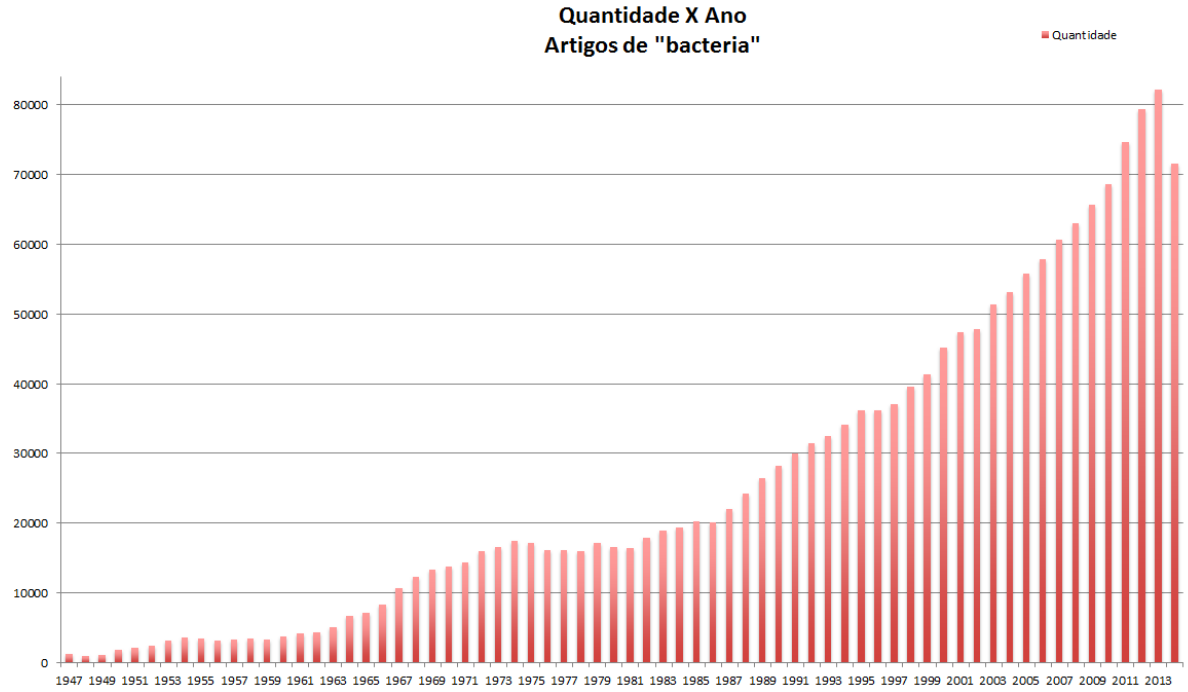


Figura 2 - Gráfico com Quantidade de Artigos por Ano relacionados a bactérias.
Fonte: Adaptado de http://www.ncbi.nlm.nih.gov/pubmed/?term=bacteria

Dentre os projetos de genomas bacterianos, há três grandes filos que possuem maior quantidade de dados depositada: *Proteobacterias* com 34%, *Firmicutes* com 21,9%, seguido por *Actiobacterias* com 12,7%, sendo que os demais filos representam 31,4% dos projetos, como é possível visualizar na Figura 3.

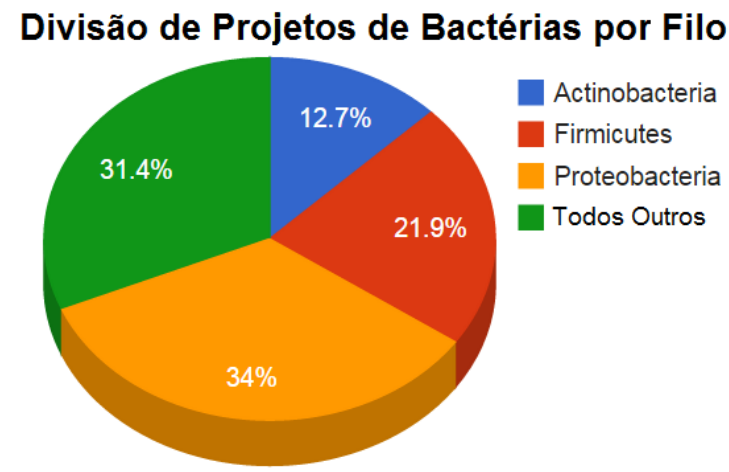


Figura 3 - Quantidade de Genomas de Bactérias por Filo.
Fonte: Adaptado de https://gold.jgi-psf.org/statistics

Dentro do filo das *Firmicutes* encontra-se o gênero *Streptococcus*, que possui elevada importância médica e veterinária. Esse gênero possui, atualmente, 110 espécies, 22 subespécies identificadas e, aproximadamente, 90 mil artigos relacionados (LPSN, 2015; PubMed Search *Streptococcus*, 2015). Dentre as espécies mais estudadas, destaca-se a *Streptococcus agalactiae*, uma bactéria patogênica que afeta diversos hospedeiros

1.2 *Streptococcus agalactiae*

O gênero *Streptococcus* é formado por bactérias em forma de cocos, microorganismos Gram-positivos, imóveis, não esporulam, são catalase e oxidase negativos. Aeróbias e anaeróbias facultativas, essas bactérias crescem em temperaturas entre 25ºC e 45ºC, podendo ser α, β hemoliticas ou não hemolítica. As dimensões variam entre 0,2 e 2,0 µm de diâmetro e são organizadas em pares ou cadeias lineares curtas. Possuem metabolismo fermentativo de açúcares que geram, como principal produto, o ácido lático e além disso, algumas espécies podem requerer CO_2 para o isolamento inicial (Ruoff, 2003; McPherson, 2007).

Existem diversas técnicas em biologia molecular capazes de comparar e classificar os isolados de *S. agalactiae* de seus hospedeiros, tais como: *Multilocus Enzyme Electrophoresis* (MLEE), *Multilocus Sequence Type* (MLST), *Random Amplification of Polymorphic DNA* (RAPD), *Restriction Fragment Lenght Polymorphism* (RFLP), Ribotipagem, *Pulsed-Field Gel Electrophoresis* (PFGE) e *Restriction Digestion Pattern* (RDP). Apesar de diferirem quanto ao funcionamento e custos, todas possuem objetivos em comum: estudar a diversidade genética das linhagens de GBS (*Group B Streptococcus*) e predizer se existe alguma associação entre as patogenias e os genótipos (Nakib *et al*., 2011).

O primeiro estudo com ferramentas de classificação de linhagens foi realizado em 1935 por Lancefield e Hare, que propuseram a criação de grupos para classificar as linhagens por meio da sorotipagem. Essa classificação baseava-se na antigenicidade de polissacarídeos capsulares (CPS), que são agrupados em 20 sorogrupos, denominados Grupos de Lancefield (A-H e K-V), os quais têm sido utilizados para caracterizar as diferentes espécies desse gênero (Lancefield *et al*., 1935; Shewmaker *et al*., 2007). O CPS do GBS possui uma estrutura molecular mais complexa que os CPS do GCS (*Group C Streptococcus*), pois possui uma estrutura central com duas unidades de ramnose, fosfato e glucitol. As cadeias laterais de ramnose, Galactose e N-acetilglicosamina são ambas ligadas na posição quatro da ramnose central (Pritchard *et al.,* 1984).

Os estudos com CPS são importantes instrumentos para auxiliar caracterização epidemiológica do GBS, pois a mesma possui uma capsula que é rica em ácido siálico, usada para classificar rapidamente uma amostra de GBS e a um baixo custo (Poyart *et al.,* 2007; Rajagopal, 2009). Porém, em algumas linhagens, a identificação dos sorotipos não é possível devido a modificações do CPS causada por mutações gênicas (Kong, 2008).

Estudos sugerem que diferentes pressões seletivas impostas pelo sistema imune do hospedeiro, têm levado a seleção de diferentes estruturas de CPS nos GBS (Cieslewicz *et al*., 2005). Além disso, a CPS é um importante fator de evasão do sistema imune dos hospedeiros, uma vez que o ácido siálico, também se encontra nas células dos vertebrados, fato que dificulta a identificação da CPS do GBS como antígeno (Cieslewicz *et al*., 2005).

Adicionalmente, os tipos capsulares variam conforme a região geográfica e o hospedeiro. Tais informações foram evidenciadas por amostras de gado em

fazendas alemãs, nas quais foi observado o predomínio dos sorotipos Ia e Ib. Em fazendas norte-americanas, o sorotipo predominante é o III, enquanto que em fazendas de gado do Brasil os sorotipos mais encontrados foram os V e III (Merl *et al*., 2003; Dogan *et al*., 2005; Pinto *et al*., 2013). Em amostras de peixe existe o predomínio quase global do sorotipo Ia (Evans *et al.,* 2008).

Nas amostras isoladas de humanos no Brasil, América Latina, Estados Unidos, Europa e Austrália, os sorotipos predominantes foram Ia, II, III e IV, enquanto que, no Japão existe um predomínio dos sorotipos VI e VIII (Winn, 2006; Kong *et al*., 2008; Dutra *et al*., 2014).

O sorogrupo B se refere à espécie de *Streptococcus agalactiae.* Essa espécie, alvo de estudo deste trabalho, foi descrita inicialmente em 1887 na França, por Nocard e Mollereau, como *Streptococcus* da mastite, e depois foi batizada com o nome de *Streptococcus agalactiae* por Lehmann e Neumann em 1896 (Metcalf, 1994).

S. agalactiae é um importante agente patogênico bacteriano causador de várias doenças em humanos, bovinos e peixes (Farley, 2001). Outras espécies também podem ser contaminadas e se tornar hospedeiros, como camelos, cães, cabras, cavalos, focas, galinhas, golfinhos, gatos, hamsters, macacos, nutrias, ovelhas, ratos e sapos, porém com menor ocorrência (Delannoy, 2013).

Dentre as principais patogenias causadas por essa bactéria em humanos, estão a pneumonia, a septicemia e a meningite em recém-nascidos, com elevadas taxas de mortalidade. *S. agalactiae* é uma das principais causadoras de sepse em gestantes e, nos últimos anos, tem aumentado o número de casos associados aos idosos e adultos imunocomprometidos (Johri *et al*., 2006; Nakib *et al*., 2011).

Na medicina veterinária, as linhagens GBS são consideradas responsáveis por altas taxas de morbidade e mortalidade, sendo observadas graves patogenias em bovídeos pela mastite, nos peixes por meningoencefalite e septicemia (Bowater *et al*., 2012; Keefe, 2012).

1.2.1 Doenças causadas por *S. agalactiae* em humanos

S. agalactiae é um agente causador de diversas patogenicidades em seres humanos em diferentes fases. Na fase infantil é frequentemente associada a meningite, sépsis neonatal e sépsis puerperal (Beitune, 2005; Rajagopal, 2009;).

O GBS é um dos maiores causadores de bacteremia e meningite em recém-nascidos (RN), e também responsável pelo maior número de infecções bacterianas fatais (Rajagopal, 2009), onde os principais fatores de risco para as doenças nos RN são: colonização materna e prematuridade. Em estudos realizados nos Estados Unidos foram observados que os casos positivos variam de 20% a 40% em grávidas, enquanto que no Brasil, variam de 15% a 25% (Pogere *et al.,* 2005; Springman *et al.,* 2009). O número de casos de septicemia nos Estados Unidos em RN foi de aproximadamente 0,3 a cada 1000, enquanto que na Europa, este valor variou entre 0,24 e 1,26 para cada 1000 RN (Spellerberg, 2000; Johri *et al.*, 2006).

A infecção nos RNs por GBS *in utero* ocorre através de uma infecção ascendente, principalmente durante o parto, devido à aspiração de fluidos vaginais contaminados e, em menor incidência, durante a amamentação (Spellerberg, 2000; Glaser *et al.,* 2002).

Quando a infecção ocorre entre as primeiras 6 a 8 horas de vida, é denominada *early-onset disease* (EOD) e nessa fase podem ocorrer bacteremia, choque séptico e falência respiratória. Quando a infecção é tardia, ocorrendo em alguns meses após o parto, é denominada *late-onset diseases* (LOD), e é nessa etapa que a meningite pode causar sérios danos neurológicos irreversíveis com elevada taxa de mortalidade (Doran *et al.,* 2004; Rajagopal, 2009).

Nos adultos, a *S. agalactiae* ocorre preferencialmente em idosos ou pessoas imunocomprometidas. Nessa fase as apresentações clínicas são: bacteremia primária, endocardite, infecção do trato urinário, peritonite, pneumonia, osteomielite e infecções nos tecidos conjuntivo, epitelial e muscular (Del Pozo *et al.,* 2000; Johri *et al.,* 2006). Os maiores fatores de risco na população adulta são câncer, doenças cardiovasculares, diabete e hepatite (Johri *et al.,* 2006). Estudos envolvendo a classificação de casos por sorotipo em humanos, demonstraram que o sorotipo Ia está mais relacionado às infecções intrauterinas e o sorotipo V a casos de infecção em homens (adultos) e mulheres (adultas e não gestantes) (Quentin *et al.,* 1995; Harrison *et al.,* 1998). A infecção por *S. agalactiae* varia de acordo com diversos fatores, como: grupo étnico, localização geográfica e idade. Estudos realizados nos Estados Unidos indicam que a infecção ocorre com uma maior frequência em mulheres hispânicas, seguidas pelas negras, mulheres mais velhas e com menor número de partos (Beitune, 2005).

A principal ferramenta na epidemiologia molecular de *S. agalactiae*, em casos de humanos, é a técnica de MLST (*Multilocus sequence type*), pois apresenta elevada robustez para análises epidemiológicas, monitoramento de patógenos, estudos evolutivos e análises de estruturas populacionais (Pavón, 2009).

Todos os dados resultantes das análises epidemiológicas utilizando o MLST estão depositados em um banco de dados (BD) e disponíveis no *website* http://pubmlst.org/sagalactiae/. Esta é uma importante vantagem na utilização da técnica de MLST, já que esse BD se comporta como enciclopédia, permitindo que os isolados bacterianos sejam comparados em todo o mundo (Chan, 2011).

Para se combater a infecção em gestantes, é utilizado o protocolo da CDC (*Centers for Disease Control and Prevention*), que recomenda a utilização de quimioprofilaxia com penicilina, ampicilina, eritromicina ou clindamicina, durante o parto (CDC, 1996). Estudos demonstram que a detecção do GBS no trato genital no final da gestação (entre a 35ª e a 40ª semanas) é uma conduta bastante efetiva para prevenir infecção dos neonatos, uma vez que a colonização por GBS é intermitente. Gestantes não colonizadas no período da gestação podem apresentar sinais positivos em período posteriores (Regan *et al.,* 1996).

Vários estudos têm demonstrado o aparecimento de linhagens GBS resistentes a antimicrobianos em diversos países, e essa resistência vem aumentando com passar dos anos (de Azavedo *et al.,* 2001; Hsueh *et al.,* 2001; d' Oliveira *et al.,* 2003). Nos últimos anos, vem ocorrendo um aumento na resistência do GBS à clindamicina. Fernandez e colaboradores (1998) demonstraram em estudos realizados nos Estados Unidos, que a resistência do GBS ao antibiótico era de 3,4% em 1998 (Fernandez *et al.,* 1998). Morales e colaboradores (1999) observaram resistência de 5%, Andrews e colaboradores (2000) encontraram 7% de resistência e Biedenbach e colaboradores (2003) 11,4%.

Levando-se em consideração os resultados avaliados por todos esses trabalhos, pode-se observar que a resistência à clindamicina quase quadruplicou em apenas cinco anos, mostrando que, possivelmente, o uso inadequado de antibióticos pode levar a resistência aos mesmos (McDonald *et al.,* 2003).

Sequenciamentos de novos isolados de humanos podem colaborar para análises comparativas mais robustas, auxiliar no desenvolvimento de novos fármacos, na criação ou melhoria de métodos de classificação das linhagens, ações essas que pode minimizar o impacto do GBS na sociedade.

1.2.2 Doenças causadas por *S. agalactiae* em bovideos

Dentre as diversas patogenias causadas pela *S. agalactiae* destaca-se a mastite crônica e infecciosa nos bovídeos e em camelídeos (Zubair, 2013). Essa doença causa grandes prejuízos ao agronegócio. Além da *S. agalactiae,* tem como principais agentes etiológicos: *Corynebacterium sp, Staphylococcus aureus, Streptococcus dysgalactiae, Streptococcus uberis* e *Escherichia coli*, ainda fungos e algas (Radostits *et al*., 2002; Holtenius, 2004). Diante da vasta biodisponibilidade de agente infectantes para a mastite, a doença ainda não foi amplamente compreendida, tornando-a ainda, em alguns aspectos, obscura. (Radostits *et al*., 2002).

Contudo, alguns estudos avaliaram informações intrínsecas na manifestação da doença, no qual se tem conhecimento que o microrganismo invade a glândula mamária, atravessando o canal do teto, e começa a se multiplicar no interior dos tecidos. Essa invasão de microrganismos pode ocorrer pela falta de higiene entre as ordenhas e pela manipulação de material contaminado no tratamento do animal. Os principais fatores responsáveis pela ocorrência da mastite são: o ambiente, o agente patogênico, a resistência do animal e o estágio de lactação (Santos *et al*., 2007).

Para a prevenção da mastite é necessário incorporar um conjunto de procedimentos simples por parte do ordenhador, por exemplo, higienização e desinfecção do ambiente, do animal, do profissional e de todos os utensílios utilizados logo após a ordenha. Tais procedimentos reduzem consideravelmente a chance de contaminação de outros animais (Embrapa, 2012).

A mastite é uma doença de grande importância epidemiológica. A maioria dos casos ocorre na forma subclínica, mas existem também as formas clínica e crônica, ambas causadoras de severos prejuízos econômicos (Radostits *et al.,*2002). A forma subclínica é responsável por reduzir a capacidade funcional das glândulas mamárias, possuindo um diagnostico mais complexo e de custo mais elevado, necessitando do isolamento do agente etiológico, de exames complementares baseados no conteúdo celular ou em modificações bioquímicas na composição do leite (Pankey *et al.,*1991; Radostits *et al.,*2002).

O tratamento das infecções intramamárias causadas pela mastite é complexo e requer a intervenção por meio de antibióticos e antimicrobianos de amplo espectro sistêmicos e com ação local (Diário Oficial, 2011).

Estudos realizados no Reino Unido indicam que mastite clínica tem uma taxa de 37,5 casos/100 vacas-ano e que as infecções intramamárias geram um caso de mastite clínica a cada 34 segundos (Leigh, 1999).

No Brasil, a média estimada de perda anual é de aproximadamente US$ 100 por vaca, sem incluir os custos de prevenção (Carneiro *et al*., 2004). Na Europa estima-se uma perda de US$ 226. Em 2011, Peres contabilizou todos os prejuízos, como por exemplo, perda de 70% causada pela redução de produção dos quartos mamários; 14% de desvalorização dos animais, descarte ou morte do animal; 8% pela perda do leite descartado; 8% pelos gastos com tratamentos e prevenção e honorários de veterinários (Peres, 2011).

Quantificar a perda associada à mastite subclínica é um processo bem complexo, mas nessa forma ela é 4 vezes mais prevalente do que a mastite clínica. As perdas da mastite subclínica foram de aproximadamente 70% a 80% enquanto que a mastite clínica foi de aproximadamente 20% a 30% (Philpot, 1991).

Por ser uma das principais causadoras da mastite em bovídeos, a bactéria *S. agalactiae* afeta a produção de diversos setores, como: de leite e lacto derivados, gado de corte, peles e lãs. Devido às consideráveis perdas financeiras, a mastite é considerada uma doença de grande importância epidemiológica e econômica.

1.2.3 Doenças causadas por *S. agalactiae* em peixes

Há alguns anos foi observado um aumento na ocorrência de casos clínicos associados a *S. agalactiae* em diversas regiões geográficas, sendo considerado um patógeno emergente em peixes de água salgada e doce (Evans *et al.,* 2002a).

O primeiro caso de infecção em peixe por *S. agalactiae* foi descrito em 1956 no Japão, onde houve um surto de septicemia em uma fazenda comercial de truta arco-íris (*Oncorhynchus mykiss*) (Hoshina *et al,* 1958). Já existem mais de 20 espécies que tiveram casos de infecções identificados (Olivares-fuster *et al.,* 2008).

As principais espécies do gênero *Streptococcus* causadoras de septicemia e meningoencefalite em peixes são: *S. agalactiae, S. iniae, S. parauberis, S. dysgalactiae, S. phocae e S. ictaluri* (Chen, 2012; Figueiredo et al., 2012). Dentre essas, duas espécies: a *S. iniae* e *S. agalactiae* são responsáveis por causar maior impacto econômico, com sinais clínicos, infecção e hospedeiros bem semelhantes (Evans et al., 2002b).

As doenças septicêmicas originadas por bactérias do gênero *Streptococcus* são responsáveis por contaminar peixes de água doce, salgada e ambientes estuarinos, são tratadas com o termo estreptococose (Mata et al., 2004).

A infecção por *S. agalactiae* ocorre quando o peixe contaminado, vivo ou morto, libera a bactéria na água colonizando a pele dos demais peixes. Outra forma de infecção é por meio do acometimento de infecções invasivas que apresentam um alto índice de mortalidade, sendo a bactéria capaz de sobreviver por longos períodos na água, na lama, no tanque, na estufa e nos equipamentos de trabalho (Suresh, 1998).

O consumo de peixes não fiscalizados, possíveis portadores de estreptococose, tem sido associado com um aumento de risco de contaminação por *S. agalactiae* com os sorotipos Ia e Ib, principalmente em humanos (Foxman, 2007). De acordo com Organização das Nações Unidas, o setor de produção animal da aquicultura mundial tem apresentado um forte crescimento nas últimas décadas para alimentação (Food and Agriculture Organization, 2010).

A aquicultura brasileira também está em alta, assim como a mundial, apresentando um aumento de 15,3% em 2010 quando comparado com a produção de 2009, sendo que a piscicultura continental representou 82,3% da produção nacional e o peixe mais produzido no país é a tilápia do Nilo, que representa aproximadamente 40% da produção nacional (Ministério da Pesca e Aquicultura, 2010).

No Brasil, existem diversos obstáculos que dificultam o crescimento da aquicultura, como a ocorrência de surtos de doenças infecciosas, principalmente as causadas por bactérias, protozoários e fungos. Tais doenças ocasionam severas perdas econômicas aos produtores, tornando a produção inviável em determinados casos (Duremdez, 2004; Mian *et al*., 2009). A infecção por *S. agalactiae* causa uma elevada taxa de mortalidade podendo chegar a 90% do plantel, sendo que normalmente a infecção é próxima da fase pré-comercialização, fase em que já ocorreu elevado consumo de ração, o componente de maior custo para produção (Evans, 2004).

Para crescimento na criação de tilápia no Brasil e no mundo é necessário adotar novas estratégias para reduzir o impacto da *S. agalactiae* na produção, como aplicação de antimicrobianos. Porém, normalmente, é tardia, quando já ocorreram grandes perdas no plantel. O uso de antimicrobianos minimiza a ocorrência da

infecção nos peixes não infectados ou animais que se encontram no início da infecção, mas não age na cura dos peixes que já possuem sinais clínicos (Heuer *et al.,* 2009; Rattanachaikunsopon, 2009). Além disso, a utilização de antimicrobianos pode causar riscos ao meio ambiente e à segurança alimentar (Heuer *et al.,* 2009), e aumentar o risco da resistência bacteriana, em casos de uso a longo prazo (Lim, 2006).

Como o tratamento pós-infecção não é efetivo em estreptococose nos peixes, é necessária a aplicação de métodos imunoprofiláticos, como a vacinação, o que seria uma alternativa mais viável para prevenção e controle das infecções por *Streptococcus*. Para tal, faz-se necessários estudos genômicos das ilhas de patogenicidade, dos fatores virulência nos diferentes hospedeiros, e principalmente das proteínas secretadas e de membrana, que possibilitariam o desenvolvimento de vacinas eficazes

1.3 Fatores de Virulência

Genes de virulência são responsáveis pela adaptação e sobrevivência da bactéria dentro hospedeiro no processo de infecção. Normalmente, são expressos em contato com ou durante a invasão do hospedeiro e a grande maioria estão presentes em ilhas de patogenicidade (Hare, 2014)

Os patógenos envolvidos na etiologia das doenças causada por GBS podem apresentar diversos fatores de virulência que facilitam a colonização e a infecção nos hospedeiros. Alguns desses patógenos são capazes de evadir as defesas do hospedeiro ao se fixarem nas células epiteliais, produzem cápsulas que atrapalham a captura e a eliminação pelos neutrófilos. Além disso, podem produzir exotoxinas e endotoxinas que inativam ou destroem os leucócitos. Podem ainda manter-se no interior das células para escapar da resposta imune do hospedeiro (Bradley, 2002; Carneiro *et al.,* 2009).

Para que ocorra a infecção por *S. agalactiae,* a mesma deve entrar em contato com o hospedeiro e atravessar as barreiras epiteliais. Para isso, são utilizados os fatores de virulência, que funcionam por meio de diferentes estratégias para que bactérias GBS possam invadir o hospedeiro (Baron *et al.,* 2005; Pietrocola *et al.,* 2005). Assim, os GBS podem utilizar as estruturas que se encontram em sua superfície ou secretar outros produtos no ambiente circundante. Além disso, para

sobreviver, a bactéria pode substituir as suas funções essenciais (Baron *et al.,* 2005).

Um dos principais fatores de virulência em *S. agalactiae* é o CPS, que auxilia na evasão bacteriana. Outro importante fator de virulência são os pili, que facilitam a adesão dos GBS em células epiteliais e no endotélio vascular (Lauer *et al,* 2005; Melin, 2011). A *C5a* peptidase, que é codificada pelo gene *scpB* de *S. agalactiae*, impede o recrutamento de neutrófilos nos locais onde ocorre a infecção e intermedia a ligação da bactéria à fibronectina humana, auxiliando a invasão nas células epiteliais (Melin, 2011; Spellerberg et al, 2000).

As proteínas da membrana externa possuem um papel muito importante na interação com o hospedeiro. Elas são capazes de mediar a invasão, e, por esse fato, podem ser consideradas como fatores de virulência (Safadi, 2010; Spellerberg, 2000). Nos GBS, as mais caracterizadas são as proteínas do antígeno C, as quais formam um complexo que contém as proteínas de superfície da família *Alpha-like proteins* (Alp): Alpha C, Rib, Epsilon (Alp1), Alp2, Alp3 e Alp4. Essas seis proteínas são caracterizadas por possuírem longas sequências de repetição e, no decorrer da infecção, esse número de repetições pode sofrer deleções internas. Ambos os fatos dificultam o reconhecimento pelo sistema imune do hospedeiro (Broker, 2004; Creti, 2004; Spellerberg, 2000). Estudos realizados com amostras do sorotipo Ia demonstraram a presença das proteínas Alpha C, Alp1 e Alp2, enquanto no sorotipo VII é encontrada a proteína Alp3 que não existe em Ia. Determinados fatores de virulência foram localizados em apenas alguns sorotipos, fato esse que apoia a teoria de que os fatores de virulência não são muito conservados entre os sorotipos (Creti, 2004).

A Tabela 1 demonstra os principais fatores de virulência descritos em amostras de GBS em humanos.

Tabela 1 - Alguns dos principais fatores de virulência de *S. agalactiae* com hospedeiro humano, funcionamento e genes envolvidos.
Fonte: Adaptado de Rajagopal (2009).

Fator de virulência	**Gene**	**Função**
Toxinas formadoras de poros		
Fator CAMP	*cfb*	Gerar poros na membrana da célula do hospedeiro
β−hemolisina / citolisina	*cylE*	Lise da celular do hospedeiro, induzir resposta

		inflamatória e apoptose
Fatores de evasão do sistema imune		
Cápsula rica em ácido siálico	*cpsA-L*	Ocultar o ácido siálico da célula do hospedeiro para dificultar o reconhecimento pelo sistema imunológico
C5a peptidase	*scpB*	Clivar a subunidade C5a do sistema Promover a aderência à fibronectina da matriz extracelular
Superoxido dismutase	*sodA*	Proteger os radicais de oxigênios e dos superóxidos
Serine protease	*cspA*	Clivar o fibrinogênio e quimiocinas
Aderência e Invasão		
Proteína ligadora ao fibrinogênio A	*fbsA*	Promover a aderência do GBS às células do hospedeiro se ligando ao fibrinogênio
Proteína ligada à laminina	*lmb*	Promover a aderência do GBS à laminina da célula do hospedeiro
Proteína serine-rich repeat 1	*srr-1*	Srr-1 promover aderência às células epiteliais
Proteína serine-rich repeat 2	*srr-2*	Srr-2 aumentar a virulência do GBS
Adesina bacteriana imunogênica	*bibA*	Promover a aderência do GBS às células do hospedeiro
Proteína C	*bca/bac*	Auxiliar a aderência do GBS às células epiteliais
Proteína de ligação ao fibrinogênio B	*fbsB*	Auxiliar a invasão do GBS nas células do hospedeiro
Gene associado à invasão e regulação	*iagA*	Auxiliar a invasão da barreira hemato-encefálica

Apesar da existência de genes responsáveis pelos fatores virulência, os mesmos podem não estar ativos, uma vez que podem estar presentes, mas não serem expressos. Isso ocorre, pois a regulação gênica pode ser diferenciada de acordo com as diferentes linhagens, diferentes sorotipos e com o ambiente (Jiang *et al.,* 2008). Sabe-se que a expressão de diversos genes de fatores de virulência é diferenciada em linhagens de distintos sorotipos (Sharma et al., 2012).

A regulação da expressão gênica em resposta aos estímulos externos é realizada pelo Sistema de Transdução de Sinal (STS), que pode regular os fatores transcricionais do DNA. O regulador mais comum em bactérias é o *two-component systems* (TCS) (Rajagopal, 2009).

Estudos realizados com sequenciamento completo de genomas de *S. agalactiae* revelaram a existência de 17 a 20 TCS. Esse elevado valor pressupõe

que as TCS podem facilitar as respostas às mudanças do GBS em diferentes hospedeiros e ambientes (Glaser *et al.*, 2002).

As regiões genômicas com maior concentração de fatores de determinados fatores como patogenicidade, resistência são caracterizadas como ilhas genômicas. Estudos indicam a existência de ilhas genômicas presentes na espécie *S. agalactiae*, uma elevada variedade de fatores, aliado a uma elevada plasticidade e uma variada distribuição de elementos móveis em diferentes linhagens de hospedeiros específicos e que alguns destes elementos podem ser característicos de isolados de um hospedeiro específico (Tettelin *et al.,* 2005, Richards *et al.,* 2011).

No entanto, são necessários estudos mais aprofundados sobre os fatores de virulência, das ilhas genômicas e as expressões gênicas sobre a interação entre o patógeno e o hospedeiro. Por isso, os estudos genômicos são importantes para a busca de informações que contribuam para o melhor conhecimento do microorganismo.

1.4 Estudos genômicos em *S. agalactiae*

O sequenciamento é uma técnica que permite revelar as sequências de bases nucléicas, permitindo assim, desenvolver estudos estruturais e funcionais do genoma, buscando compreender a complexa maquinaria biológica (Ribeiro et al., 2012).

O primeiro genoma completo de *S. agalactiae* foi depositado no banco de dados do GenBank no ano de 2002. O genoma de *S. agalactiae* NEM316, do sorotipo III, isolado de caso fatal de humano com septicemia, foi sequenciado utilizando o método de Sanger, e obteve-se um genoma com tamanho de 2.211.485 bp e 2.118 genes (Glaser *et al*., 2002).

A partir do sequenciamento desse genoma foi possível comparar o genoma de *S. agalactiae* com o genoma completo de *S. pyogenes,* observando-se uma similaridade de 25%. Além disso, foi observado que *S. agalactiae* era uma espécie diferente das demais espécies de *Streptococcus*, principalmente nas regiões gênicas responsáveis pela codificação de proteínas relacionadas aos elementos móveis (Tettelin *et al.,* 2005). Essas regiões gênicas possuem *transposons*, bacteriófagos e elementos de inserção. Acredita-se que esses genes estão fortemente associados à adaptação e à capacidade de colonizar diversos ambientes e causar doenças em

vários hospedeiros. Os mesmos são adquiridos por transferência horizontal de outras espécies de bactérias (Tettelin, *et al.,* 2005; Zubair *et al.,* 2013).

Estudos do pan-genoma de *S. agalactiae* revelaram que o genoma central, ou seja, o conjunto de genes presentes em todas as espécies em estudo, é formado por 80% do total de genes. Esses estão associados às funções reguladoras e aos genes essenciais (*housekeeping genes*). O genoma dispensável, ou seja, o conjunto de genes ausentes em algumas amostras, é formado por elementos móveis e extracromossomais. A formação do genoma dispensável possivelmente ocorre por meio da aquisição dos genes de outras espécies por transferência horizontal (Tettelini *et al.,* 2005; Richards *et al.,*2011).

Estudos realizados com modelos estatísticos predisseram que o pan-genoma do GBS ainda se encontra aberto, pois novos genes continuam a ser identificados e descritos a cada novo isolado sequenciado. Aproximadamente, 33 novos genes são adicionados para cada novo genoma descrito (Tettelin *et al.,* 2005).

Até o momento (1 de fevereiro de 2015), 16 genomas completos de *S. agalactiae* podem ser encontrados no banco de dados do National Center for Biotechnology Information (NCBI). Dentre as linhagens com os genomas completos, 10 foram isoladas de humanos, 4 de peixes, 2 de camelos e 1 de bovino. As linhagens de *S. agalactiae* possuem em média, o genoma variando de 1,8Mb a 2,2 Mb, quantidade de tRNA variando de 63 a 81, rRNA de 18 a 21 e conteúdo GC de 35,3% a 35,9% (Tabela 2).

Tabela 2 - Informações de genomas completos da *S. agalactiae* no NCBI.
Fonte: http://www.ncbi.nlm.nih.gov/genome/genomegroups/186 (acesso em: 1 de fevereiro de 2015).

Linhagem	Tamanho (Mb)	GC%	Hospedeiro	Sorotipo	Gene	tRNA	rRNA
2603V/R	2.16027	35.60	Humano	V	2279	80	21
A909	2.12784	35.60	Humano	IC	2154	80	21
COH1	2.06507	35.40	Humano	III	2069	80	21
NGBS061	2.22121	35.50	Humano	IV	2249	81	21
NGBS572	2.06143	35.50	Humano	IV	2069	81	21
GBS6	2.23148	35.80	Humano	II	2223	80	21
GBS2NM	2.2143	35.90	Humano	II	2202	80	21
GBS1NY	2.24371	35.90	Humano	II	2227	81	21
CNCTC 10/84	2.01384	35.40	Humano	V	2035	80	21

09mas018883	2.13869	35.50	*Bos taurus*	Não Informado	2153	80	21
ILRI112	2.0292	35.30	*Camelus dromedarius*	Não Informado	2173	79	21
ILRI005	2.10976	35.40	*Camelus dromedarius*	Não Informado	2180	80	21
GD201008-001	2.06311	35.60	Peixe Tilápia	IA	2061	77	21
SA20-06	1.82089	35.60	Peixe Tilápia	IB	1863	79	21
138P	1.8387	35.50	Peixe Tilápia	Não Informado	1862	69	16
138spar	1.83813	35.50	Peixe Tilápia	Não Informado	1859	67	16

1.4.1 Métodos de sequenciamento genômico

Na década de 70 foram criados os primeiros métodos de sequenciamento: o de terminação de cadeira por Sanger e Coulson, e alguns anos após, foi criado o método químico de degradação de bases por Maxam e Gilbert. Esses métodos foram um importante marco para o sequenciamento genômico (Kaur, 2013).

Os métodos de sequenciamento passaram por melhorias significativas, principalmente em relação à acurácia e ao custo. O método de Sanger foi responsável pelo primeiro sequenciamento de genoma completo em 1995: da bactéria *Haemophilus influenzae* (Fleischmann *et al*., 1995). No entanto, esse método possui um tempo de execução longo e o custo de sequenciamento por par de bases (pb) é alto, o que inviabiliza o sequenciamento de grandes genomas (Bonetta, 2006).

Um avanço alcançado na ciência foi o sequenciamento do genoma humano em 2001, o qual tinha como objetivos identificar os genes dos 23 pares de cromossomos, mapear todos os genes responsáveis pelas características normais e pelas patológicas, armazenar as informações em bancos de dados, desenvolver novas ferramentas eficientes para analisar os dados e criar meios para usá-los em estudos biológicos e medicinais. Foram necessários 13 anos para ser finalizado e teve um custo aproximado de três bilhões de dólares (Husemann, 2001).

Passados quatro anos do término do projeto do genoma humano, em 2005, foram lançadas as plataformas de sequenciamento de próxima geração, ou *Next-Generation Sequencing* (NGS). As NGS são responsáveis pelo aumento da cobertura do genoma, por leituras de DNA cada vez mais acuradas, pela redução no tempo de corrida, pela facilidade em sequenciar genomas completos e pela redução

significativa no custo por sequenciamento (Metzker, 2010; Loman *et al*., 2012; El-Metwally *et al.,* 2013).

As plataformas NGS apresentam metodologias diferentes do método de Sanger e têm como principal vantagem o alto volume de dados gerados por corrida. Algumas plataformas são capazes de gerar em 24 horas, uma quantidade de dados igual ou superior à que seria gerada por centenas de sequenciadores do tipo capilar. Essas vantagens possibilitaram um crescimento significativo na quantidade de projetos de sequenciamento de genomas completos, principalmente para organismos procariotos (Husemann, 2011).

Dentre as plataformas NGS pertencentes à segunda geração destacam-se: 454 GS FLX *system* (Roche), Illumina GA llx (Illumina), HeliScope (Helicos) e SOLiD 5500 XL *system* (ABI). Dentre as plataformas de terceira geração destacam-se: a plataforma PacBio RS *system* (Pacific Biosciences) e a Ion Personal Genome Machine – Ion PGM ™ (Life Technologies), sendo essa última considerada uma tecnologia de transição entre a segunda e terceira geração (Kaur, 2013).

1.4.2 Plataforma Ion Torrent

A plataforma Ion PGM (Life Technologies) foi lançada em 2011 e possui um sequenciamento que funciona por meio de detecção das alterações de pH dentro de um *chip* de silício durante a reação de sequenciamento. Os principais fatores que se alteram entre os modelos de *chip* são o volume de dados gerados e o tamanho das leituras. Uma das versões mais utilizadas é o *chip* 318 versão 2, o qual é capaz de produzir de 0,6 a 1Gb de dados com leituras de 200pb. A etapa de processamento das leituras sequenciadas possui uma execução veloz de aproximadamente duas horas (Jünemann *et al.,* 2013; Life Technologies, 2015).

Antes do processo de sequenciamento é necessário a construção da biblioteca genômica (Figura 4). Nesse processo, a amostra de DNA é fragmentada, e cada fragmento é ligado a adaptadores, e em seguida a uma *bead*, e será amplificado por PCR em emulsão. O produto de cada PCR é separado e inserido em um dos poços do *chip* de silício, que possui um tamanho um pouco maior que a bead coberta com o produtos de amplificação para que cada poço tenha somente copias do mesmo fragmento a ser sequenciado (Rothberg *et al.,* 2011). Dentro de cada poço ocorre a reação de sequenciamento em rodadas. Em cada rodada, o sequenciador disponibiliza uma determinada quantidade de um dos quatros tipos de

dNTP: dATP, dTTP, dGTP e dCTP. Várias DNA polimerases incorporam o dNTP à nova fita de DNA a ser sintetizada, caso este seja complementar à fita molde. A Figura 5 demonstra a incorporação de um novo nucleotídeo por meio da reação de DNA polimerase (Rothberg et al., 2011).

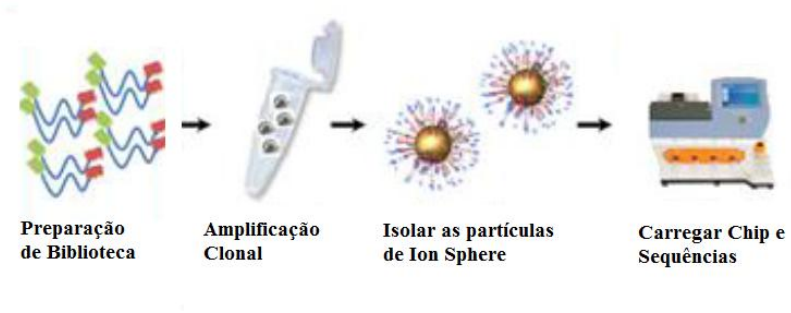


Figura 4 - *Workflow* de passos resumidos do processo de sequenciamento do Ion PGM.
Fonte: Adaptado de Life Technologies, 2015

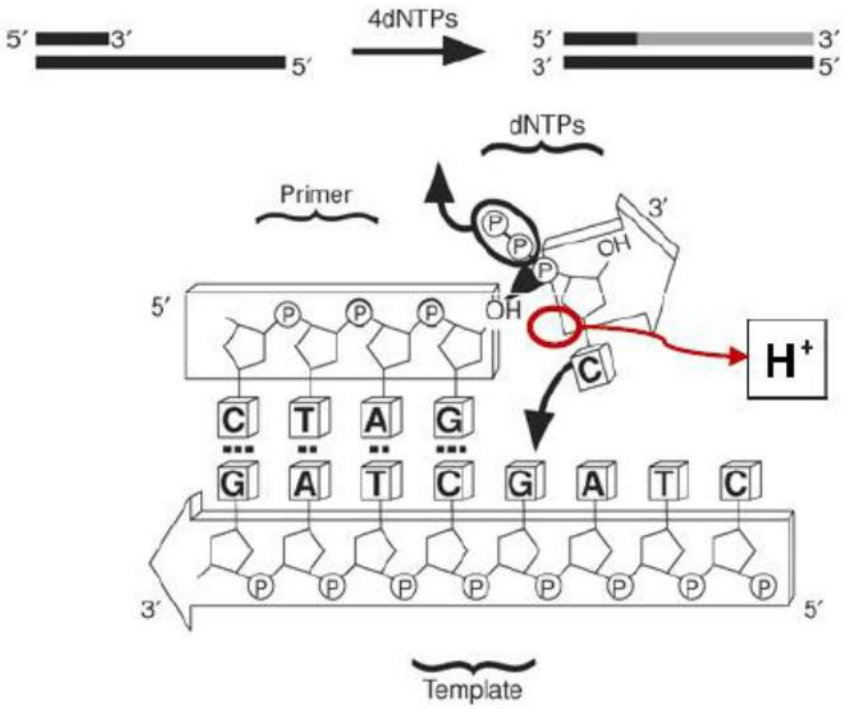


Figura 5 - *Template* da Reação DNA Polimerase
Fonte: Adaptado de Life Technologies (2015)

A reação de polimerização libera em cada ligação um átomo de hidrogênio que é responsável por alterar o pH de cada poço do *chip*. O transistor ISFET detecta essa alteração de pH e gera um sinal elétrico com intensidade variável de acordo

com a base, como é possível visualizar na Figura 6 (Loman *et al.,* 2012; Life Technologies, 2015).

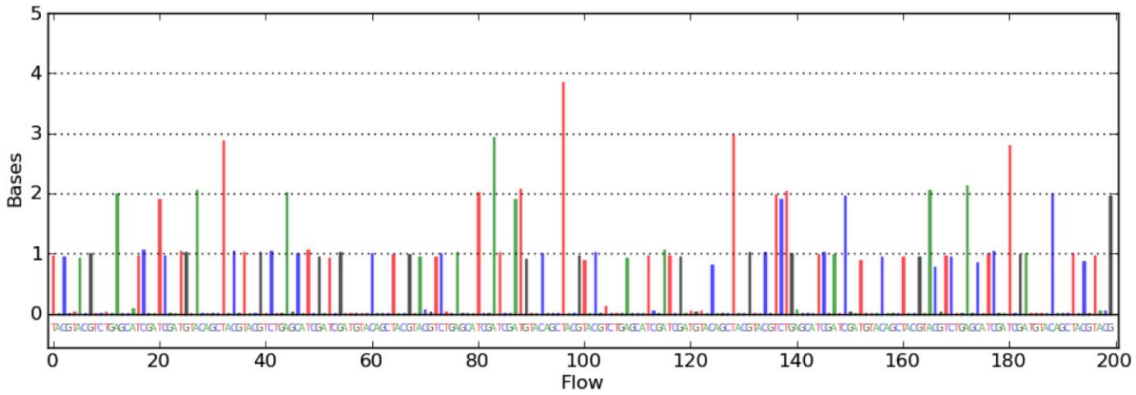


Figura 6 - Relatório dos sinais elétricos gerados pelo transistor ISFET.
Fonte: Life Technologies, 2015

Um fator limitante das plataformas de sequenciamento são os homopolímeros, ou seja, sequências repetitivas de um único nucleotídeo que são incorporadas em uma única rodada. O sensor do transistor ISFET não é capaz de detectar com precisão tais regiões (Rothberg *et al.,* 2011). Isso acarreta uma alta taxa de erros de inserção e deleção de sequências (*indel*) (aproximadamente 0,39% na versão 2 do chip 318). Apesar dessas limitações, a plataforma Ion PGM possui uma boa eficácia para o sequenciamento de genomas completos de procariotos, melhor custo benefício e menor custo por pb quando comparado a outras plataformas NGS (Ramos *et al.,* 2012; Loman *et al.,* 2012b; Jünemann *et al.*, 2013; Mariano, 2015). Essas vantagens são importantes diferenciais na escolha de uma plataforma de sequenciamento NGS.

A utilização de plataformas NGS para sequenciamento de genomas completos vem auxiliando a microbiologia a compreender os surtos epidêmicos, a imunologia auxiliando na criação de novas vacinas e fármacos e diversas outras funções (Loman *et al.,* 2012a).

Porém, com o surgimento destas novas tecnologias aplicadas nas plataformas NGS, surgiram novos desafios a serem resolvidos, como obter equipamentos robustos que possuam recursos computacionais suficientes para armazenar e analisar os dados gerados, realizar montagens de genoma com pequenas leituras e suportar as repetições existentes no genoma (Metzker, 2010).

1.4.3 Montagem e finalização de genomas

A montagem de genomas constitui no processo de ordenação e orientação das leituras obtidas por sequenciamento NGS (Miller *et al.*, 2010). Os processos de montagem e finalização de genomas a partir de metodologias recentes podem ser divididos em três etapas: (i) análise de dados; (ii) montagem *de novo* e (iii) *scaffolding* (Mariano, 2015).

Um dos processos essenciais para montagem de genoma é a análise de dados ou homogeneização de dados. Antes de se processar montagens é necessário converter os dados do sequenciamento para padrões utilizados pelo *software* de montagem. O formato FASTQ é comumente utilizado por esses *softwares*, uma vez que o mesmo apresenta as sequências de nucleotídeos e dados de qualidade de cada base (Ramos, 2011; Mariano, 2015). Essa qualidade de base é calculada de acordo com o índice probabilístico do algoritmo Phred (Tabela 3). Através do *software* FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) é possível visualizar a média de qualidade de base e assim, avaliar a qualidade do sequenciamento.

Tabela 3 - Probabilidade de erro por base pelo algoritmo Phred.
Fonte: Ramos (2011).

Qualidade Phred	Precisão em %	Probabilidade de Erro
10	90%	1 em 10
20	99%	1 em 100
30	99,9%	1 em 1000
40	99,99%	1 em 10000
50	99,999%	1 em 100000

A etapa de montagem *de novo* compreende a sobreposição de todas as sequências, a fim de formar sequências contínuas (*contigs*). *Software* de montagem *de novo* podem utilizar os modelos de algoritmos: (i) *overlap-layout-consensus*: Mira, Newbler; e (ii) grafo de bruijn: Minia. A última etapa é o processo de *scaffolding*, ou finalização de montagem, que pode ser realizada a partir de um genoma referência, como por exemplo, a partir do *software* CONTIGuator. Dentre as ferramentas para auxílio nos processos de montagens de genomas, pode-se citar a ferramenta SIMBA

(Mariano, 2015). SIMBA possui os *software* integrados: Mira v3.9, Mira v4.0, Minia e Newbler, e permite fazer montagens com qualquer um desses *software* através de uma interface simples.

1.5 Análises comparativas com ferramentas de Bioinformática

Após obtenção dos genomas, é importante iniciar análises que possam acrescentar informações que visam melhorar o conhecimento biológico dos organismos. Uma abordagem para se obter estas informações ocorre a partir de uma comparação entre genomas de gênero ou linhagens da mesma espécie.

Análises comparativas com ferramentas de bioinformática são importantes métodos para se analisar e selecionar possíveis alvos de estudos *in vivo e in vitro*. Uma das grandes vantagens da bioinformática é a possibilidade de se trabalhar com alto volume de dados, com custo menor quando comparado com os outros tipos de estudos e com um menor tempo de resposta. Por isso, atualmente, antes de se realizar estudos *in vivo* ou *in vitro* são realizadas as análises *in silico*. Dentre as diversas ferramentas de análise comparativa em bioinformática, destacam-se as análises de sintenia, MLST e filogenia.

1.5.1 Análises de sintenia

Análises de sintenia são essenciais para se visualizar similaridades entre os blocos gênicos de diferentes organismos. Mesmo em casos onde a ordem gênica não seja muito conservada ainda possível visualizar e a comparar pequenos blocos. Através dessas técnicas de visualização é possível comparar o genoma de diferentes organismos.

1.5.2 Análises Funcional Genômica

As análises funcionais genômica são importantes ferramentas para se caracterizar os processos e domínios dos genes. Pois através dessa caracterização é mais fácil compreender as funções genicas, a distribuição deles em cada genoma e para futuras comparações em diversos organismos.

1.5.3 Análises de MLST

O MLST mede a variação de nucleotídeos entre as sequências de DNA de vários genes *housekeeping* e logo depois as classifica de acordo com perfil de cada

alelo. O número de genes utilizados no MLST varia de acordo com espécie, e em *S. agalactiae* são sete: *adhP* (*Alcohol dehydrogenase*), *atr* (*Amino acid transporter*), *glcK* (*Glucose kinase*), *glnA* (*Glutamine synthetase*), *pheS* (*Phenylalanyl tRNA synthetase*), *sdhA* (*Serine dehydratase*) e *tkt* (*Transketolase*) com tamanho de sequência variando de 459 pb a 519 bp. Esses genes são usados para amostras de diversos hospedeiros e em todos os sorotipos. Cada gene tem um valor de alelo, e o somatório desses valores é o ST (*sequence type*), que é usado para caracterizar cada amostra (Jones, 2003). Esse valor de cada alelo é escolhido por ordem de descoberta. A combinação dos valores dos sete alelos é usado para gerar um *allelic profile* (AP), que é representado por um valor de ST. Logo após gerar o AP de cada amostra, são realizadas análises entre elas para detectar a proximidade, que pode ser observada pela comparação entre os valores dos AP (Urwin *et al.*, 2003).

Caso sejam encontrados grupos de STs que possuam relacionamentos entre si e possuam um ancestral comum, esses são agrupados em complexos clonais (CC) (Chan, 2011). Esse agrupamento ocorre através do algoritmo *BURST* (*Based Upon Related Sequences*), que é executado para identificar CCs e designar o genótipo central (Feil *et al.,* 2004). Os STs que não foram agrupados em nenhum CCs são chamados de *singletons* (Pavón *et al.,* 2009). A identificação dos CCs é uma importante ferramenta para as análises epidemiológicas (Urwin, 2003). Atualmente, o BD Pubmlst se armazena 536 ST, 751 MLST e 1273 amostras.

1.5.4 Filogenia molecular

A partir da década 70, com as técnicas de sequenciamento de DNA, houve uma grande evolução que gerou no refinamento e na melhoria da classificação dos grupos taxonômicos. Assim, começou a se utilizar as sequências do gene ribossomal *16S,* e o mesmo se tornou o padrão utilizado para construção da filogenia dos procariontes (Ludwig, 2001). Tentativas com outros genes essenciais foram realizadas como, por exemplo, utilizando o gene *rpoB*, o qual codifica a subunidade β da RNA polimerase bacteriana. Porém, observaram que a utilização de apenas um gene não resolveu os problemas taxonômicos, pois ainda eram encontrados organismos com características fenotípicas bem divergentes e que eram agrupadas numa mesma espécie (Gevers, 2005). Atualmente, para auxiliar essa classificação é utilizado um conjunto de genes essenciais da espécie, pois com

o maior número de genes é esperado que as distâncias entre os ramos da árvore filogenética sejam mais acuradas (Ludwig, 2001).

Para criação de uma árvore filogenética são necessários vários passos como: selecionar os genes alvos, organizar e alinhar as sequências, pesquisar o melhor modelo de inferência evolutiva e a escolher o método estatístico.

Os estudos filogenéticos são de elevada importância para se caracterizar, agrupar e comparar os novos objetos biológicos com os existentes afim de notar os padrões evolutivos.

2. Justificativa

Streptococcus agalactiae é uma bactéria de grande importância médica e veterinária, devido ao alto impacto econômico e social, associado ao elevado número de patogenicidade em diversos hospedeiros e à alta diversidade genética entre as linhagens. Atualmente, existem aproximadamente 300 genomas de *S. agalactiae* depositados no banco de dados público do NCBI, porém a grande maioria é incompleta, sendo apenas 16 genomas completos. Além disso, não existe nenhum genoma completo do sorotipo IA de humanos. Estudos com todos os sorotipos de GBS são de elevada importância para evitar crises epidemiológicas e auxiliar no desenvolvimento de drogas e vacinas mais eficazes. Outro fator importante é que a variação dos sorotipos ocorre de acordo com hospedeiro e com a região geográfica, e em diversos estudos realizados na Argentina, Brasil, Estados Unidos, Europa e Austrália é descrito que sorotipo IA de humanos é o mais predominante. Em estudos experimentais realizados anteriormente com a linhagem GBS85147 verificou-se que a mesma conseguiu sobreviver por mais 24 horas em macrófagos, por meio da ativação da NADPH-oxidase. Além disso, foi capaz de aderir e invadir as células endoteliais. Esses fatos demonstram o alto potencial patogênico da linhagem.

Para melhor compressão das funções e das estruturas da linhagem GBS85147 é de suma importância sequenciar, montar e anotar a mesma. Tais trabalhos podem auxiliar os futuros estudos comparativos entre os gêneros, espécies e linhagens.

3. Objetivos

3.1 - Objetivo Geral

Sequenciar, montar e anotar o genoma da *Streptococcus agalactiae* linhagem GBS85147 para auxiliar posteriores análises de genômica comparativa entre esta linhagem e as demais linhagens completas depositadas no banco de dados do NCBI.

3.2 - Objetivos Específicos

- Sequenciar o genoma de *S. agalactiae* GBS85147 utilizando a plataforma de sequenciamento de nova geração Ion PGM;
- Montar o genoma de *S. agalactiae* GBS85147;
- Predizer as regiões codificadoras e ribossomais;
- Realizar a anotação funcional de genes e os produtos gênicos;
- Depositar o genoma de *S. agalactiae* GBS85147 no banco de dados GenBank;
- Realizar análise genômica comparativa das sequências, estruturas e funções, entre os genomas completos de *S. agalactiae* depositadas no NCBI;
- Realizar comparações de conservação sintênica entre as linhagens;
- Realizar análises filogenéticas entre as linhagens.

4. Capítulos

4.1 Capítulo 1 - Artigo submetido à revista “Standards in Genomic Sciences”

Este capítulo é a versão em inglês do manuscrito ”*Complete Genome Sequence of Streptococcus agalactiae Strain GBS85147 Sorotype Ia Isolated From Human Oropharynx*”, submetido no dia 07 de maio de 2015 à publicação no periódico SIGS - Standards In Genomic Sciences (ISSN: 1944-3277).

SIGS é uma revista que publica pesquisas voltadas a todos os campos da biologia, possibilitando aos autores a publicação de análises de genomas. Esse periódico possui circulação internacional e fator de impacto 3,17.

Complete Genome Sequence of *Streptococcus agalactiae* Strain GBS85147 Sorotype Ia Isolated From Human Oropharynx

Edgar Lacerda de Aguiar[1], Diego César Batista Mariano[1], Marcus Vinícius Canário Viana[1], Leandro de Jesus Benevides[1], Flávia de Souza Rocha[1], Letícia de Castro Oliveira[1], Felipe Luiz Pereira[2], Siomar de Castro Soares[2], Fernanda Alves Dorella[2], Carlos Augusto Gomes Leal[2], Alex Fiorini de Carvalho[2], Camila Azevedo Antunes[3], Ana Luiza Mattos-Guaraldi[4], Prescilla Emy Nagao[3], Anne Cybele Pinto[1], Henrique César Pereira Figueiredo[2], Vasco Azevedo[1*]

[1] Laboratory of Cellular and Molecular Genetics (LGCM) - Federal University of Minas Gerais, Belo Horizonte, Brazil.

[2] National Reference Laboratory for Aquatic Animal Diseases (AQUACEN), Ministry of Fisheries and Aquaculture - Federal University of Minas Gerais, Belo Horizonte, Brazil.

[3] Laboratory of Molecular Biology and Physiology of *Streptococci* - State University of Rio de Janeiro, Rio de Janeiro, Brazil.

[4] Faculty of Medical Sciences - State University of Rio de Janeiro, Rio de Janeiro, Brazil

*Correspondence: Prof. Vasco Azevedo (vasco@icb.ufmg.br)

Keywords: *Streptococcus agalactiae*, Human Pathogenic Bacteria, Oropharynx, Complete Genome Sequence, Ion Torrent.

Abbreviations

bp - Base Pair

CDS - Coding Sequence

CPS - Capsular polysaccharides

DNA - Deoxyribonucleic acid

GBS - Group B *Streptococcus*

Kb - Kilobase

Mb – Megabase

MIGS - Minimum Information about a Genome Sequence

NCBI - National Center for Biotechnology Information

NT - Not Type

PGM - Personal Genome Machine

Abstract

Streptococcus agalactiae is a Gram-positive and a coccoid bacterial pathogen. The specie can cause diseases in humans, cattles, fishes and some other hosts. In humans, it is associated with neonatal sepsis, meningitis and early or late-onset diseases. The pathogen can also infect adults with underlying disease, particularly the elderly and immunocompromised ones. Because of the high medical and veterinary importance of this microorganism, it became necessary to start the genomic study by sequencing the genome of this pathogen. Here, we used Ion Torrent PGM platform with fragment library 200bp sequencing kit. The sequencing generated 578,082,183bp, distributed in 2,973,022 reads, resulting in a mean coverage depth of approximately of 246-fold and was assembled using Mira Assembler v3.9.18. The *S. agalactiae* strain GBS85147 is comprised of a circular chromosome with a final genome length of 1,996,151bp containing 1,925 coding sequences (CDS), from which 18 are rRNA genes, 63 tRNA genes, and 2 pseudogenes, with a G+C content of 35.48%.

Introduction

Streptococcus agalactiae is a worldwide distributed bacterial pathogen that causes diseases in humans and animals [52]. In humans, it is frequently associated with meningitis, neonatal sepsis and may also affect immunocompromised adults and elderly [2]. *S. agalactiae* is responsible for the most fatal bacterial infections in human newborns [3]. In fish, the pathogen causes meningoencephalitis and septicemia worldwide in both fresh water and salt-water species [4; 19]. Consumption of fish has been associated with an increased risk of *S. agalactiae* serotypes Ia and Ib colonization in people [44]. Due to high mortality rate, it is being responsible for large global economic losses for fish producers and fishermen [5]. *S. agalactiae* continues to be a major cause of subclinical mastitis in dairy cattle, which is the dominant health disorder affecting milk production within the dairy industry and is responsible for substantial financial losses to industry worldwide [6].

S. agalactiae is a bacterium of great medical and veterinary importance due to a high social and economic impact [1], together with the number of pathogenicity in different hosts [9]. The incidence of invasive infections unrelated to pregnancy in human adults and animals seems to be increasing worldwide [45], justifying an increase in the number of studies in the area. Since the 1990s, serotype V emerged in the United States as the most frequent GBS serotype causing invasive disease in nonpregnant adults [46]. Later, other serotypes, such as Ia and III have also been recognized worldwide as significant causes of invasive disease [47].

In the post genomic era, it became possible to perform comparative genomic studies among various strains of *S. agalactiae* to better understand the biological complexity of this organism. One such reason drove this study for genome sequencing, assembly and annotation of a new GBS85147 strain, serotype Ia, Sequence Type 103 (ST-103), isolated from the oropharynx of a female human patient in Rio de Janeiro, Brazil. In previous studies, GBS85147 strain was found to survive 24 hours inside macrophages with NADPH oxidase activation [48; 49] and it is also able to adhere and invade endothelial cells [50], showing the pathogenic potential of this strain.

Organism Information
Classification and Features

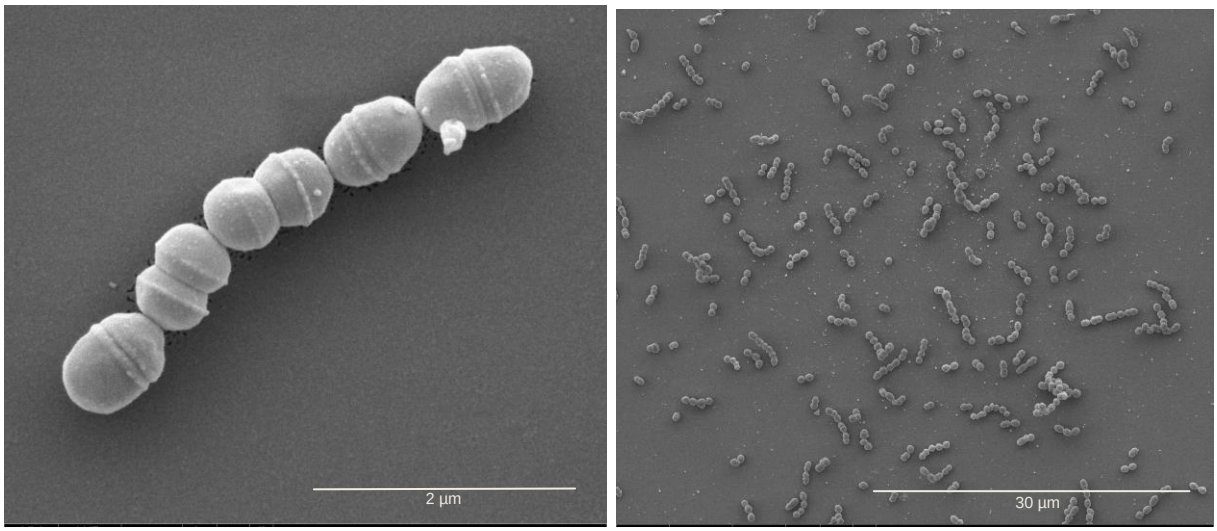


Figure 1. Scanning electron microscopy of *Streptococcus agalactiae* strain GBS85147 grown in liquid media after 8 hours. Scale bars, 2 and 30 µm.

S. agalactiae is a gram-positive non-sporulated bacterium that has a spherical shape with dimensions ranging from 0.2 to 1.0 microns [10]. In solid medium, *S. agalactiae* may form short chains or can make group in two double cocci. In liquid cells, the microorganism can be grouped to form long chains (Figure 1). The bacterium is a facultative anaerobe, catalase and oxidase negative, and is capable of performing fermentation of lactic sugars [11]. Lancefield identified the group B antigen, a peptidoglycan-anchored antigen (rhamnose, galactose, N-acetylglucosamine, and glucitol) that defines the *S. agalactiae* species [12, 51].

The capsular polysaccharide antigen (CPS) is used for the classification of *S. agalactiae* strains into serotypes [14]. The structure of CPS is determined by genes encoding enzymes responsible for its synthesis [13]. The serotype classification process is performed based on the capsular antigenic differences that are detected by PCR or immunodiffusion techniques [15]. Currently, ten serotypes have been described (Ia, Ib, II, III, IV, V, VI, VII, VIII, IX), being the serotype IX was identified in 2007 [16]. In some strains, the identification of serotypes is not possible due to the absence of the polysaccharide caused by a mutation in the capsular genes [17]. The high degree of variation in the capsular structure is related to the virulence of different strains of *S. agalactiae* [18]. Those variations in the capsular structure may also explain the infections in unusual hosts such as camels, dogs, horses, seals, chickens, dolphins, cats, hamsters, frogs, and monkeys [9].

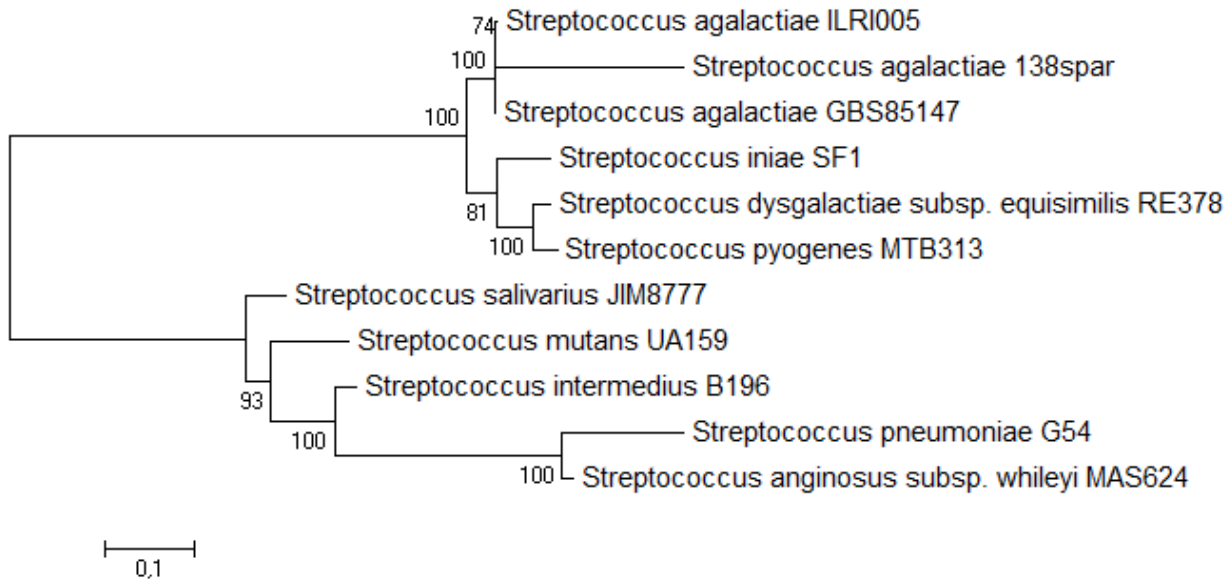


Figure 2. The phylogenetic tree was generated with S. agalactiae Gbs85147 strain and 10 other strains of Streptococcus genus, by concatenating the 16S rRNA and rpoB genes. The sequences were aligned automatically using ClustalW2 software [41] and after manual editing, the dataset contained sequences with 4810 characters. With the aligned sequences, we performed likelihood ratio test in Mega6 [42] software to determine the best model for evolutionary inference for the dataset in question. The evolutionary model General Time Reversible [43] + Gamma was considered the best choice. The phylogenetic test Bootstrap method with value 1000 replications and statistical method Maximum Likelihood were applied.

In order to observe an evolutionary distinction among different strains of GBS and other strains, a phylogenetic tree was created (Figure 2) using two genes: *rpoB* and 16S; to increase the accuracy of the Maximum Likelihood model.

The other taxonomic information can be viewed in Table 1.

Table 1. Classification and general features of *Streptococcus agalactiae* strain GBS85147 according to the MIGS recommendations [7]

MIGS ID	**Property**	**Term**	**Evidence code***
	Classification	Domain *Bacteria*	TAS[20]
		Phylum *Firmicutes*	TAS[21]
		Class *Bacilli*	TAS[22]
		Order *Lactobacillales*	TAS[23]
		Family *Streptococcaceae*	TAS[24]
		Genus *Streptococcus*	TAS[25]
		Specie *Streptococcus agalactiae*	TAS[26]
		Type strain GBS85147	NAS
		Sorotype Ia	NAS
	Gram stain	Positive	TAS[10]
	Cell shape	Coccus-shaped	TAS[11]
	Motility	Non-motile	TAS[10]
	Sporulation	Non-sporulating	TAS[10]
	Temperature range	Mesophile	TAS[27]
	Optimum	37ºC	IDA

	temperature		
	pH range; Optimum	5.4 – 9.4; 7.4	IDA
	Carbon source	Beta-glucoside, Cellobiose, Fructose, Glucose, Lactose, Mannose, Mannitol, N-acetylgalactosamine and Trehalose.	TAS [28]
MIGS-6	Habitat	Host-associated	TAS [1]
MIGS-6.3	Salinity	4.0% to 6.0%	TAS [27]
MIGS-22	Oxygen requirement	Facultative anaerobe	TAS[11]
MIGS-15	Biotic relationship	Symbiotic	TAS[1]
MIGS-14	Pathogenicity	Pathogen	TAS[48]
MIGS-4	Geographic location	Rio de Janeiro, Brazil	
MIGS-5	Sample collection time	Not reported	
MIGS-4.1	Latitude	Not reported	
MIGS-4.2	Longitude	Not reported	
MIGS-4.4	Altitude	Not reported	

* Evidence codes - IDA: Inferred from Direct Assay; TAS: Traceable Author Statement (i.e., a direct report exists in the literature); NAS: Non-traceable Author Statement (i.e., not directly observed for the living, isolated sample, but based on a generally accepted property for the species, or anecdotal evidence). These evidence codes are from the Gene Ontology project [8].

Genome sequencing information

Genome project history

S. agalactiae strain GBS85147 was isolated from human oropharynx in the Laboratory of Molecular Biology and Physiology of *Streptococci* in the city of Rio de Janeiro/RJ, Brazil. The genome was sequenced, assembled and annotated at the Laboratory of Cellular and Molecular Genetics (LGCM) in collaboration with the National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture (AQUACEN), both located at the Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil.

The project information and its association with MIGS version 2.0 compliance [7] are summarized in Table 2.

Table 2. Project information.

MIGS ID	**Property**	**Term**
MIGS-31	Finishing quality	Finished
MIGS-28	Libraries used	Fragment
MIGS-29	Sequencing platforms	Ion Torrent™ PGM System
MIGS-31.2	Fold coverage	246x
MIGS-30	Assemblers	Mira v3.9.18
MIGS-32	Gene calling method	FgenesB
	Locus tag	GBS85147
	Genbank ID	CP010319
	Genbank Date of Release	2015/05/01
	GOLD ID	

	BIOPROJECT	PRJNA263907
MIGS 13	Source Material Identifier	SAMN03108598
	Project relevance	Medical, Veterinary, Biotechnological

Growth conditions and genomic DNA preparation

S. agalactiae GBS85147 was obtained from the Streptococcus bacterial collection of the Laboratory of Molecular Biology and Physiology of *Streptococcus*. The sample was grown in brain-heart-infusion (BHI-HiMedia Laboratories Pvt. Ltda, India) under rotation for 48 hours at 37º C. The chromosomal DNA was extracted from 30 ml of culture. Briefly, the culture was centrifuged at 4°C and 4000rpm for 15 min. Cell pellets were re-suspended in 600µL Tris/EDTA/NaCl [10 mMTris/HCl (pH7.0), 10 mM EDTA (pH8.0), and 300 mMNaCl], placed in tubes containing cell lysis with Precellys® 2 times at rotations of 6500 rpm during 30 seconds. The DNA was purified with phenol/chloroform/isoamyl alcohol (25:24:1) and precipitate using ethanol/NaCl/glycogen (2.5vethanol, 10% NaCl and 1% glycogen) and re-suspended in 30µL MilliQ®. Finally, the DNA was stained using ethidium bromide and visualized in 1 % agarose gel. [29]

Genome sequencing and assembly

The genome sequencing was performed using a fragment library with Ion Torrent™ PGM System, 200bp sequencing kit. The sequencing produced a total of 578,082,183 bp, distributed in 2,973,022 reads, with average genome coverage depth of 246x and a Phred quality greater than or equal to 20 in 91.25% of bases. *De novo* assembly was performed using the software Mira v3.9.18 [30][31]. The assembly resulted 104 contigs, accounting for 2,032,890bp and a N50 of 104.996bp.

The contigs were ordered and oriented using the software CONTIGuatorv2 [32] with *S. agalactiae* GD201008 as a reference genome. From this processes 31 scaffolds. The remmaining gaps were closed via consensus sequences obtained by mapping the raw data against the reference-genome using the software CLC Workbench v7 [33] and BlastN [34], remmaining 2 pseudogenes.

Genome annotation

The structural gene prediction was performed using the software FGENESB [35] with *S. agalactiae* 09mas018883 [36] as the reference, resulting in 1,616 genes. The genome annotation was performed manually with Artemis [37] software, UniProt databases (http://www.uniprot.org/) and Interproscan 5 [38]. After manual annotation, 299 additional genes were created, with a total of 1,915 genes. For the prediction of rRNA and tRNA the software RNAmmer v1.2 [39] and tRNAscan-SE [40] were used, respectively.

Genome properties

The genome has one circular chromosome with 1,999,151pb and 35.48% of G+C content. A total of 1,998 genes including 1,915 protein coding, 18 rRNAs and 63 tRNAs genes. A circular map of the genome was generated using the CGView Comparison Tool (http://stothard.afns.ualberta.ca/downloads/CCT/), showed in Figure 3. The genome statistics are summarized in Tables 3-4.

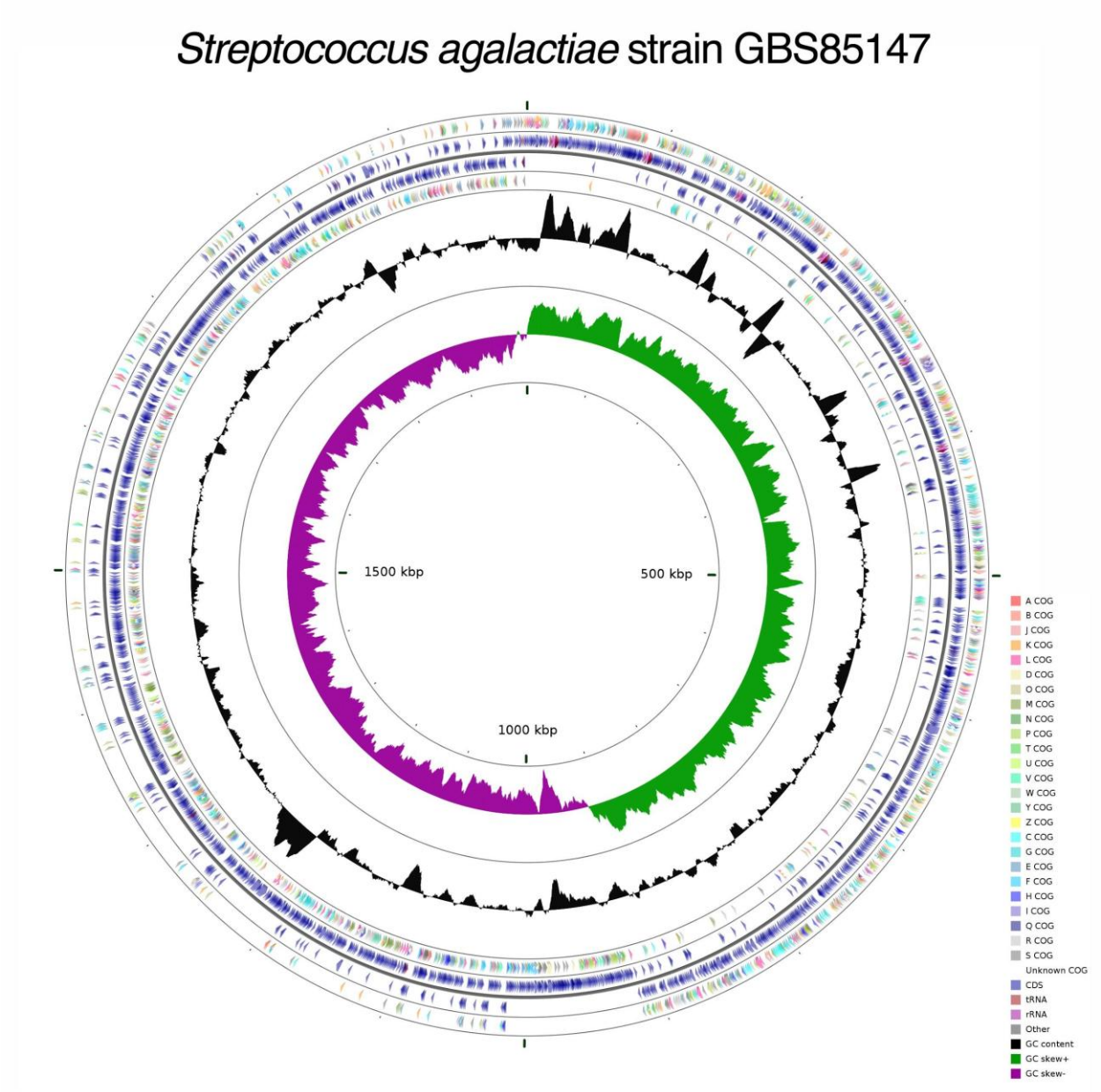


Figure 3. Map of Circular genome generated with CGview comparison tool.

In the outermost ring the genes identified by the COG, followed by Blue CDS, tRNAs in orange, rRNAs in pink, other RNAs in gray. In the intermediate ring GC content in black and the innermost ring represent the GC skew+ in green and GC skew- in purple.

Table 3. Genome statistics.

Attribute	**Value**	**% of Total**
Genome size (bp)	1,996,151	100
DNA coding (bp)	1,804,165	90.38
DNA G+C (bp)	708,380	35.48
DNA scaffolds	1	100
Total genes	1,998	100
Protein coding genes	1,915	95.84
RNA genes	81	4.05
Pseudogenes	2	0.1
Genes in internal clusters	26	1.30
Genes with function prediction	1,713	85.73
Genes assigned to COGs	1,564	78.27
Genes with Pfam domains	1,651	82.63
Genes with signal peptides	111	5.55
Genes with transmembrane helices	511	25.57
CRISPR repeats	1	

Table 4. Number of genes associated with general COG functional categories.

Code	Value	% of total	Description
J	144	6,63	Translation
A	0	0,00	RNA processing and modification
K	115	5,29	Transcription
L	97	4,46	Replication, recombination and repair
B	1	0,05	Chromatin structure and dynamics
D	20	0,92	Cell cycle control, mitosis and meiosis
V	36	1,66	Defense mechanisms
T	61	2,81	Signal transduction mechanisms
M	106	4,88	Cell wall/membrane biogenesis
N	7	0,32	Cell motility
U	30	1,38	Intracellular trafficking and secretion
O	58	2,67	Posttranslational modification, protein turnover, chaperones
C	52	2,39	Energy production and conversion
G	173	7,96	Carbohydrate transport and metabolism
E	155	7,13	Amino acid transport and metabolism
F	74	3,41	Nucleotide transport and metabolism
H	50	2,30	Coenzyme transport and metabolism
I	49	2,25	Lipid transport and metabolism
P	102	4,69	Inorganic ion transport and metabolism

Q	23	1,06	Secondary metabolites biosynthesis, transport and catabolism
R	219	10,08	General function prediction only
S	167	7,69	Function unknown
-	434	19,97	Not in COGs

Conclusion

The genome sequence of *S. agalactiae* GBS85147 obtained by Ion Torrent PGM platform with a coverage of approximately 246 folds was completely finished, manually annotated, its pseudogenes were manually curated and the resulting genome file was deposited in NCBI. Nowadays GBS85147 is the only serotype Ia strain isolated from a human host deposited in NCBI.

After manual annotation of CDSs, we identified the function of 1,713 (85.73%) genes and after frameshift manual curation only two pseudogenes remained. The final size of the genome is ~2Mb with a G + C content of 35.48%, values that are consistent with the genomes of other strains of the *S. agalactiae* species.

The functional analysis using the COG base showed that approximately 27% of the genes do not have any described function, which consists in the sum of genes with unknown functions (7.69%) and genes that were not found in the database (19.97%). This lack of information shows that a significant amount of genes requires further functional studies. In the phylogenetic tree, it is possible to observe a divergence between the strains of *S. agalactiae* GBS85147 isolated from human, *S. agalactiae* 138spar isolated from fish and *S. agalactiae* ILRI005 isolated from camel.

The structural, functional and evolutionary genomic information of the strain GBS85147 are of paramount importance for future comparative studies with other strains, considering the hosts, serotypes and virulence factors.

References

1.Mian GF, Godoy DT, Leal CA, Yuhara TY, et al. (2009). Aspects of the natural history and virulence of S. agalactiae infection in Nile tilapia. Vet. Microbiol. 136: 180-183.

2. Park SE, Jiang S and Wessels MR (2012). CsrRS and environmental pH regulate group B Streptococcus adherence to human epithelial cells and extracellular matrix. Infect. Immun. 80: 3975-3984.

3. Rajagopal L. Understanding the regulation of Group B Streptococcal virulence factors. Future Microbiol 2009; 4:201-221.

4. Chen M, Wang R, Li LP, Liang WW, et al. (2012). Screening vaccine candidate strains against *Streptococcus agalactiae* of tilapia based on PFGE genotype. Vaccine 30: 6088-6092.

5. Duremdez R, Al-Marzouk A, Qasem JA, Al-Harbi A, Gharabally H. Isolation of *Streptococcus agalactiae* from cultured silver pomfret, Pampusargenteus (Euphrasen), in Kuwait. J Fish Dis 2004; 27:307-310.

6.Richards VP, Lang P, Bitar PD, Lefebure T, et al. (2011). Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted *Streptococcus agalactiae*. Infect. Genet. Evol. 11: 1263-1275.

7. Field D, Garrity G, Gray T, et al. The minimum information about a genome sequence (MIGS) specification. Nat. Biotechnol. 2008;26:541–7.

8. Anon. Gene Ontology Consortium-Experimental EvidenceCodes.

9. Pereira UP, Mian GF, Oliveira ICM, Benchetrit LC, Costa GM, Figueiredo HCP. Genotyping of *Streptococcus agalactiae* strains isolated from fish, human and cattle and their virulence potential in Nile tilapia. Vet Microbiol 2010; 140:186-192.

10. Phares CR, Lynfield R, Farley MM, Mohle-Boetani J, Harrison LH, Petit S, et al. Epidemiology of invasive group B streptococcal disease in the United States, 1999-2005. JAMA. 2008;299(17):2056-65.

11. Schuchat A. Epidemiology of group B streptococcal disease in the United States: shifting paradigms. *Clin Microbiol Rev* 1998; 11:497-513.

12. Glaser, P., C. Rusniok, C. Buchrieser, F. Chevalier, L. Frangeul, T. Msadek, M. Zouine, E. Couvé, L. Lalioui, C. Poyart, P. Trieu-Cuot, and F. Kunst. 2002. Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. Molecular Microbiology. 45:1499-1513

13. Ramaswamy, S.V., P. Ferrieri, L. C. Madoff, A. E. Flores, N. Kumar, H. Tettelin, and L. C. Paoletti. 2006. Identification of a novel cps locus polymorphisms in nontypable group B Streptococcus. Journal of Medical Microbiology. 55:775-783.

14. Quentin, R., H. Huet, F. S. Wang, P. Geslin, A. Goudeau, and R. K. Selander. 1995. Characterization of *Streptococcus agalactiae* strains by multilocus enzyme genotype and serotype: identification of multiple virulent clone families that cause invasive neonatal disease. Journal of Clinical Microbiology. 33:2576-2581.

15. Verani JR, McGee L, Schrag SJ; Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention (CDC). Prevention of perinatal group B streptococcal disease - revised guidelines from CDC, 2010. MMWR Recomm Rep. 2010;59(RR-10):1-36.

16. Slotved HC, Kong F, Lambertsen L, Sauer S, Gilbert GL. Serotype IX, a proposed new Streptococcus agalactiae serotype. J Clin Microbiol. 2007;45(9):2929-36.

17. Kong, F., L. M. Lambertsen, H-C. Slotved, D. Ko, H. Wang, and G. L. Gilbert. 2008. Use of Phenotypic and Molecular Serotype Identification Methods To Characterize Previously Nonserotypeable Group B Streptococci. Journal of Clinical Microbiology. 46:2745-2750.

18. Winn WC Jr, Alen SD, Janda WW, Koneman EW, Procop GV, Schrenkenberger PC, et al. Koneman's color atlas and textbook of diagnostic microbiology. 6th ed. Philadelphia: Lippincott Williams & Wilkins; 2006. Chapter 13: Gram-Positive Cocci: Part II: Streptococci, Enterococci and the "Streptococcus-Like" Bacteria; p. 683-713.

19. Delannoy CM, Crumlish M, Fontaine MC, Pollock J, Foster G, Dagleish MP, Turnbull JF, Zadoks RN. Human *Streptococcus agalactiae* strains in aquatic mammals and fish. BMC Microbiol. 2013; 13: 41.

20. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the do-mains Archaea, Bacteria, and Eucarya. ProcNatlAcadSci USA 1990; 87:4576-4579.

21. Murray RGE. The Higher Taxa, or, a Place for Everything...? In: Holt JG (ed), Bergey's Manual of Systematic Bacteriology, First Edition, Volume 1, The Williams and Wilkins Co., Baltimore, 1984, p. 31-34.

22. Euzéby J. List of new names and new combina-tions previously effectively, but not validly, pub-lished. List no. 132. Int J SystEvolMicrobiol 2010; 60:469-472.

23. Ludwig W, Schleifer KH, Whitman WB. Order II. Lactobacillales ord. nov. In: De Vos P, Garrity G, Jones D, Krieg NR, Ludwig W, Rainey FA, Schleifer KH, Whitman WB (eds), Bergey's Manu-al of Systematic Bacteriology, Second Edition, Volume 3, Springer-Verlag, New York, 2009, p. 464.

24. Deibel RH, Seeley HW. Family II. Streptococcaceae. In: Buchanan RE, Gibbons NE (eds), Bergey's Manual of Determinative Bacteri-ology, Eighth Edition, The Williams and Wilkins Co., Baltimore, 1974, p. 490-515.

25. Rosenbach FJ. In: Bergmann JF (ed), Microorganismenbei den Wund-Infections-Krankheiten des Menschen., Wiesbaden, 1884, p. 1-122.

26. Lehmann KB, Neumann R. Atlas und Grundriss der Bakteriologie und Lehrbuch der speziellen bakteriologischen Diagnostik, First Edition, J.F. Lehmann, München, 1896, p. 1-448.

27. Whiley RA, Hardie JM. The Firmicutes. In: Garrity GM, Boone DR, Castenholz RW (eds), Bergey’s Manual of Systematic Bacteriology, Second Edi-tion, Volume 3, Springer, New York, 2001, p. 655-735.

28. Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, Zouine M, Couvé E, Lalioui L, Poyart C. Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease. MolMicrobiol 2002; 45:1499-1513.

29. Sambrook, J. Russel, D. W. 2001. Molecular Cloning. 3rd edition. 3 vol. Cold Spring Harbor Laboratory Press, New York.

30. Chevreux, B.; Wetter, T.; Suhai, S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, pp. 45-56. 1999.

31. Liu G, Zhang W, Lu C. Complete Genome Sequence of Streptococcus agalactiae GD201008-001, Isolated in China from Tilapia with Meningoencephalitis. Journal of Bacteriology 2012;194(23):6653. doi:10.1128/JB.01788-12.

32. Galardini M, Biondi EG, Bazzicalupo M, Mengoni A. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol. Med. 2011;6:11.

33. Anon. CLC Genomics Workbench. Available at: http://www.clcbio.com/products/clc-main-workbench/.

34. 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.

35. V. Solovyev, A Salamov (2011) Automatic Annotation of Microbial Genomes and Metagenomic Sequences. In Metagenomics and its Applications in Agriculture, Biomedicine and Environmental Studies (Ed. R.W. Li), Nova Science Publishers, p. 61-78

36. Zubair S, de Villiers EP, Fuxelius HH, et al. Genome Sequence of Streptococcus agalactiae Strain 09mas018883, Isolated from a Swedish Cow. Genome Announcements. 2013;1(4):e00456-13. doi:10.1128/genomeA.00456-13.

37. Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation. Bioinformatics 2000;16:944–945.

38. Jones, Philip. Binns, David. Chang, Hsin Yu, et al. InterProScan 5: genome-scale protein function classification 2014.

39. Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T Ussery DW RNammer: consistent annotation of rRNA genes in genomic sequences - Nucleic Acids Res. 2007 Apr 22.

40. Lowe, T.M. and Eddy, S.R. (1997) Nucleic Acids Res, 25: 955-964.

41. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wal-lace IM, Wilm A, Lopez R, et al. Clustal W and Clustal X version 2.0. Bioinformatics 2007; 23:2947-2948.

42. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. 2013;30:2725–9

43. Nei M, Kumar S. Molecular Evolution and Phylogenetics. New York: Oxford University Press; 2000:333.

44. Foxman B, Gillespie BW, Manning SD, Marrs CF: Risk factors for group B streptococcal colonization: potential for different transmission systems by capsular type. Ann Epidemiol 2007; 17:854–862.

45. Phares CR, Lynfield R, Farley MM, Mohle-Boetani J, Harrison LH, Petit S, Craig AS, Schaffner W, Zansky SM, Gershman K, Stefonek KR, Albanese BA, Zell ER, Schuchat A, Schrag SJ. Epidemiology of invasive group B streptococcal disease in the United States, 1999-2005. JAMA. 2008; 299:2056-65.

46. Skoff TH, Farley MM, Petit S, Craig AS, Schaffner W, Gershman K, Harrison LH, Lynfield R, Mohle-Boetani J, Zansky S, Albanese BA, Stefonek K, Zell ER, Jackson D, Thompson T, Schrag SJ. Increasing burden of invasive group B streptococcal disease in nonpregnant adults, 1990 –2007. Clin. Infect. Dis. 2009; 49:85–92.

47. Martins ER, Melo-Cristino J, Ramirez M. Dominance of serotype Ia among group B Streptococci causing invasive infections in nonpregnant adults in Portugal. J Clin Microbiol. 2012; 50:1219-27.

48. Teixeira CF, Azevedo NL, Carvalho TMU, Fuentes J, Nagao PE. Cytochemical study of *Streptococcus agalactiae* and macrophage interaction. Microscopy Res. Tech. 2001; 54:254–259.

49. Monteiro GCTS, Hirata Jr R, Andrade AFB, Mattos-Guaraldi AL, Nagao PE. Surface carbohydrates as recognition determinants in non-opsonic interactions and intracellular viability of group B *Streptococcus* strains in murine macrophages. Int. J. Mol. Med. 2004; 13:175-180.

50. Lione VOF, Santos GS, Hirata Jr R, Mattos-Guaraldi AL, Nagao PE. Involvement of intercellular adhesion molecule-1 and ß1 integrin in the internalization process to human endothelial cells of group B *Streptococcus* clinical isolates. Int. J. Mol. Med. 2005; 15: 153-157.

51. Caliot É, Dramsi S, Chapot-Chartier MP, Courtin P, Kulakauskas S, Péchoux C, Trieu-Cuot P, Mistou MY. Role of the Group B antigen of *Streptococcus agalactiae*: a peptidoglycan-anchored polysaccharide involved in cell wall biogenesis. PLoS Pathog. 2012;8:e1002756.

52. Farley MM. 2001. Group B streptococcal disease in nonpregnant adults. Clin. Infect. Dis. 33:556–561.

4.2 Capítulo 2 - Análises comparativas

Neste capítulo serão apresentadas as análises complementares do genoma de *S. agalactiae* linhagem GBS85147 em comparação com outros genomas completos da espécie depositados no NCBI. Tais análises não foram abordadas no capítulo anterior devido a restrições nos formatos requeridos pela revista. Serão apresentadas as análises essenciais para um melhor esclarecimento das diferenças genômicas entre GBS85147 e as demais linhagens.

4.2.1 Materiais e métodos

4.2.1.1 Obtenção de Dados de Referências

Para se realizar as análises comparativas genômicas é necessária à obtenção de dados no formato compatível com *software* de análise. Esses dados são obtidos em base de dados genômicos. Neste trabalho optou-se por utilizar a Base NCBI que é mantida pelo Centro de Biotecnologia do Governo Norte Americano, caracterizada por ser uma base de dados robusta e estável quando comparada às demais bases de dados.

O NCBI fornece os dados genômicos das linhagens submetidas a eles de duas maneiras: a primeira pelo FTP (*File Transfer Protocol*) através do endereço ftp://ftp.ncbi.nih.gov/genomes/. Em casos onde se faz necessária a obtenção de alto volume de dados e variados formatos, o uso FTP é de uso essencial para agilizar a obtenção dos dados. O segundo método é através do website http://www.ncbi.nlm.nih.gov/genome/, onde é possível efetuar procuras por espécies, gêneros e ate mesmo linhagens e outras taxonomias. Algumas espécies possuem páginas específicas onde é possível filtrar e visualizar dados como tamanho do genoma, quantidade de proteínas, data de submissão entre outros. Nessa mesma página é possível organizar as linhagens por *status* do genoma: completo ou incompleto, processo que facilita a procura por genomas completos, que é o principal objetivo na etapa de obtenção de dados.

4.2.1.2 Formato dos Dados

O formato dos dados obtidos é de extrema importância, pois alguns *software* só são compatíveis com um determinado formato. O FASTA apresenta sequências

após um cabeçalho iniciado pelo caractere “>”, podendo ser encontrando nas extensões FASTA, FAS, FA, FAA, FFN e FNA (NCBI-Blast, 2015). Um outro formato bastante utilizado é o *GenBank Flat File Format* (GBK), constituído de dados de sequências nucleicas, proteicas, anotações e características do genoma (GenBank Format, 2015).

4.2.1.3 Montagem

Para realizar a montagem do genoma de *S. agalactiae* GBS85147 foi utilizado o *software* SIMBA (http://ufmg-simba.sourceforge.net), e os *software* de montagem Mira v3.9, Mira 4.0, Minia v1.7 e Newbler v2.9. Os seguintes parâmetros foram utilizados:

Tabela 4 - Lista de Parâmetros utilizados no SIMBA.

Software	**Parâmetros**
Mira 3.9	project = t1 job = genome,denovo,accurate parameters = -GE:not=16 IONTOR_SETTINGS -AS:mrpc=100 readgroup = fragment technology = iontor data = *.fastq *.xml
Newbler 2.9	project = t2 assembler = newbler processors = 16 %cluster = 100\%
Minia 1.7	project = t3 assembler = minia k_mer = 31 length genome = 3000000
Mira 4.0	project = t4 job = genome,denovo,accurate parameters = -GE:not=16 -NW:cac=warn IONTOR_SETTINGS -AS:mrpc=100 readgroup = fragment technology = iontor data = *fastq *.xml

Para etapa de finalização foi utilizada a montagem realizada com Mira v3.9. A finalização da montagem foi realizada com CONTIGuator v2.7 usando como referência de *S. agalactiae* GD201008. Sobreposições entre contigs foram removidas usando *in-house scripts* e os últimos foram curados manualmente usando a extração da sequência consenso do mapeamento dos dados brutos sobre o genoma das referências *S. agalactiae* GD201008 e *S. agalactiae* 09mas018883.

4.2.1.4 Predição gênica

A etapa de predição gênica compreende a etapa de delimitação de regiões codificadoras (CDS). Nessa etapa também são preditos o posicionamento de RNAs estruturais. Para predição de CDS foi utilizado o *software* FgenesB, tendo como

entrada uma sequência fasta, e um genoma referência para que se possa realizar a predição. O FgenesB pode ser encontrado no endereço http://www.softberry.com/berry.phtml?topic=fgenesb&group=programs&subgroup=gfindb. Para a predição de tRNAs foi utilizado o tRNAscan-SE (http://lowelab.ucsc.edu/tRNAscan-SE/). No site é necessário apenas incluir a sequência fasta do genoma que o mesmo retorna os dados e localização dos tRNAs. O RNAmmer (http://www.cbs.dtu.dk/services/RNAmmer/) é um *web software* responsável por realizar as predições de rRNA. No site é incluído a sequencia fasta, seleciona-se o Reino a qual pertence o organismo e o mesmo retorna os dados e a localização dos rRNAs.

4.2.1.2 Anotação Genômica

Após a predição de regiões codificadoras e RNAs estruturais é necessário o processo de anotação. Para essa etapa foi utilizado o *software* Artemis (http://www.sanger.ac.uk/resources/software/artemis) e a base de dados do Uniprot (http://www.uniprot.org/), um banco de dados biológicos que contém informações sobre sequências de proteínas e RNAs de outros bancos de dados como, TrEMBL e Swiss-Prot. Para anotação, a sequência de cada gene deve ser submetida ao banco de dados e o mesmo utiliza a ferramenta BLAST (*Basic Local Alignment Search Tool*) (Altschul, 1990), o qual retorna uma lista com as sequências com maior similaridade com a sequencia de busca. O Interproscan 5 foi utilizado para reanalisar proteínas hipotéticas encontradas no genoma.

4.2.1.3 Curadoria de frameshift

Após a anotação gênica é importante realizar a curadoria dos *frameshifts.* A curadoria de *frameshift* é o processo de correção de inserções, deleções e substituições inseridas na etapa de montagem. Para essa etapa foi utilizado o *software* CLC Genomics Workbench v7 (http://www.clcbio.com/products/clc-genomics-workbench/), uma suíte de visualização, análises e tratamento de dados gerados por plataformas NGS, para visualização dos dados de cobertura de dados brutos e assim correção da sequência de DNA.

4.2.1.4 Predição Funcional Genômica com *Blast2GO*

Blast2GO é *software* de bioinformática para anotação funcional automática de dados de sequências de genes e proteínas. Através da ferramenta BLAST são comparadas as sequências alvo com as sequências depositadas, e em seguida, ocorre uma transferência da anotação funcional caracterizando as sequências alvo. A informação funcional é representada através do *Gene Ontology* (GO), um vocabulário controlado de atributos funcionais. O *software* Blast2Go pode ser encontrado no *website* *https://www.blast2go.com*. Foram realizadas análises com a ferramenta Blast2GO. Alguns testes também foram realizados com COG, e os resultados podem ser visualizados nos anexos.

4.2.1.5 Análises comparativas e de sintenia

Para etapa de comparação genômica foi realizado o download de todas as linhagens de *S. agalactiae* através da plataforma Web do NCBI no endereço http://www.ncbi.nlm.nih.gov/genome/genomegroups/186.

Em todas as análises desenvolvidas nesse trabalho foram utilizadas apenas os dados das linhagens completas, descartando-se as demais para não ocorrer nenhum viés por informações gênicas ausentes. Para análises de sintenia foram utilizados os *softwares* CONTIGuator, Mauve, Brig.

4.2.1.5.1 CONTIGuator

O CONTIGuator executa o mapeamento entre os contigs do genoma alvo contra um outro genoma referência e após esta etapa, gera um mapa com esses resultados em formato PDF. CONTIGuator é um *software* bastante utilizado na finalização de montagens por referência, entretanto seu gráfico de sintenia pode ser utilizado para visualizar diferenças sintênicas entre genomas. O *software* CONTIGuator pode ser encontrado no *website* http://contiguator.sourceforge.net/.

4.2.1.5.2 BRIG

O software BRIG (*BLAST Ring Image Generator*) foi usado para gerar visualização da estrutura circular de genomas completos. Os dados de entrada do BRIG são sequências genômicas em formato fasta (extensão fna), que foram

utilizadas para gerar um banco de dados por meio de comparações entre as sequências genômicas através da ferramenta *blastn* (BLAST de nucleotídeos). Foram utilizados os valores padrões indicados pelo software, entre 50% a 70% de similaridade entre as sequências. O programa permite escolher a disposição dos genomas, e permite a inserção de arquivos tabulares gerados por *software* externos. Ele tem como arquivo de saída uma figura circular a qual os genomas são dispostos em anéis. O *software* pode ser encontrado no *website* http://brig.sourceforge.net/.

4.2.1.5.3 Mauve

O *software* Mauve é utilizado para gerar análise de sintenia e o mesmo funciona com alinhador de múltiplas sequências genômicas. Os arquivos de entrada devem estar no formato fasta, contendo as sequências nucleicas de genomas completos ou incompletos. Nesse trabalho foram utilizados dois parâmetros para personalização do alinhamento: o padrão, quando se utiliza espécies diferentes, e o *progressiveMauve*, quando se utiliza organismos relacionados. Após os alinhamentos, os genomas são dispostos na forma de blocos gênicos coloridos indicando homologia entre determinadas trechos das sequências de DNA, o qual facilita a visualização e a comparação entre os blocos gênicos. O *software* pode ser encontrado no *website* http://darlinglab.org/mauve/mauve.html.

4.2.1.6 Predição de Ilhas Genômicas com GIPSy

GIPsy é um *software* de predição de ilhas genômicas capaz predizer ilhas patogênicas, ilhas de resistência, ilhas simbióticas, ilhas de metabolismo. O mesmo suporta arquivos do tipo GBK, e compara genoma não patogênico com um patogênico cruzando os dados esses dados ao DB. GIPsy pode ser encontrado em http://www.bioinformatics.org/groups/?group_id=1180.

4.2.1.7 Filogenia

A primeira etapa do processo da análises de filogenia é obtenção do genes alvos. Foi utilizado o gene ribossomal *16s* de *S. agalactiae* linhagem GBS85147como referência para buscar os demais genes *16s* pertencentes a outras linhagens utilizando a ferramenta *blastn* do NCBI. O mesmo processo ocorreu para

obtenção dos dados do gene *rpoB,* e no caso do *rpoB* + 16s, as sequências foram concatenadas.

A segunda etapa foi o alinhamento, onde criou-se um arquivo contendo as sequências nucleicas dos genes alvos de todas as linhagens selecionadas. A seguir, as sequências foram alinhadas e editadas com *software* ClustalW2 (Ludwig, 2001).

A terceira etapa foi a busca do melhor modelo evolutivo para o conjunto de dados utilizando o *likelihood ratio test* no *software* MEGA6 que pode ser encontrado no endereço http://www.megasoftware.net/ (Tamura, 2013). Nessa etapa ocorre o processo de teste com os 24 modelos de inferência evolutiva e suas sub variações no *dataset* alinhado.

Na quarta etapa, foi realizado o processo de construção de árvore filogenética. Após a predição do modelo evolutivo mais propício para *dataset* alinhado, no modo Construct *Maximum Likelihood* Tree, foi selecionado o modelo evolutivo com as suas variações, em seguida seleciona-se o *Bootstrap method* com valor de 1000 replicações e com método estatístico *Maximum Likelihood*.

Após a montagem da árvore é possível editar a visualização no software Jalview Desktop que pode ser encontrado no endereço: http://www.jalview.org/, assim como exportar a árvore para editar em outros softwares.

4.2.1.8 Análises de MLST

Para realizar as análises de MLST de S. *agalactiae* foi necessário localizar cada um dos sete genes, e incluí-los em arquivo no formato fasta e submete-los ao banco de dados de alelos dos MLST de *S. agalactiae* (*http://pubmlst.org/sagalactiae/*).

Com os valores de cada alelo dos sete genes, uma nova pesquisa foi realizada para descobrir o valor do ST. Após as análises de cada linhagem foi retornado o valor de ST e o valor de cada alelo. Com esses valores de ST é possível diferenciar ou agrupar as linhagens de maneira rápida, a um baixo custo e até mesmo compartilhar os resultados com outros pesquisadores.

4.2.1.9 Análises adicionais

Foram realizadas duas análises adicionais: de predição de profagos e predição de vias metabólicas, ambas podem ser encontradas no Anexo 1.

4.2.2 Resultados e discussões

4.2.2.1 Montagem com SIMBA

Foram realizadas montagens de genoma utilizando o gerenciador de *softwares* SIMBA e obtidos os seguintes resultados para cada *software* integrado ao SIMBA (Mariano, 2015).

Tabela 5 - Resultados dos Montadores com SIMBA.

Software	Contig	Tamanho (pb)	Min contig	Max contig	N50
Mira v3.9	104	2.032.890	332	231.496	104.996
Newbler	15	1.967.233	4.076	683.614	453.172
Minia	23.505	1.831.151	63	408	71
Mira v4.0	108	2.009.080	278	120.039	59.123

Foram realizadas quatro tentativas de montagem com quatro diferentes montadores. Minia apresentou elevado números de *contigs* e um tamanho de genoma inferior aos demais. Newbler apresentou o menor número de *contigs*, porém o tamanho do genoma foi aproximadamente 42 Kb menor que a montagem com Mira v4.0. Já resultados do Mira v3.9 e Mira v4.0 foram bastante similares em tamanho do genoma e número de *contigs*, entretanto a versão 3.9 apresentou um menor número de *contigs* que a versão 4.0, além de um maior tamanho de genoma. Assim, a montagem realizada com Mira v3.9 foi escolhida para etapa de finalização de montagem. Os resultados da finalização da montagem foram apresentados no capítulo anterior.

Assim, SIMBA se mostrou uma ferramenta simples e eficiente para realização de montagens de genomas, permitindo uma fácil comparação entre os resultados.

4.2.2.2 Predição Funcional Genômica

Blast2Go padroniza a representação dos genes e assim como seus produtos, subdividindo-os em três categorias:

- processo biológico – atividade biológica com o qual o gene/produto contribui;
- função molecular – atividade bioquímica do gene/produto;
- componente celular – região celular onde o gene/produto é ativo.

Foram realizadas todas as análises usando a linhagem GBS85147, porém destacou-se nesse estudo as análises de processos biológicos, a fim de se

identificar as atividades biológicas em que possuem maior participação no genoma. Para verificar estes detalhes em níveis diferentes dentro do processo biológico, selecionou-se os níveis dois e três para análise da sequencia genômica.

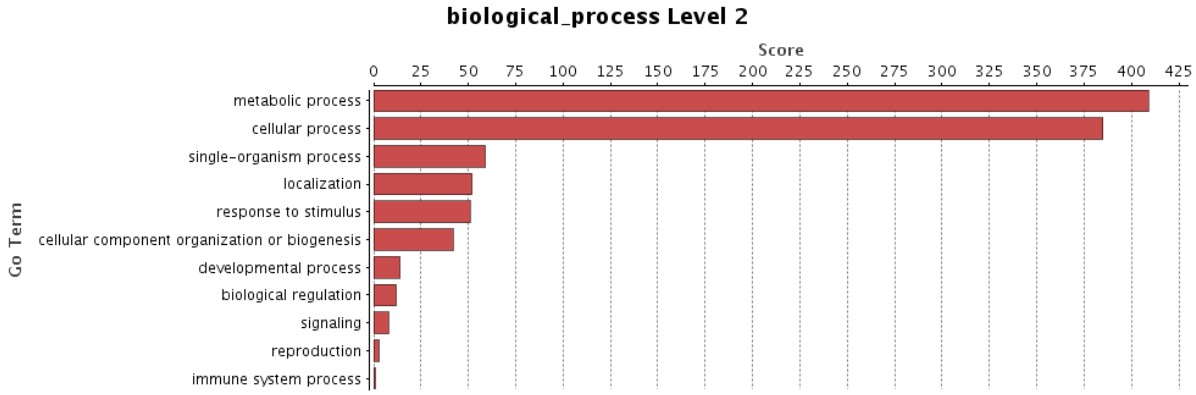


Figura 7 - Processo Biológicos Nível 2 com Linhagem GBS85147.

Na Figura 7 é possível visualizar os resultados do mapeamento dos processos biológicos nível 2, que no caso foram 11, onde o “processo metabólico” aparece em primeiro lugar com mais de 400 genes envolvidos, seguido por “processo celular”, com aproximadamente 385 genes. Observa-se que mais de 50% dos processos estão concentrados nesses dois itens, onde foram agrupados os genes de maior importância, que participam de processos metabólicos e celulares, pois são processos essenciais para sobrevivência da GBS8517.

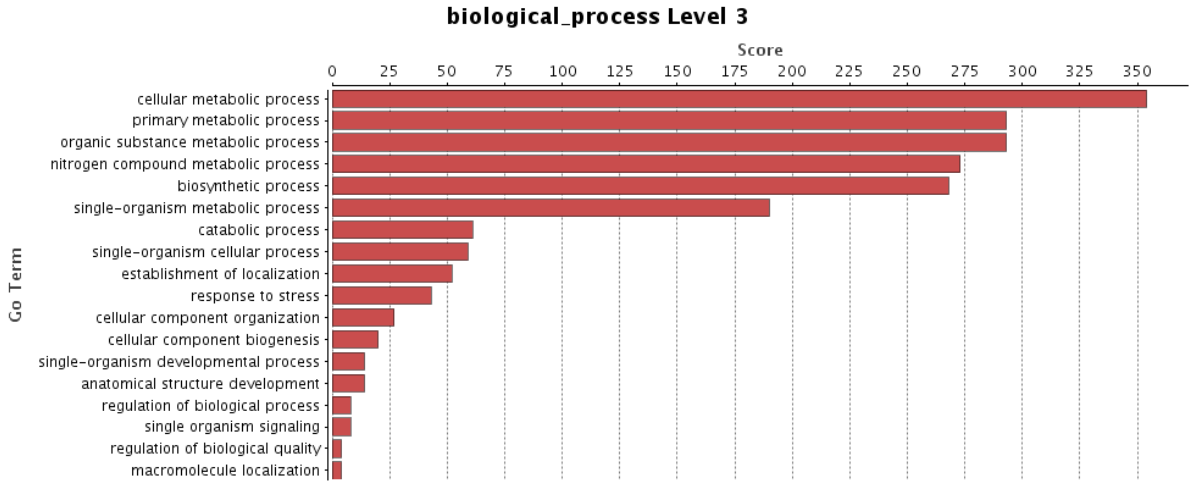


Figura 8 - Processo Biológicos Nível 3 com Linhagem GBS85147.

As análises com os processos biológicos em nível 3 demonstraram que houve um aumento no número processo, de 11 para 18, devido a especificação deles, onde ocorreu nova distribuição e concentração dos genes (Figura 8). Foram seis

processos com mais de 75% dos genes preditos. Além disso, os quatro maiores processos estão relacionados ao metabolismo.

Essa distribuição reforça a informação gerada no gráfico do nível 2, sobre a elevada importância dos processos metabólicos para GBS85147 já que se mantiveram como os processos mais encontrados em ambos os níveis.

Quando comparamos a distribuição dos processos biológicos do nível 2 e 3 é possível notar que nível 2 agrupou alguns processos e o nível 3 apresentou distribuição mais especifica dos processos gênicos. O nível 3 facilitou a visualização dos detalhes de processos biológicos em relação ao nível 2 onde estão mais agrupados.

O *software* Blast2Go se mostrou uma boa opção para mapeamento de funções e processos biológicos da linhagem GBS85147, o mesmo foi capaz de gerar uma extensa gama de visualizações. Porém um dos principais problemas do Blast2Go é o desempenho. Blast2Go apresenta uma extensa variedade de análises, fato que resulta num tempo de execução elevado para as análises de cada linhagem. Para as análises com linhagem GBS85147 foram gastos mais de uma semana num servidor robusto de alto desempenho (S.O. CentOS 6.4, processador 64 bit com 64 cores, 1 TB de memória RAM e 30 TB de armazenamento em disco). Estima-se que para testes com as análises comparativas envolvendo as 17 linhagens completas seria gasto um tempo de aproximadamente cinco meses. Esse alto tempo de execução dificulta a realização de análises comparativas entre as linhagens. Além disso, Blast2Go não é um *software* gratuito.

Uma análise adicional envolvendo predição dos processo gênicos com 3 linhagens (GBS85147, 138spar e ILRI005) através do BD do COG pode ser encontrada no Anexo 1, Item 1.1 Análises do COG. As análises envolvendo as linhagens com o BD do COG são motivadas principalmente pela agilidade em gerar os dados e gratuidade de acesso. Porém essas análises só retornam os dados em arquivo tabular, sendo necessário desenvolvimento de visualizações para apresentar os resultados.

4.2.2.3 Sintenia

Após obtenção do genoma completo, a próxima etapa foi a obtenção de resultados comparativos entre as linhagens (17 disponíveis no NCBI). Para esse fim, foram realizadas análises através dos *software* Mauve, CONTIGuator e BRIG.

Apesar das 17 linhagens completas do NCBI, a comparação individual de todas contra todas usando CONTIGuator e Mauve foi considerada inviável por apresentar um grande número de comparações. Por esse motivo, foi necessário uma nova abordagem selecionando apenas uma linhagem de cada hospedeiro. Foram utilizadas as linhagens Gbs85147 de Humano, 138spar de Peixe e ILRI005 de Camelo. Para uma correta comparação entre as três linhagens foi necessário o uso da fita reversa do genoma da linhagem 138spar, pois detectou-se que essa linhagem em comparação com outras linhagens de *S. agalactiae* possuía a fita reversa depositada no NCBI.

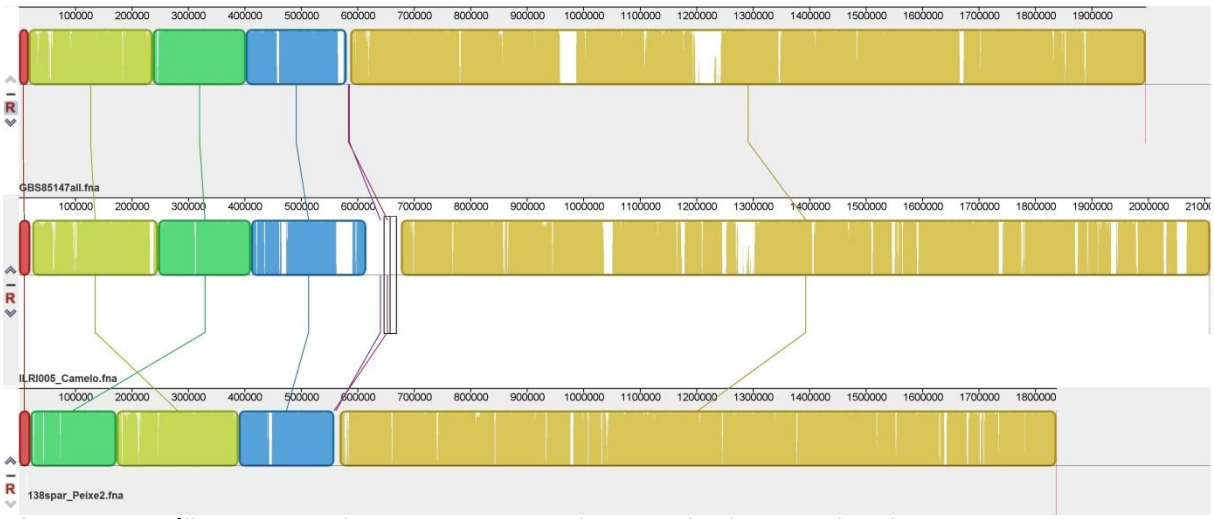


Figura 9 – Análise comparativa entre genomas de *S. agalactiae* usando a ferramenta Mauve. Nessa comparação cada linha representa as linhagens Gbs85147 de Humano, 138spar de Peixe e ILRI005 de Camelo. Nessa técnica de visualização, cada bloco de uma determinada cor representa um trecho sintênico em comparação com as outras. Trechos em branco nos blocos indicam regiões únicas.

No método de análise *progressiveMauve* do *software* Mauve são criados blocos gênicos e assim como a conservação entre esses blocos nas demais linhagens. Com a criação desses blocos é mais fácil visualizar deleções e inversões, o que não é possível notar no método análise padrão. Na linhagem 138spar, além do bloco final ser menor que as demais, ocorreu uma inversão entre o segundo e terceiro bloco que não ocorre nas outras duas linhagens. Quando comparado as

linhagens GBS85147 e ILRI005 a ordem dos blocos são mantidos, porém o tamanho não é conservado, além disso a figura indica a existência de diversas regiões únicas em cada linhagem (Figura 9).

Apesar de ser uma comparação com amostras reduzidas, é possível inferir que existem diferenças notáveis entre linhagens de diferentes hospedeiros. Esses resultados podem permitir futuras análises envolvendo outras combinações de linhagens, separando por sorotipo ou outras características morfológicas. Essas análises poderiam auxiliar à busca por blocos gênicos mais conservados que poderiam ser possíveis alvos de estudos para desenvolvimento de fármacos.

Além disso, foram realizadas novas análises com o software CONTIGuator para avaliar se as diferenças vistas com *software* Mauve se mantinham. O *software* CONTIGuator possui uma limitação na quantidade de linhagens a ser comparadas, onde não suporta múltiplas comparações, mas é útil para visualização de pequenas inversões e trechos repetitivos, que não são notáveis através do Mauve.

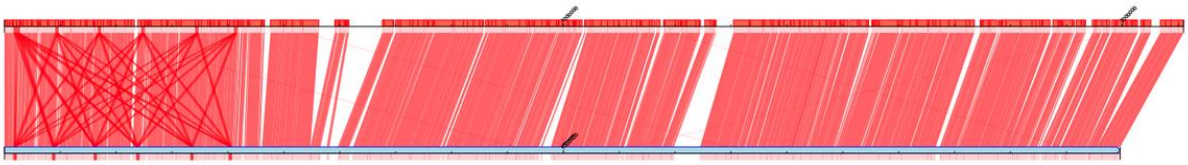


Figura 10 – Análise comparativa entre genomas de *S. agalactiae* GBS85147 (abaixo) com a linhagem *ILRI005* (acima) usando a ferramenta CONTIGuator.
Os blocos na cor vermelha que interligam os dois genomas indicam regiões sintênicas. Linhas em vermelho escuro interligando diferentes pontos indicam regiões de sequência repetitiva.

Na Figura 10 é possível visualizar as comparações entre as linhagens GBS85147 e ILRI005 pelo *software* CONTIguator. Na região inicial (à esquerda) é possível notar seis regiões a qual estão presentes os genes ribossomais. Essas regiões são bastante conservadas. É possível visualizar também muitas regiões únicas, indicando que as linhagens possuem diferenças nas sequências.

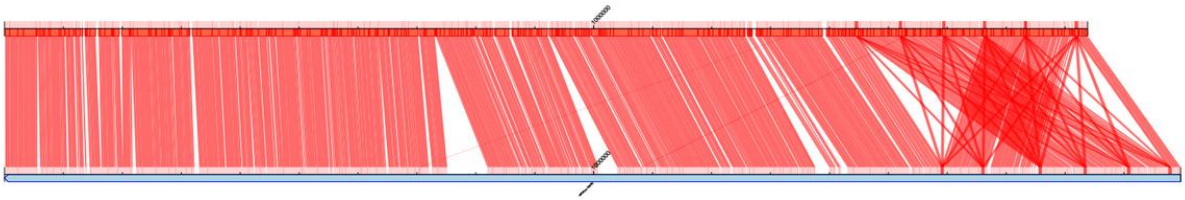


Figura 11 – Análise comparativa entre genomas de *S. agalactiae* GBS85147 (acima) com a linhagem 138spar (abaixo) usando a ferramenta CONTIGuator. Para uma melhor visualização foi necessário o uso da fita reversa de GBS85147.

Já na Figura 11, onde houve comparação entre as linhagens GBS85147 e 138spar, é possível notar que existem diversas regiões únicas em ambos os genomas, além de uma grande inversão próxima aos operons codificadores de RNA ribossomal.

Para as análises de sintenia realizadas com *software* BRIGs, foi possível avaliar todas as 17 linhagens de *S. agalactiae* que apresentam o genoma completo disponível no NCBI, o que nos permitiu obter uma visualização global das características existentes entre essas linhagens.

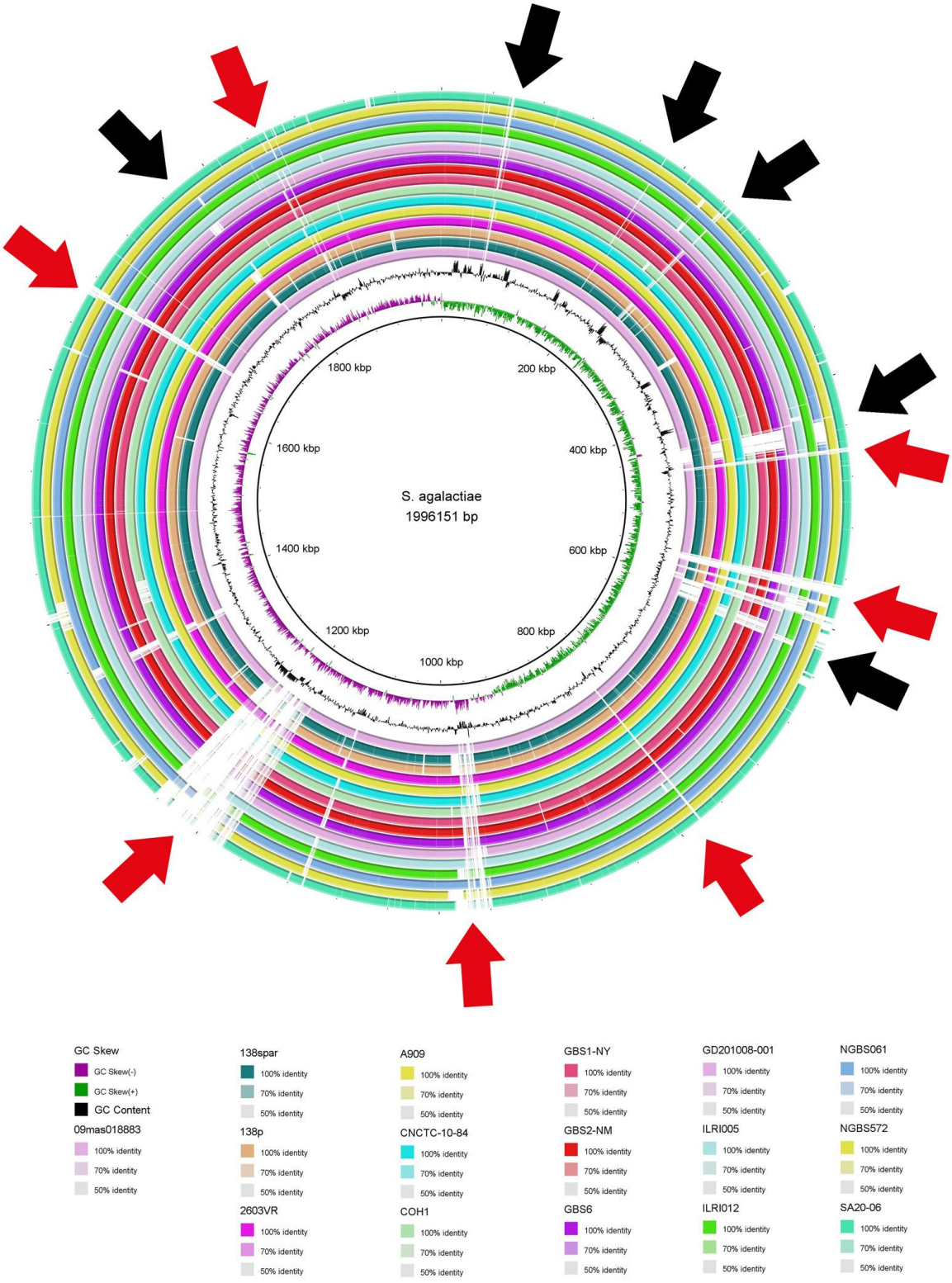


Figura 12 - Análise sintenia entre genomas de *S. agalactiae* usando a ferramenta BRIG.
O círculo central representa o genoma de *S. agalactiae* GBS85147. Os círculos ao redor representam as outras linhagens presentes no banco de dados do NCBI. As regiões indicadas pelas setas vermelhas representam sequências que são encontradas em GBS85147. Regiões indicadas pelas setas na cor preta representam regiões que aparecem em GBS85147 e em apenas algumas outras linhagens.

Pode-se observar na figura 12 que o anel mais interno é formado pela sequência da linhagem GBS85147, seguido pelo seu GC *Content* e CG *Skew*. As demais linhagens são organizadas em ordem alfabética. Cada anel pode apresentar diferentes tons de acordo com a porcentagem de identidade da região comparada com GBS85147. Através da análise do BRIG é possível notar sete regiões (setas vermelhas) que existem apenas na GBS85147. Análises envolvendo essas regiões seriam importantes para diferenciar genes únicos que poderiam ser usados para classificar os fatores de virulência desse sorotipo.

Foram encontradas seis regiões (setas pretas) que são compartilhadas entre Gbs85147 e algumas linhagens, mas não em todas.

Também foi possível ter visão geral das diferenças sintênicas entre as linhagens. Por GBS85147 ser a primeira linhagem de humano com sorotipo Ia era esperado encontrar genes e regiões genômicas diferentes.

A visualização em forma de anéis do *software* BRIG se mostrou a melhor opção para análises sintênicas com múltiplos genomas.

Cabe ressaltar que essas regiões únicas detectadas pelas análises de sintenia com Mauve, CONTIGuator e BRIG podem apresentar genes de grande importância para os processos de infecção em humanos. Assim, as análises de predição de ilhas genômicas podem auxiliar numa melhor compreensão do conteúdo dessas regiões.

4.2.2.4 Predição de Ilhas de Genômicas

As predições de ilhas genômicas foram realizadas com o software Gipsy. Os resultados do Gipsy foram divididos em três principais imagens de acordo com tipo de ilha predita, sendo: genômica, patogênica e resistência. Gipsy apresenta seus resultados em formato tabular, utilizando a ferramenta BRIG para visualizar o posicionamento das ilhas.

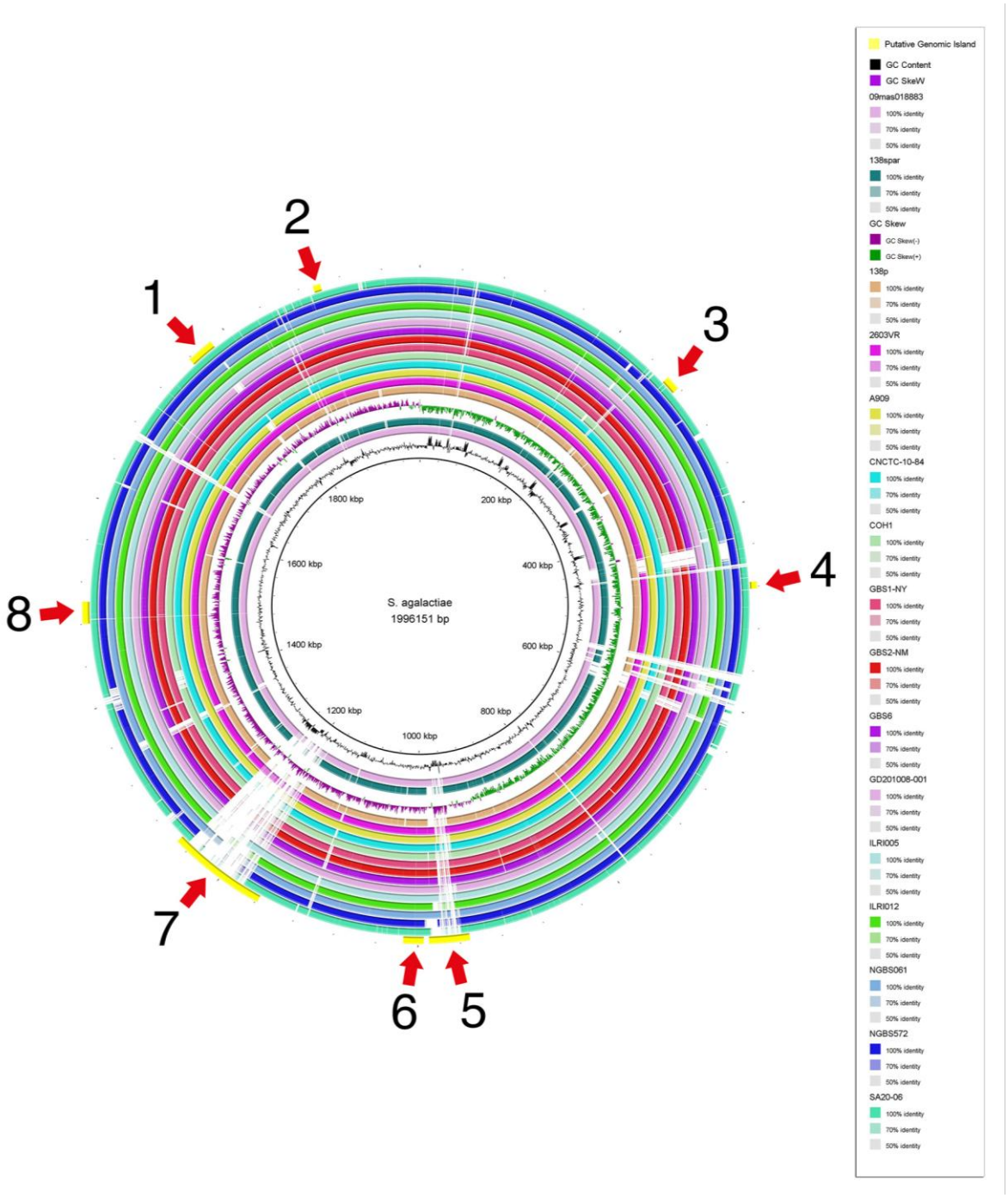


Figura 13 – Visualização de ilhas genômicas obtidas pela ferramenta Gipsy.

Na Figura 13 são apresentadas oito ilhas genômicas, porém de classificação desconhecida. Esse resultado indica que Gipsy reconheceu a região como uma provável ilha genômica, mas não conseguiu identifica-la. Entre as ilhas genômicas, destaca-se a ilha 5, que apresenta noventa genes existentes em GBS85147 e alguns também são encontrados na linhagem 2603VR. A ilha 7 destaca-se por apresentar a maior região única encontrada em GBS85147 (23 genes). Numa análise aprofundada dos genes presentes nessa ilha foi possível constatar a

existência de muitos genes com produtos classificados como proteínas hipotéticas. Conclui-se a importância de realizar estudos de genes presentes nessa região a fim de melhor caracterizar suas funções.

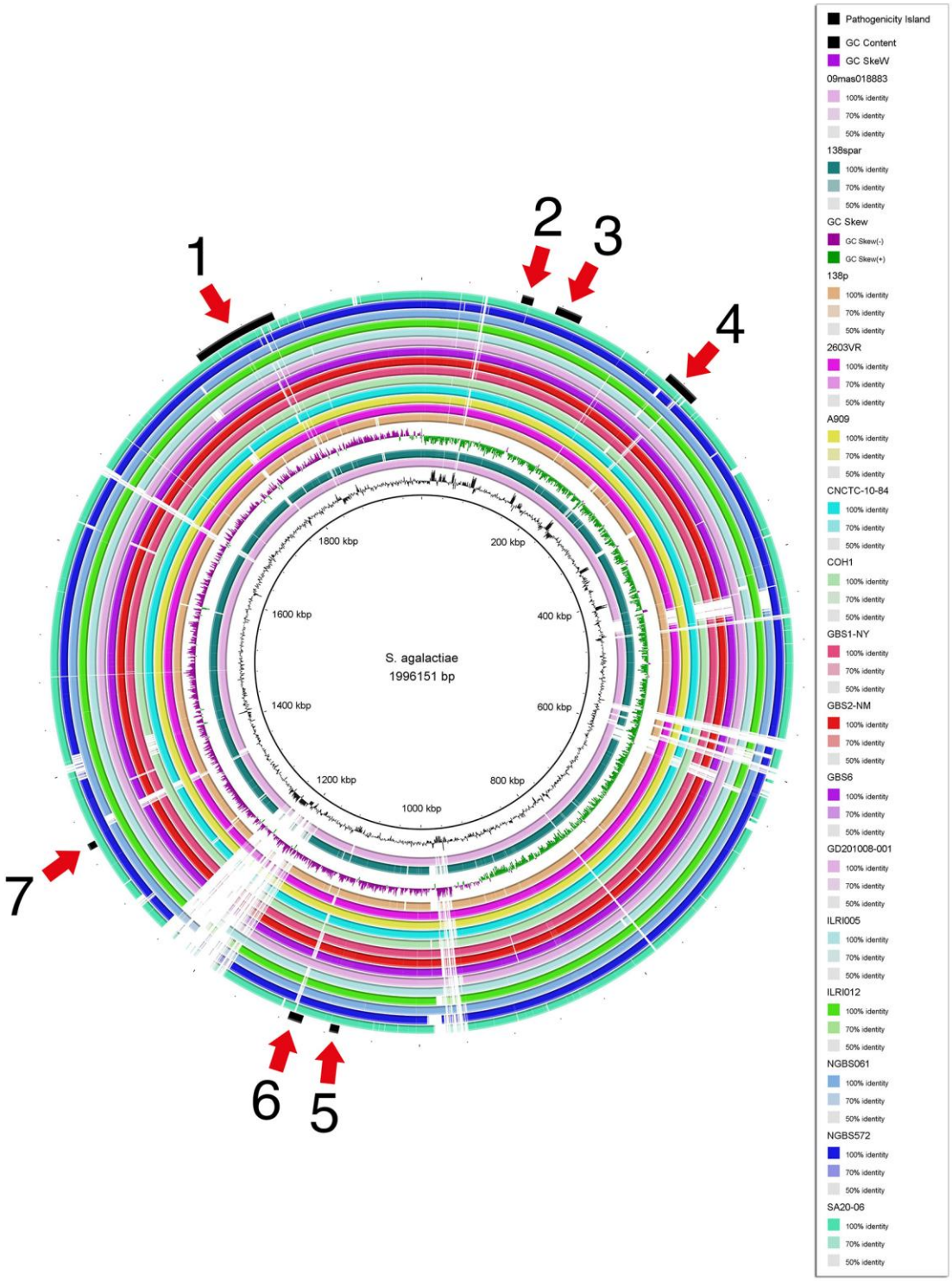


Figura 14 – Visualização de ilhas patogênicas obtidas pela ferramenta Gipsy.

A análise de ilhas patogênicas através do Gipsy demonstrou a existência de sete ilhas em todas as linhagens (Figura 14). Dentre essas, as ilhas 4 e 6 apresentam

pequenas regiões genômicas que existem apenas na GBS85147. A importância da análise de ilhas de patogenicidade, que podem ser consideradas como segmentos de DNA inseridos no cromossomo bacteriano, deve-se ao fato de serem capazes de atribuir uma variedade de características de virulência aos microorganismos que a possuem, como a capacidade de aderir e invadir o epitélio da célula hospedeira, produzir toxinas, captar ferro do meio ambiente e sintetizar dispositivos moleculares que permitam a translocação de moléculas efetoras para o interior da célula hospedeira (Vieira, 2009).

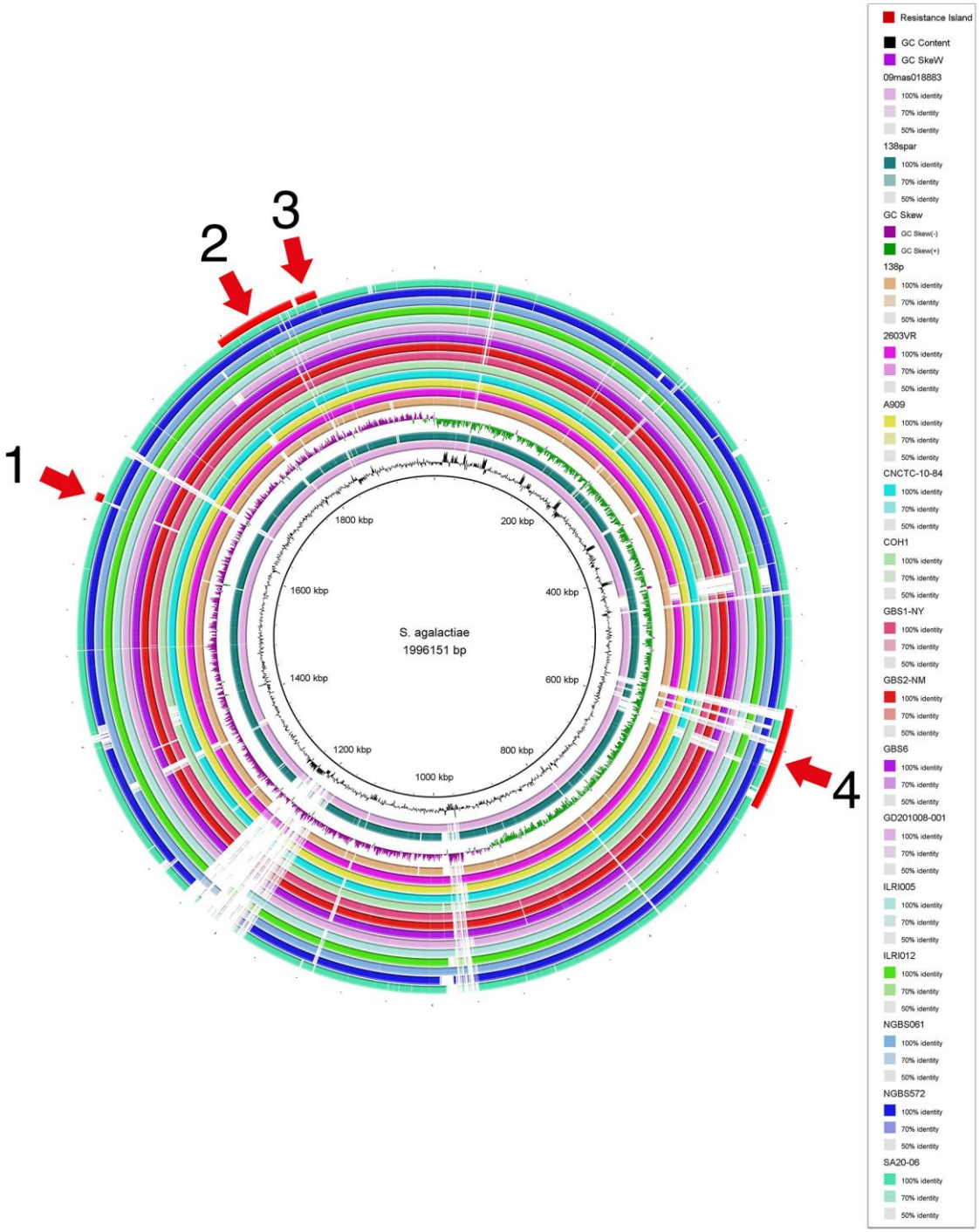


Figura 15 – Visualização de ilhas resistência obtidas pela ferramenta Gipsy.

Por fim, na predição das ilhas de resistência foram localizadas quatro ilhas, sendo que as ilhas 1 e 3 estão localizadas em todas as linhagens. Essas ilhas podem ser um indício que esses genes associados à resistência são bem conservados entre as espécie (Figura 15). A ilha 4 está presente completamente apenas em GBS85147. Grande parte da ilha 2 de resistência também foi predita como ilha de patogenicidade na análise anterior.

Os resultados apresentados nas análises de ilhas de patogenicidade e ilhas de resistência podem auxiliar no processo de desenvolvimento de fármacos. Além disso, foram selecionadas duas ilhas de patogenicidade, a 4 e a 6, como alvos de análises mais detalhadas.

A ilha 4 de patogenicidade é composta por seis genes, quatro hipotéticos e dois não conservados em todas as linhagens.

O primeiro é "*Streptokinase*", uma enzima que normalmente é secretada pelas espécies *Streptococcus*, possui alto potencial terapêutico para combater trombólise, e atualmente é utilizada no combate ao ataque cardíaco e embolia pulmonar (Sikri, 2007).

O segundo "*Glycine betaine/proline transport system*", faz parte do complexo de transporte de Glicina e betaína (Kegg, 2015). A Glicina está envolvida na formação de peptidoglicano na parede celular de bactérias Gram-positivas e também auxilia na fixação de estruturas celulares externas (Madigan, 2004).

Dos cinco genes presente na ilha 6 de patogenicidade foram encontrados três genes relacionados ao transporte de Ferro, um hipotético e transportador de ATP.

Nessa ilha destacam-se as proteínas responsáveis pelo complexo de transporte de Ferro. Durante a infecção, os patógenos necessitam captar uma diversidade de elementos a partir do organismo hospedeiro com o intuito de suprir suas necessidades nutricionais. Um deles é o ferro, pois é um elemento muito importante para o crescimento e metabolismo bacteriano, micronutriente necessário para a manutenção do organismo em níveis extremamente pequenos (Madigan, 2004). O ferro atua na manutenção do funcionamento de vias metabólicas essenciais, presentes nos organismos vivos em geral (Sheftel *et al.* 2010). Além disso, o ferro possui uma elevada importância nas reações celulares devido ao potencial redox desse íon, pois o mesmo tem a capacidade de alternar entre as formas Fe^{2+} e Fe^{3+}. Característica essas encontradas na catálise de várias reações

de oxi-redução, podendo proporcionar a manutenção nos sistemas biológicos (Ganz, 2006).

Ressalta-se que seriam necessárias análises mais profundas para verificar outros genes presentes nas demais ilhas e estudos mais aprofundados de regiões codificadores de proteínas anotadas como hipotéticas para identificá-las e elucidar seu papel na célula.

4.2.2.5 Filogenia

O objetivo das análises filogenéticas foi caracterizar as divergências evolutivas entre as linhagens e redistribuir as linhagens testando os métodos de filogenia clássicos.

4.2.2.5.1 16s

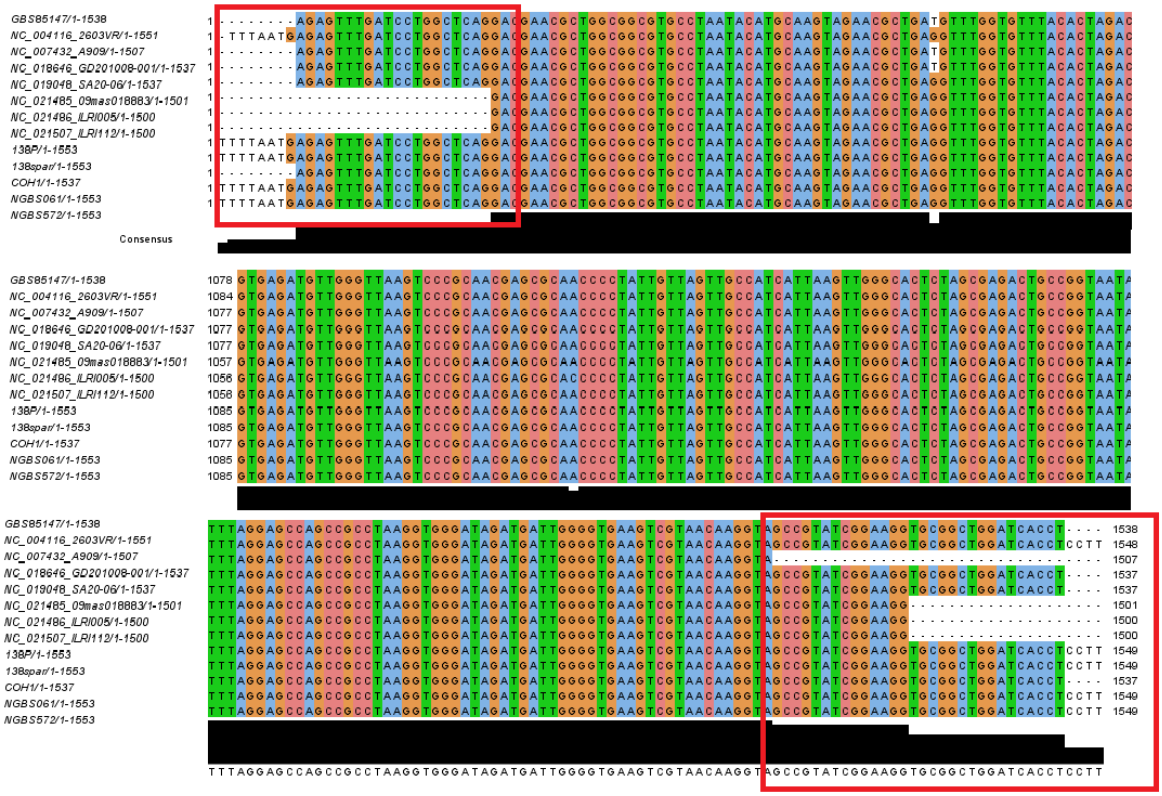


Figura 16 – Alinhamento dos genes 16s das 13 linhagens visualizado pelo *software* Jalview.

As linhagens de *S. agalactiae* utilizadas para montagem do *dataset* foram: 2603V/R, A909, GD201008001, SA2006, 09mas018883, ILRI112, ILRI005, 138P, COH1, 138spar, NGBS061 e NGBS572 e a GBS85147.

Após a montagem do *dataset* contendo as sequências do gene *16s* das 13 linhagens de *S. agalactiae,* foi utilizado o alinhador Clustal W2. Com as sequências alinhadas, elas foram importadas para o *software* Jalview para uma melhor visualização e edição do resultado do alinhamento (Figura 16). Devido a extensão do gene, foram exibidas apenas as extremidades que são as regiões de menor conservação (trechos em destaque). Os demais trechos se mostram muito conservados. Após o alinhamento, descobriu-se que três sequências pertencentes às linhagens eram exatamente iguais: 138spar, 138P e NGBS061.Porém, com a edição pós alinhamento, as sequências não se mostraram tão conservadas.

Jalview se mostrou um bom *software* para edição pós alinhamento e visualização das sequências alinhadas.

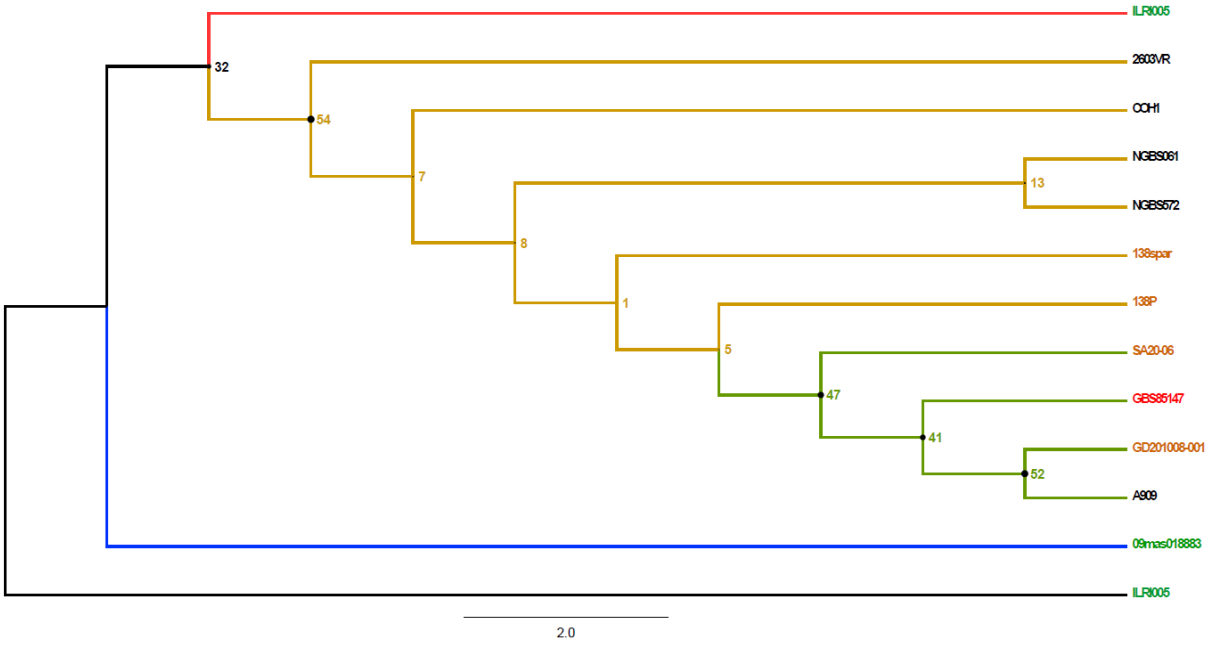


Figura 17 – Árvore filogenética com genes 16s das 13 linhagens.

Após a edição do alinhamento das sequências foi executado o *likelihood ratio test* no *software* MEGA6, a fim de determinar o melhor modelo de inferência evolutiva para as sequências em questão. O modelo evolutivo indicado foi o Felsenstein 1981 (F81) sem variações (Felsenstein, 1981). Logo, foi aplicado o teste filogenético B*ootstrap method* com valor 1000 de replicações e com método estatístico *Maximum likelihood*.

Após a geração da árvore filogenética, a mesma foi editada no MEGA6 para facilitar a visualização dos nodos (Figura 17). Os valores de *bootstrap* obtidos em todos os nós desta árvore ficaram abaixo do valor de 70%, mostrando que os ramos da árvore não tem boa sustentação.

Como essa primeira árvore não foi enraizada, não foi possível afirmar em que táxon foi iniciada a ramificação. Fez-se necessário então, a inclusão de um grupo externo na estrutura filogenética. Para a montagem do grupo externo foram selecionadas três espécies de *Streptococcus*: *Streptococcus dysgalactiae* subsp*. equisimilis* RE378, que faz parte do grupo C, *Streptococcus iniae* SF1, que não faz parte de nenhum grupo e *Streptococcus pyogenes* MTB313, que faz parte do grupo A. Na etapa de pesquisa dos genes do grupo externo através da ferramenta blastn do NCBI foram identificadas nos resultados, 4 novas linhagens completas de *S. Agalactiae*: GBS1-NY, GBS2-NM, GBS6 e CNCTC 10/84. Além das 3 linhagens externas foram incluídas essas 4 novas linhagens para montagem do novo *dataset*. Após a montagem do *dataset* as vinte sequências foram novamente alinhadas com Clustal W2. A inclusão das quatro linhagens *S. Agalactiae* modificou pouco o alinhamento, mas com a inclusão do grupo externo foi possível notar diferenças significativas nas sequências.

Após a inclusão desses novos genes e construção de um novo *dataset* e alinhamento com Clustal W2, ocorreu um novo processo de busca pelo modelo de inferência evolutiva no Mega6, sendo indicado o modelo Kimura 2 *Parameter* (K2P) com as variações +G e +I, utilizando *Bootstrap method* com 1000 replicações e com método estatístico *Maximum Likelihood.*

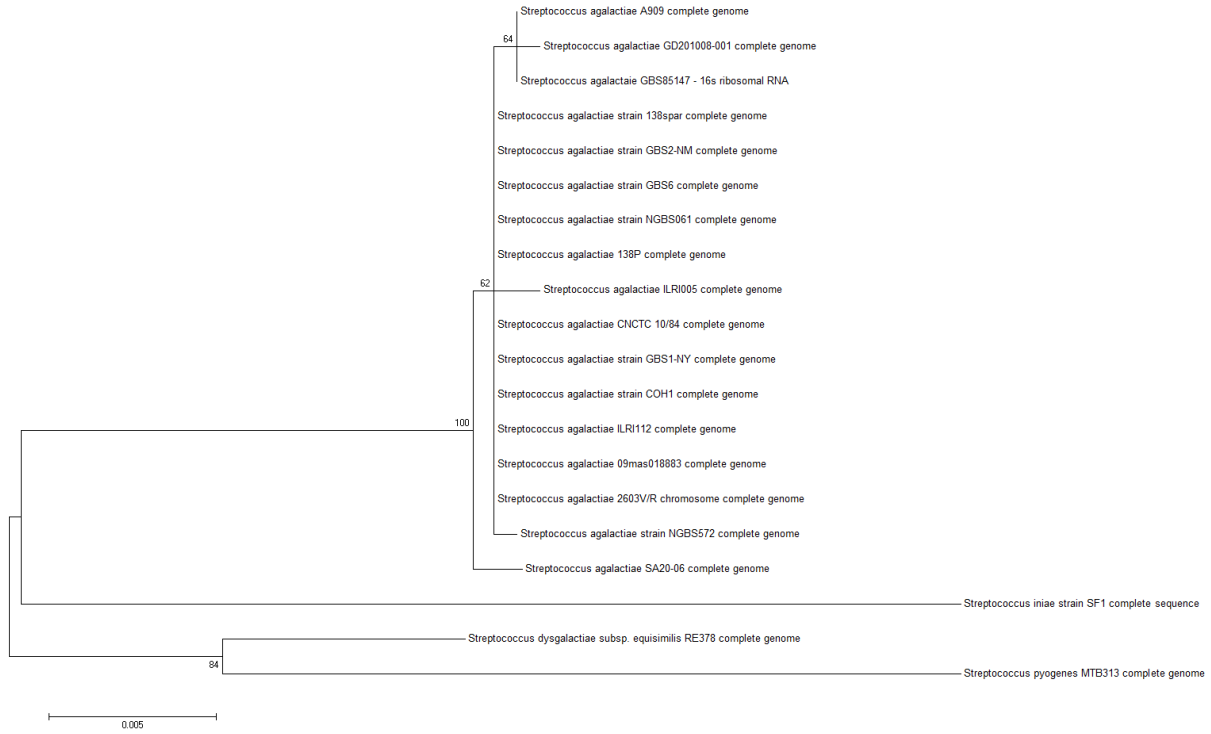


Figura 18 – Árvore filogenética com genes 16s das 17 linhagens, mais grupo externo.

Com a inclusão das novas linhagens e do grupo externo ocorreu um impacto nos *datasets*, pois a predição de modelo evolutivo foi alterado do F81 sem variações para K2P com variações +G e +I, isso confirma que inclusão dessas novas linhagens foi uma mudança significativa para construção do dataset e da árvore.

Através dessa segunda reconstrução filogenética foi possível corroborar informações da literatura de que o gene *16s* é eficiente para estudos filogenéticos, mas não para todos os casos, visto que quando há comparação entre grupos taxonômicos de mesma espécie, a acurácia da análise cai consideravelmente (Figura 18) (Cohan, 2004).

4.2.2.5.2 rpoB

Uma nova tentativa de reconstrução filogenética foi realizada, porém utilizando o gene *housekeeping rpoB*, uma vez que o mesmo evolui lentamente, assim como o gene *16s*, e tem maior resistência a transferência horizontal (Thompson, 2009).

Com o *dataset* montado com as sequências do gene *rpoB* dos 20 genomas, sendo as 17 *S. agalactiae* e 3 grupos externos, foi feito alinhamento com ClustalW2, usando os parâmetros padrão do ClustalW2.

Após alinhamento e a edição do *dataset* foi pesquisado o melhor modelo de inferência evolutiva no MEGA6, sendo indicado o GTR com a variação +G.

A árvore foi montada utilizando *Bootstrap method* com valor 1000 de replicações e com método estatístico Maximum Likelihood (Figura 19).

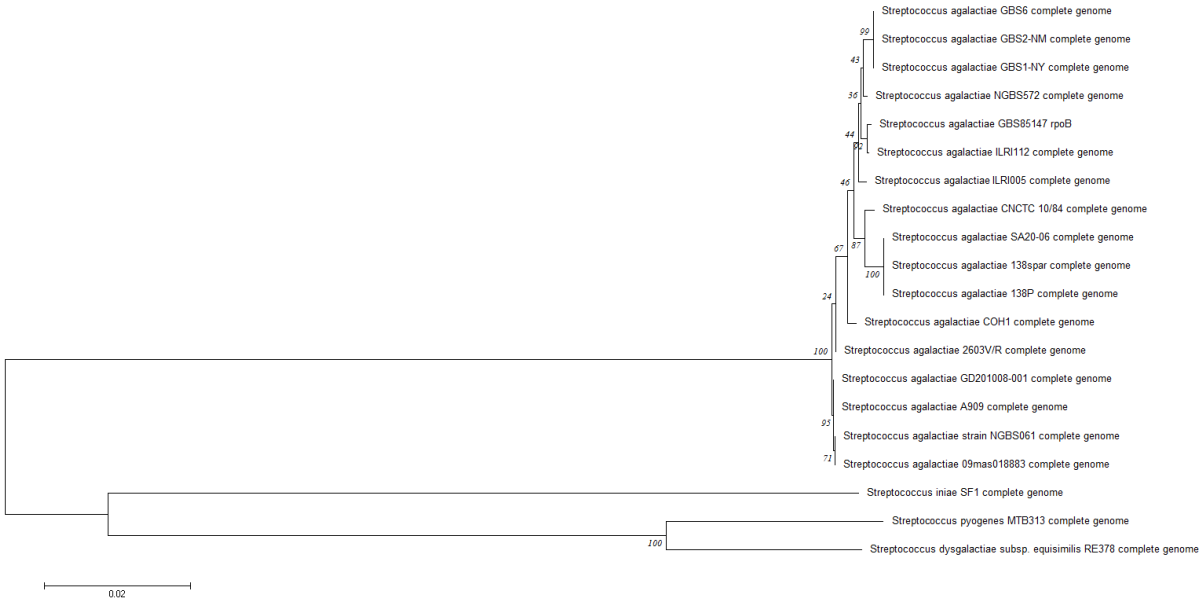


Figura 19 - Árvore filogenética com gene *rpoB* das 17 linhagens, mais grupo externo.

Na árvore com *rpoB* fica evidente que quando comparada com a árvore do 16s, a do *rpoB* teve um resultado mais promissor, pois conseguiu desagrupar melhor algumas linhagens. Porém, ainda existe clados com *bootstrap* abaixo do valor mínimo aceitável (Figura 19).

4.2.2.5.3 rpoB + 16s

Uma terceira abordagem envolvendo as sequências dos genes 16s e *rpoB* concatenadas, foi realizada para tentar montar uma árvore com melhor sustentação dos ramos.

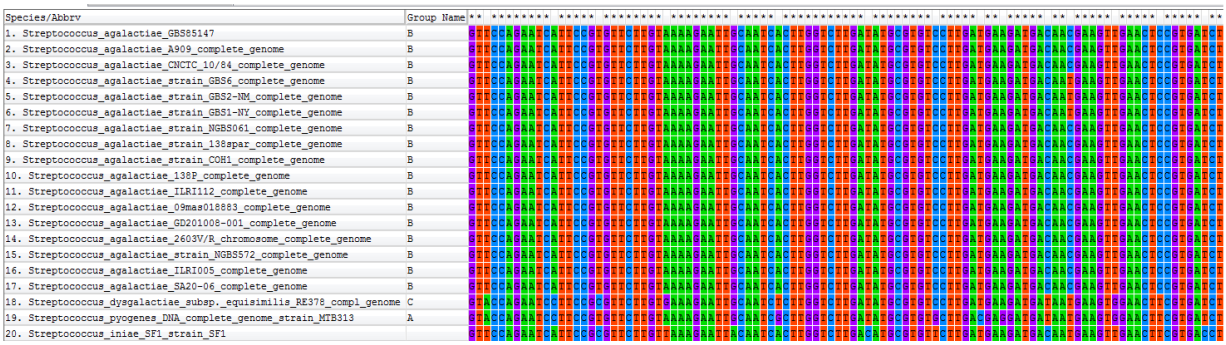


Figura 20 – Alinhamento dos genes *16s* e *rpoB* das 20 linhagens visualizado pelo Mega6.

Na etapa de montagem do *dataset* após a concatenação das sequências dos genes já foi possível notar uma redução na conservação que é bom para diferenciação evolutiva, como é possível visualizar na Figura 20.

O *dataset* foi alinhado com Clustal W2, seguido da busca pelo melhor modelo de inferência evolutiva no MEGA6, sendo indicado o modelo GTR com a variação +G.

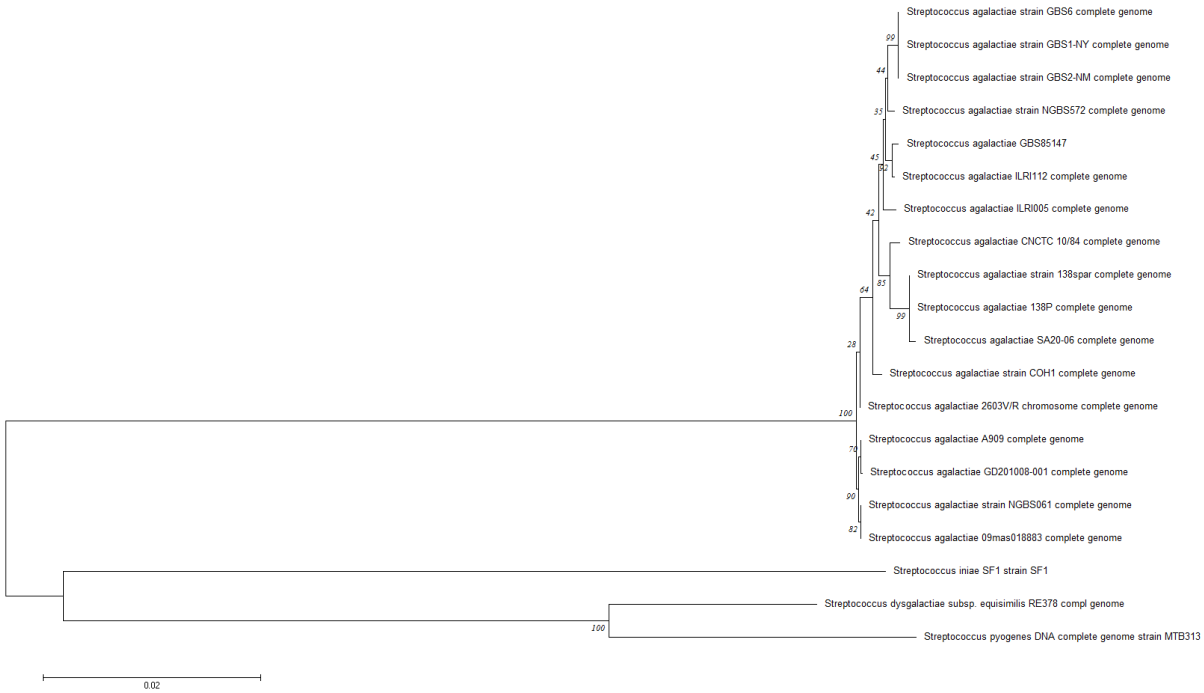


Figura 21 - Árvore filogenética com os gene *rpoB* e 16s das 17 Linhagens, mais grupo externo.

Para a montagem da árvore foi utilizado Bootstrap method com valor de 1000 replicações e com método estatístico Maximum Likelihood (Figura 21).

Na árvore com 16s + *rpoB* foi possível desagrupar grande parte das linhagens. Mas algumas ainda se mantiveram agrupadas, por exemplo as linhagens 138p e 138spar que são amostras de peixe. Assim como GBS6, GBS1-NY e a

GBS2-NM linhagens de humanos. E a mais divergente 09mas018883 que é uma linhagem de Boi e a linhagem NGB061 de humano. Aproximadamente 41% das linhagens se mantiveram agrupadas.

A abordagem de concatenar os dois genes se mostrou interessante, pois após a montagem da árvore alguns dos clados que possuíam um valor de bootstrap baixo passaram a ter um valor aceitável, isso aumenta o peso filogenético sobre os clados e também ocorreu uma melhor distinção entre as linhagens que não ocorreu com genes 16s e *rpoB* isoladamente (Cohan, 2004).

Para analisar eventos evolutivos seria necessária uma análise de inferência de redes filogenéticas e não somente de árvores. Para montagem do dataset poderia ser usado um maior conjunto de genes mais abrangente.

4.2.2.6 Análises de MLST

Como as análises com genes 16s e *rpoB* não foram suficientes para caracterizar corretamente todas as 17 linhagens, foi realizada uma nova abordagem utilizando MLST. Para a montagem do dataset foram utilizados os setes genes essenciais das linhagens de *S. agalactiae.* Objetivo dessa análise de MSLT foi verificar a possibilidade de diferenciar as linhagens agrupadas erroneamente.

Tabela 6 - Resultado MLST.

Linhagem	*adhP*	*atr*	*glcK*	*glnA*	*pheS*	*sdhA*	*tkt*
GBS85147 - ST 103	16	6	9	2	1	9	2
09mas018883 - ST 1	1	2	2	1	1	1	2
138spar - ST 261	54	31	25	4	17	26	19
138p - ST 261	54	31	25	4	17	26	19
2603VR - ST 110	1	3	2	2	1	2	9
A909 - ST 7	10	2	2	1	1	3	2
CNCTC-10-84 - ST 26	1	5	4	4	1	1	6
COH1 - ST 17	2	1	1	2	1	1	1
GBS1-NY - ST 22	13	1	1	3	3	1	1
GBS2-NM - ST 22	13	1	1	3	3	1	1
GBS6 - ST 22	13	1	1	3	3	1	1
GD201008-001 - ST 7	10	2	2	1	1	3	2
ILRI005 - ST 609	56	4	53	66	40	1	4
ILRI012 - ST 617	13	68	52	65	40	3	51

NGBS061 - ST 459	1	3	12	1	1	41	2
NGBS572 - ST 452	5	4	3	3	25	2	3
SA20-06 - ST 552	52	31	*26	4	17	26	19

Através das análises de MLST foi possível classificar ST de todas linhagens, única exceção foi na detecção do gene *glcK* da linhagem SA20-06, pois quando submetido com as demais sequências o *web software* não a localizou, mas no processo de busca individual foi identificado o valor de AP do mesmo. Para evitar esse tipo de erro, seria aconselhável busca por genes individualmente, porém em casos onde o número de linhagens é alto, seria aconselhável uso de script ou BD para controle e melhor organização desses dados.

Apesar da classificação de todas as linhagens, houve uma melhoria de 11% com separação das linhagens 09mas018883 e a linhagem NGB061 em dois ST diferentes, mesmo não ocorreu em análises filogenéticas anteriores.

Porém em 30% das linhagens ainda apresentaram o mesmo o mesmo problema das análises filogenéticas anteriores com genes 16s e *rpoB,* foram agrupadas no mesmo ST, a GBS1-NY, GBS2-NM e GBS6 com ST 22 e a 138spar e 138p com ST 261. Mesmo aumentando o número de dois para seis genes ainda não foi bastante para se caracterizar todas as divergências sintênicas existentes entre as linhagens como foi possível analisar anteriormente nesse trabalho.

Diante de tais fatos é necessário novas análises filogenéticas com genomas completos para se observar as diferenças e classificar corretamente as linhagens de *S. agalactiae*.

5. Conclusões finais

Neste trabalho foi apresentado o genoma completo de *S. agalactiae* linhagem GBS85147, a primeira linhagem do sorotipo Ia isolada de humano depositada no banco de dados do NCBI, sob número de acesso CP010319.

Os resultados da montagem, além de diversas análises do genoma, foram apresentados no capítulo 1 através do artigo submetido à revista SIGS.

Além disso, realizamos análises estruturais, que permitiram a identificação de regiões únicas em GBS85147, e que existem diferenças entre as linhagens quando são organizadas por hospedeiros.

As análises de ilhas genômicas demonstraram a existência de ilhas de patogenicidade que são encontradas em todas linhagens, fato que poderia facilitar a busca por alvos terapêuticos.

Através das análises filogenéticas utilizando os genes *16s, rpoB, 16s + rpoB, e* os genes do MLST, ainda não foi possível diferenciar todas as linhagens de *S. agalactiae*, pois um gene ou mesmo conjunto de genes, não foi suficiente para diferenciar corretamente as linhagens, pois diversas linhagens não foram desagrupadas apesar de possuírem uma significativa diferença sintênica.

5.1. Perspectivas para o futuro

O sequenciamento de novas linhagens de *S. agalactiae* de diferentes hospedeiros e sorotipos podem ser importantes para auxiliar estudos de patogenicidade na espécie. Auxiliar na enumeração de características genômicas, tais como funções e processos, além de auxiliar nos estudos etiológicos em diferentes hospedeiros. A elucidação desses aspectos é essencial para o desenvolvimento de estratégias terapêuticas, além de realizar novas comparações de genomas entre linhagens, que podem trazer informações úteis para minimizar os impactos socioeconômicos da bactéria na sociedade.

Pretendemos aperfeiçoar as análises de genes presentes nas ilhas genômicas. Análise de imunoinformática envolvendo a predição e seleção dos genes alvos nas ilhas genômicas, é importante para melhor compreensão dos mecanismos de patogenicidade da espécie. Realizar novas análises filogenéticas envolvendo genoma completo e análise de inferência de redes filogenéticas. Além de realizar um estudo de pangenoma de *S. agalactiae*, a fim de auxiliar nas pesquisas por alvos para o desenvolvimento de fármacos e na melhor caracterização da espécie.

6. Referências bibliográficas

Alexopoulos, C.J., Mims, C.W., Blackwell, M. **Introductory mycology.** 4.ed. New York: John Wiley & Sons, 1996.

Alvim M.J., Paciullo D.S.C., Carvalho M.M.et al**. Sistema de produção de leite com recria de novilhas em sistemas silvipatoris**. Embrapa Gado de Leite – Sistema de Produção nº 7. Versão eletrônica 2005. Disponível em: http://sistemasdeproducao.cnptia.embrapa.br/FontesHTML/Leite/LeiteRecriadeNovilhas/racas.htm. Acessado em: 20, Jan. 2014.

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.

Andrews, J. I., *et al*. **Group B streptococci causing neonatal bloodstream infection: antimicrobial susceptibility and serotyping resusts from sentry centers in the Western Hemisphere**. Am J Obstet Gynecol, 183(4): 859-862, 2000.

Baron, M. J., et al. **Anchors away: contribution of a glycolipid anchor to bacterial invasion of host cells**. The Journal of Clinical Investigation. United States, 115(9): 2325-2327, 2005.

Beitune, P., G. Duarte., and C. M. L. Maffei. 2005. **Colonization by Streptococcus agalactiae During Pregnancy: Maternal and Perinatal Prognosis**. The Brazilian Jounal of Infectious Diseases. 9:276-282.

Berry, E.A. et al. 2004. **Decision tree analysis to evaluate dry cow strategies under UK conditions**. J. Dairy Res. 71: 409-418.

Biedenbach D. J. et al,. **Antimicrobial susceptibility profile among beta-haemolytic Streptococcus spp. collected in the sentry Antimicrobial Surveillance Program - North America, 2001**. Diag Microl and Infect Disease, 46(4):291-294, 2003.

Bisno, A. L.; Rijn, I. van de. **Classification of streptococci**. In: Mandell, G. L.; Douglas; Bennett's. **Principles and practive of infectious diseases**. New York: Churchill Livingstone, 1995. p. 1784-1785.

Bonetta, L. **Genome sequencing in the fast lane**. Nat Methods, 3:141-147, 2006.

Bradley, A. **Bovine mastitis: an evolving disease**. Veterinary Journal, London, v. 164, n. 2, p. 116-128, Mar. 2002.

Brandão, Aloísio. **Bactérias essas velhas, perigosas e benefícios conhecidas**. Revista Pharmacia Brasileira , São Paulo, p.19, novembro. 2011.

Bras, J. **Antony van Leeuwenhoek: inventor do microscópio.** Patol. Med. Lab**.**, Rio de Janeiro, v. 45, n. 2, Apr. 2009 . Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1676-24442009000200001&lng=en&nrm=iso>. Acesso em: 16 Mar. 2015.

Broker, G. and B. Spellerberg. **Surface proteins of Streptococcus agalactiae and horizontal gene transfer**. International Journal of Medical Microbiology. 294:169-175, 2004.

Carneiro, A. V.; Stock, L. A.; Oliveira, V. M.; Zoccal, R.; Carvalho, G. R.; Martins, P.C.; Yamaguchi, L.C.T. 2004. **Mastite clinica: prevalência e custo de tratamento em rebanho leiteiro**. Disponível em: <http://www.cileite. com.br/sites/default/files/mastite_clinica_prevalencia_e_custo_de_tratamento_em_rebanho_leiteiro.pdf>. Acesso em: 20 de abril de 2015.

Carneiro, D. M. V. F. et al. **Imunidade inata da glândula mamária bovina: resposta à infecção**. Ciência Rural, Santa Maria, v. 39, n. 6, p. 1934-1943, Nov. 2009.

Centers for Disease Control and Prevention (CDC). **Prevention of Perinatal Group B Streptococcal Disease: Revised Guidelines from CDC**, 2010 November 19, 2010/59 (RR10); 1-32

Chan, M. S.; JOLLEY, K. **Streptococcus agalactiae (group B streptococcus GBS) MLST Database**. Disponível em: <http://pubmlst.org/sagalactiae/>. Acesso em: 1 jul. 2011.

Chen M., et al. (2012). **Screening vaccine candidate strains against *Streptococcus agalactiae* of tilapia based on PFGE genotype**. Vaccine 30: 6088-6092.

Cieslewicz, M. J. et al. **Structural and genetic diversity of group B Streptococcus Capsular Polysaccharides.** Infection and Immunity, Washington, v.73, n. 5, p. 3096-3103, 2005.

Cohan F. **Concepts of bacterial biodiversity for the age of genomics**. Chapter 11, pages 175–194. Springer-Verlag New York, LLC, 2004.

Creti, R., F. Fabretti, G. Orefici, and C. von Hunolstein. **Multiplex PCR Assay for direct Identification of Group B Streptococcal Alpha-Protein-Like Protein Genes**. Journal of clinical microbiology. 42:1326-1329, 2004.

D'Oliveira R. E. et al,. **Susceptibility to antimicrobials and mechanisms of erythromycin resistance in clinical isolates of Streptococcus agalactiae from Rio de Janeiro, Brazil**. J Med Microbiol, 52: 1029-1030, 2003.

De Azevedo J. C. S., et al. **Prevalence and mechanisms of macrolide resistance in invasive and noninvasive group B Streptococcus isolates from Ontario, Canada**. Antimicrob Agents Chemother, 45: 3504-08, 2001.

Dechen, T. C. et al. **Correlates of vaginal colonization with group B streptococci among pregnant women**. Journal of Global Infectious Diseases. v.2, n.3, p.236-241, Aug. 2010.

Delannoy CM, Crumlish M, Fontaine MC, et al. **Human Streptococcus agalactiae strains in aquatic mammals and fish**. BMC Microbiology. 2013;13:41.

Del Pozo, J. S. G. et al. **Vertebral osteomyelitis caused by *Streptococcus agalactiae*.** Journal of Infection, London, v. 41, n. 1, p. 84-90, July 2000.

Diário Oficial do Brasil. **Instrução Normativa n° 62, de 29 de dezembro de 2011. Estabelece o regulamento fixar os requisitos mínimos que devem ser observados para a produção, a identidade e a qualidade do leite**. Diário Oficial da Republica Federativa do Brasil, Poder Executivo, Brasília, DF, 30 dez. 2011. Seção 1, p. 6. Acesso em: 20 Abril. 2015

Dogan, B. et al. **Distribution of serotypes and antimicrobial resistance genes among Streptococcus agalactiae isolates from bovine and human hosts**. Journal of Clinical Microbiology, Washington, v. 43, n. 2, p. 5899–5906, 2005.

Doran, K. S. et al. **Molecular pathogenesis of neonatal group B streptococcal infection: no longer in its infancy**. Molecular Microbiology, Salem, v. 54, n. 1, p. 23-31, Oct. 2004.

Duremdez R, Al-Marzouk A, Qasem JA, Al-Harbi A, Gharabally H. **Isolation of *Streptococcus agalactiae* from cultured silver pomfret, Pampusargenteus (Euphrasen), in Kuwait**. J Fish Dis 2004; 27:307-310.

El-Metwally, Sara. et al. **Next-Generation Sequence Assembly: Four Stages of Data Processing and Computational Challenges**. PLOS Computational Biology, 2013.

Embrapa Portal - Publicações 2012 - **Banco de dados**. Disponível em: <http://www.cpatc.embrapa.br/publicacoes_2012/doc_170.pdf>.
Acessado em: 20, Abril. 2015.

Evans, A. C. 1936. **Studies on hemolytic streptococci. Methods of classification**. Journal Bacteriol 31:423-37.

Evans, J.J.; Wiedenmayer, A.A.; Klesius, P.H**. A transport system for maintenance of viability of Acinetobacter calcoaceticus, Streptococcus iniae, and Streptococcus agalactiae over varying time periods**. Bull. Eur. Assoc. Fish Pathol., v.22, p. 238-246, 2002a.

Evans, J.J.; Klesius, P.H.; Gilbert, P.M et al. **Characterization of b-haemolytic group B Streptococcus agalactiae in cultured seabream, Sparus auratus L., and wild mullet, Liza klunzingeri (Day), in Kuwait**. J. Fish Dis., v.25, p.505-513, 2002b.

Evans JJ, et al. 2004. **Efficacy of Streptococcus agalactiae (group B) vaccine in tilapia (Oreochromis niloticus) by intraperitoneal and bath immersion administration**. Vaccine. 22:3769-3773.

Evans, J. J.; Bohnsack, J. F.; Klesius, P. H.; Whiting, A. A.; Garcia, J. C.; Shoemaker, C. A.; Takahashi, S.. **Phylogenetic relationship among Streptococcus agalactiae from piscine, dolphin, bovine and human sources: a dolphin and piscine lineage associated with a fish epidemic in Kuwait is also associated with human neonatal infection in Japan**. Journal of Medical Microbiology v.57, p.1369-1376, 2008.

Farley MM. 2001. **Group B streptococcal disease in nonpregnant adults**. Clin. Infect. Dis. 33:556–561.

Feil, E. J. et al. **eBurst: inferring patterns of evolutionary descent among cluster of related bacterial genotypes from multilocus sequence typing data**. Journal of Bacteriology, Washington, v. 186, n. 5, p. 1518-1530, Mar. 2004.

Felsenstein, J. **Evolutionary trees from DNA sequences: a maximum likelihood approach.** J Mol Evol. 1981;17(6):368-76.

Fernandez M, et al. **Antimicrobial susceptibilities of group B streptococci isolated between 1992 and 1996 from patients with bacteremia or meningitis**. Antimicrob Agents Chemother, 42: 1517-1519, 1998.

Figueiredo, H. C. P. et al. **Streptococcus iniae outbreak in Brazilian Nile tilapia (Oreochromis niloticus) farms.** Brazilian Journal of Microbiology, São Paulo, v. 43, p. 576-580, 2012.

Fleischmann R, Adams M, White O, Clayton R, Kirkness E et al. (1995) **Whole-genome random sequencing and assembly of Haemophilus influenzae**. *Science* 269: 496-512.

Food and Agriculture Organization. **The state of world fisheries and aquaculture**. Rome, 2010. 197 p.

Foxman B, Gillespie BW, Manning SD, Marrs CF**: Risk factors for group B streptococcal colonization: potential for different transmission systems by capsular type.** Ann Epidemiol 2007; 17:854–862.

Ganz T. **Regulation of iron acquisition and iron distribution in mammals.** Biochim Biophys Acta 1763: 690-699, 2006.

Gevers, D. et al. **Opinion: Re-evaluating prokaryotic species.** Nat Rev Microbiol, 3(9):733{739, Sep 2005.

Gold. **Statistics Gold**. 2015. Web. 20 Mar. 2015. Disponível em: < https://gold.jgi-psf.org/statistics>.

Glaser, P. et al. **Genome sequence of Streptococcus agalactiae, a pathogen causing invasive neonatal disease**. Molecular Microbiology, Salem, v.45, p. 1499-1513, 2002.

Guimarães, Marcos Pezzi. Fasciola hepatica. In: Neves, David Pereira (Ed.). **Parasitologia humana**. 11. ed. São Paulo: Atheneu, 2005.

Hare R.F. **Francisella novicida Pathogenicity Island Encoded Proteins Were Secreted during Infection of Macrophage-Like Cells**. PLoS ONE 9(8): e105773, 2014.

Harrison L H, et al. **The Maryland emerging infectious. Serotype distribuition of invasive Group B Streptococcal isolates in Maryland: implications for vaccine formulation**. J Infect Dis , 177: 998-1002, 1998.

Heuer, O. E. et al. **Human health consequences of use of antimicrobial agents in aquaculture**. Clinical Infectious Diseases, Chicago, v. 49, n. 8, p. 1248-1253, 2009.

Holtenius, K.; Waller, KP.; Essen-Gustavsson, B.; Holtenius, P.; Hallen, C. S. **Metabolic parameters and blood leukocyte profiles in cows from herds with high or low mastitis incidence**. Vet. J. v. 168, p. 65-73, 2004.

Hoshina, T.; Sano, T.; Morimoto, Y. A**. Streptococcus pathogenic to fish**. Journal of Tokyo University of Fisheries, v. 44, p. 57-68, 1958.

Hsueh P., et al. **High incidence of erythromycin resistance among clinical isolates of Streptococcus agalactiae in Taiwan**. Antimicrob Agents Chemother, 45: 3205-08, 2001.

Human Genome, Science Genome Map, Science 16 February 2001: 291 (5507), 1218.

Husemann, P. **Bioinformatic Approaches for Genome Finishing**. Bielefeld University, Alemanha, 2011.

Instituto brasileiro de geografia e estatística – IBGE (2012). **Banco de dados**. Disponível em: ftp://ftp.ibge.gov.br/Producao_Pecuaria/Producao_da_Pecuaria_Municipal/2012/tabelas_pdf/tab06.pdf. Acessado em: 10, Abril. 2015.

Jiang, S. M. et al. **Variation in the group B Streptococcus CsrRS regulon and effects on pathogenicity**. Journal of Bacteriology, Washington, v. 190, n. 6, p. 1956-1965, Mar. 2008.

Johri, A. K. et al. **Group B *Streptococcus*: global incidence and vaccine development**. Nature Reviews Microbiology, London, v. 4, n. 12, p. 932-942, Dec. 2006.

Jones N, ohnsack JF, Takahashi S, Oliver KA, Chan MS, Kunst F, et al. **Multilocus sequence typing system for group B streptococcus**. J Clin Microbiol. 2003;41: 2530–2536 12791877

Jünemann, S. et al. **Updating benchtop sequencing performance comparison**. Nature Biotechnol, 31, 294–296, 2013.
Kanehisa, M. **KEGG: Kyoto Encyclopedia of Genes and Genomes**. Nucleic Acids Research, London, v. 28, p. 27-30, 2000.

Kaur R. **Next Generation Sequencing: A Revolution in Gene Sequencing**. *CIBTech Journal of Biotechnology* 2 (4): 1-20, 2013.

Kegg - **Glycine betaine/proline transport system** - <http://www.kegg.jp/module/M00208>.
Acessado em: 20, Junho. 2015.

Kong, F., L. M. Lambertsen, H-C. Slotved, D. Ko, H. Wang, and G. L. Gilbert. 2008. **Use of Phenotypic and Molecular Serotype Identification Methods To Characterize Previously Nonserotypeable Group B Streptococci**. Journal of Clinical Microbiology. 46:2745-2750.

Lancefield, R.C., and R. Hare. 1935**. The serological differentiation of pathogenic and non-pathogenic strains of hemolytic Streptococci from parturient women**. Journal of Experimental Medicine. 61:335-349.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wal-lace IM, Wilm A, Lopez R, et al. **Clustal W and Clustal X version 2.0**. Bioinformatics 2007; 23:2947-2948.

Lauer, P., C. D. et al,. **Genome Analysis Reveals Pili in Group B Streptococcus**. Science. 309:105, 2005.

Leigh, J. A**. Streptococcus uberis: a permanent barrier to the control of bovine mastitis**. The Veterinary Journal, v. 157, n. 3, p. 225-238, 1999.

Life Technologies. **Ion PGM™ and Ion Proton™ System *Chip*s**. Disponível em: <http://www.lifetechnologies.com/br/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-pgm-ion-proton-system-*chip*s.html>. Acesso em: 20 de abril, 2015.

Lim C, et al. 2006. **Tilapia: biology, culture and nutrition**. An Imprint of the Haworth Press, New York, United States: 678.

Loman, N. J. et al. **High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity**. Nature reviews. Microbiology, Nature Publishing Group, v. 10, n. 9, p. 599–606, set. 2012a.

Loman, N. J. et al. **Performance comparison of benchtop high-throughput sequencing platforms**. Nature Biotechnology 30, 434–439, April, 2012b.

LPSN. **List of Prokaryotic names - Genus Streptococcus Online**. 2015. Web. 20 Mar. 2015. Disponível em: <http://www.bacterio.net/streptococcus.html>.

Ludwig W. and Klenk H. **Overview: a phylogenetic backbone and taxonomic framework for procaryotic systematics.** In In Bergey's Manual of Systematics Bacteriology. Second Edition., pag 49-65. Springer-Verlag. Berlin., 2001.

Madigan, T. M.; Martinko, J. M.; Parker, J. **Microbiologia de Brock**. 10a edição, São Paulo: Prentice Hall, 2004.

Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms**. *Proc Natl Acad Sci USA* 1998, 95:3140-3145.

Mariano, Diego César Batista. **SIMBA: uma ferramenta Web para gerenciamento de montagens de genomas bacterianos**. Dissertação de mestrado. Programa de Pós-Graduação em Bioinformática da UFMG. Belo Horizonte (MG), 2015.

Mata, A. I. et al. **Multiplex PCR Assay for Detection of Bacterial Pathogens Associated with Warm-Water Streptococcosis in Fish**. Applied and Environmental Microbiology, v. 70, n. 5, p. 3183-3187, 2004.

Mcdonald L. C. et al,. **Peripartum transmission of penicin-resistant Streptococcus pneumoniae**. J Clin Microbiol, 41 (5): 2258-2260, 2003.

McPherson, R. A. and M. R. Pincus. 2007. **Henry's Clinical Diagnosis and Management by Laboratory Methods (21 ed.)**. Saunders, Philadelphia.

Melin, P. **Neonatal group B Streptococcal disease: from pathogenesis to preventive strategies**. Clinical Microbiology and Infection. 17:1294-1303, 2011.

Metcalf HE, Luchsinger DW, Ray WC. Brucellosis. In: Beran GW, Steele JH. Handbook of zoonoses**. Section A: bacterial, rickettsial, chlamydial, and mycotic**. 2nd ed. Boca Raton: CRC Press; 1994. p.167.

Metzker, M.L et al. **Sequencing technologies - the next generation**. Nat Ver Genet, 11(1):31-46, 2010.

Miller, J. R., et al. **Assembly algorithms for next-generation sequencing data**. Genomics 95: 315-327, 2010.

Ministério da Agricultura - **Caprinos e Ovinos.** Disponível em: http://www.agricultura.gov.br/animal/especies/caprinos-e-ovinos Acessado em: 10, Abril. 2015.

Ministério da Pesca e Aquicultura – **Informações e Estatísticas**. Banco de dados. Disponível em: http://www.mpa.gov.br/files/Docs/Informacoes_e_Estatisticas/Boletim%20Estat%C3%ADstico%20MPA%202010.pdf - Acessado em: 10, Abril. 2015.

Merl, K. et al. **Determination of epidemiological relationships of Streptococcus agalactiae isolated from bovine mastitis**. FEMS Microbiology Letters, London, v. 226, p. 87-92, 2003.

Mian, G. F. et al. **Aspects of the natural history and virulence of S. agalactiae infection in Nile tilapia**. Veterinary Microbiology, Amsterdam, v. 136, n. 1-2, p. 180-183, 2009.

Morales W. J. et al. **Change in antibiotic resistance of group B Streptococcus: impact on intrapartum management**. Am J Obstet Gynecol, 181(2): 310-314, 1999.

Nakib, M. et al. **Comparison of the Diversilab system with multi-locus sequence typing and pulsed-field gel electrophoresis for the characterization of *Streptococcus agalactiae* invasive strains**. Journal of Microbiological. Methods, Amsterdam, v. 85, n. 2, p. 137-142, May 2011.

Narváez. J. et al. **Group B streptococcal spondylodiscitis in adults: 2 case reports**. Joint Bone Spine, Paris, v. 71, n. 4, p. 338-343, July 2004.

Nei M. and Kumar S. (2000). **Molecular Evolution and Phylogenetics**. Oxford University Press, New York.

Olivares-fuster, O. et al. **Molecular typing of Streptococcus agalactiae isolates from fish**. Journal of Fish Diseases, Oxford, v. 31, n. 4, p. 277-283, 2008.

Pavón, A. B. I. et al. **Multilocus sequence typing. In: Molecular epidemiology of microorganisms.** Clifton: Humana Press, 2009. chap. 11, p. 129-140. (Methods in Molecular Microbiology, v. 551).

Park SE, Jiang S and Wessels MR (2012). **CsrRS and environmental pH regulate group B Streptococcus adherence to human epithelial cells and extracellular matrix.** Infect. Immun. 80: 3975-3984.

Pelczar JR, M.J.; Chan, E.C.S.; Krieg, N.R. **Microbiologia: conceitos e aplicações**. Traducao de Sueli Yamada, Tania Ueda Nakamura, Benedito Prado Dias Filho. São Paulo: Makron Books, 1996. 524 p. 1 v.

Peres Neto, F.; Zappa,V. **Mastite em vacas leiteiras**. Revista Cientifica Eletrônica de Medicina Veterinária, Graça, SP, a. 9, n. 16, 2011.

Philpot, W. N.; Nickerson, S. C. **Mastitis: counter attack. A strategy to combat mastitis.** Naperville: Babson Bros. Co. , 1991. 150 p.

Pietrocola, G., et al. **FbsA, a fibrinogenbinding protein from Streptococcus agalactiae, mediates platelet aggregation**. Blood. United States, 105(3):1052–1059, 2005.

Pinto, T. C. A. et al. **Distribution of serotypes and evaluation of antimicrobial susceptibility among human and bovine Streptococcus agalactiae strains isolated in Brazil between 1980 and 2006**. Brazilian Journal Infectious Diseases, São Paulo, 2013.

Pogere, A. et al. **Prevalência da colonização pelo streptococo do grupo B em gestantes atendidas em ambulatório de pré-natal**. Revista Brasileira de Ginecologia e Obstetrícia, Rio de Janeiro. v. 27, n. 4, p.174-170, abr. 2005.

Pop, M. **Genome assembly reborn: recent computational challenges**. Briefings in bioinformatics, v. 10, n. 4, p. 354–66, 2009.

Portal Brasil - **Produção de pescado no País**. Disponível em: http://www.brasil.gov.br/governo/2013/03/producao-de-pescado-no-pais-cresce Acessado em: 10, Abril. 2015.

Poyart, C. et al. **Multiplex PCR assays from rapid and accurate capsular typing of group B streptococci**. Journal of Clinical Microbiology, Washington, v. 45, p. 1985-1988, 2007.

Pritchard D. G et al. **Characterization of the groupspecific polysaccharide of group B Streptococcus**. Arch biochem and Biophys, 1984, 235(2):385-392.

Quinn, P.J.; Markey, B.K., Carter, M.E. et al. (Eds). **Microbiologia veterinária e doenças infecciosas**. Porto Alegre: Artmed, 2005. 512p.

Quentin, R., H. et al. **Characterization of *Streptococcus agalactiae* strains by multilocus enzyme genotype and serotype: identification of multiple virulent clone families that cause invasive neonatal disease**. Journal of Clinical Microbiology. 33:2576-2581, 1995.

Radostits, O. M.; Blood D.C.; Gay, C.C. **Um tratado de doenças dos bovinos, ovinos, suínos, caprinos e eqüinos**. 9. ed. Rio de Janeiro: Guanabara Koogan, 2002. 1737 p.

Ramaswamy, S.V., P. Ferrieri, L. C. Madoff, A. E. Flores, N. Kumar, H. Tettelin, and L. C. Paoletti. 2006. **Identification of a novel cps locus polymorphisms in nontypable group B Streptococcus**. Journal of Medical Microbiology. 55:775-783.

Ramos, Rommel Thiago Jucá. **Desenvolvimento de um "suíte" de aplicativos computacionais para suporte à montagem de genomas bacterianos a partir de leituras curtas**. Dissertação de mestrado. Programa de Pós-Graduação em Genética e Biologia Molecular da UFPA. Belém (PA), 2011.

Rajagopal L. L. **Understanding the regulation of Group B Streptoccocal virulence factors**. Future Microbiology, London, v. 4. p. 201-221, 2009.

Rattanachaikunsopon, P.; Phumkhachorn, P. **Prophylactic effect of Andrographis paniculata extracts against Streptococcus agalactiae infection in Nile tilapia (Oreochromis niloticus).** Journal of Bioscience and Bioengineering, Osaka, v. 107, n. 5, p. 579-582, 2009.

Regan J. A. et al,. **Colonization with group B streptococci in pregnancy and adverse outcome**. Am J Obstet Gynecol, 74:1354-1360, 1996.

Ribeiro, F. J. et al. **Finished bacterial genomes from shotgun sequence data.** Genome research, v. 22, n. 11, p. 2270–7, 2012.

Richards, V. P. et al. **Comparative genomics and the role of lateral gene transfer in the evolution of bovine adapted Streptococcus agalactiae**. Infection, Genetics and Evolution, Amsterdam, v. 11, n. 6, p. 1263-1275, 2011.

Rothberg, JM. et al. **An integrated semiconductor device enabling non-optical genome sequencing**. Nature, 475(7356):348–352, April 2011.

Ruoff, K. L., R. A. Whiley, and D. Beighton. 2003. Streptococcus, p.405-417. In P. R. Murray, E. J. Baron, J. H. Jorgensen, M. A. Pfaller, and R. H. Yolken (8 ed.), **Manual of Clinical Microbiology, vol. 1**. ASM Press, Washington, DC.

Safadi, R., S. Amor, G. Héry-Arnaud, B. Spellerberg, P. Lanotte, L. Mereghetti, F. Gannier, R. Quentin, and A. Rosenau. **Enhanced expression of lmb gene encoding laminin-binding protein in Streptococcus agalactiae strains harboring IS1548 in scpB-lmb intergenic region**. PLoS One. 5: e10794, 2010.

Santos, E. M. P.; Brito, M. A. V. P.; Lange, C. C.; Brito, J. R. F.; Cerqueira, M. M. O. **P. Streptococcus e gêneros relacionados como agentes etiológicos de mastite bovina**. Acta Scientiae Veterinariae, v. 35, n. 1, p. 17-27,2007.

Sharma, P. et al. **Role of pili proteins in adherence and invasion of Streptococcus agalactiae to the lung and cervical epithelial cells**. The Journal of Biological Chemistry, Bethesda, v. 228, n. 6, p. 4023-4034, Dec. 2012.

Sheftel A, Stehling O, Lill R 2010. **Iron–sulfur proteins in health and disease**. Trends in Endocrin Met 21: 302-314.

Shewmaker, P. L. et al. **Streptococcus ictaluri sp. nov., isolated from Channel Catfish Ictalurus punctatus broodstock**. International Journal of Systematic and Evolutionary Microbiology, Reading, v. 57, n. 7, p. 1603-1606, 2007.

Sikri N, Bardia A. ***A History of Streptokinase Use in Acute Myocardial Infarction.*** *Texas Heart Institute Journal*. 2007;34(3):318-327.

Slotved HC, Kong F, Lambertsen L, Sauer S, Gilbert GL. **Serotype IX, a proposed new Streptococcus agalactiae serotype**. J Clin Microbiol. 2007;45(9):2929-36.

Suresh AV. 1998. **Tilapia update 1998**. World Aquaculture. 30:8-68.

Spellerberg, B. **Pathogenesis of neonatal Streptococus agalactiae infections**. Microbes and Infection, Paris, v. 2, n. 14, p. 1733-1742, Nov. 2000.

Springman, A. C. et al. **Selection, recombination and virulence genes among group B streptococcal genotype.** Journal of Bacteriology, Washington, v. 191, n. 17, p. 5419-5427, Sept. 2009.

Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. **MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.** Mol. Biol. Evol. 2013; 30:2725–9

Tettelin, H. et al. **Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial pan-genome**. P. Acad. Nat Sci. Washington, v. 102, p. 13950-13955, 2005.

Thompson, C. C., Vicente, A. C. P., Souza, R. C. , *et al*. **Genomic taxonomy of vibrios**. BMC Evol Biol, 9:258, 2009.

Thompson, F. L., Thompson, C. C., *et al***. Photobacterium rosenbergii sp. nov. and Enterovibrio coralii sp. nov., vibrios associated with coral bleaching**. Int J Syst Evol Microbiol, 55(Pt 2):913{917, Mar 2005.

Tortora, G.J.; Funke, B.R.; Case, CL. **Microbiologia**. 10. ed., Porto Alegre: Artmed, 2010.

Urwin, R. et al. **Multi-locus sequence typing: a tool for global epidemiology**. Trends in Microbiology, Cambridge, v. 11, n. 10, p. 479-487, Oct. 2003.

Vandamme, P. et al. **Polyphasic taxonomy, a consensus approach to bacterial systematics**. Microbiol Rev, 60(2), Jun 1996.

Verani JR, McGee L, Schrag SJ; **Division of Bacterial Diseases, National Center for Immunization and Respiratory Diseases, Centers for Disease Control and Prevention (CDC).** Prevention of perinatal group B streptococcal disease - revised guidelines from CDC, 2010. MMWR Recomm Rep. 2010;59(RR-10):1-36.

Vieira, Mônica Aparecida Midolli. **Ilhas de patogenicidade.** O Mundo da Saúde, São Paulo: 2009;33(4):406-414.

Waterhouse, A.M., Procter, J.B., Martin, D.M.A, Clamp, M. and Barton, G. J. (2009) "**Jalview Version 2 - a multiple sequence alignment editor and analysis workbench**". Bioinformatics25 (9) 1189-1191.

Winn WC Jr, Alen SD, Janda WW, Koneman EW, Procop GV, Schrenkenberger PC, et al. **Koneman's color atlas and textbook of diagnostic microbiology**. 6th ed. Philadelphia: Lippincott Williams & Wilkins; Chapter 13: Gram-Positive Cocci: Part II: Streptococci, Enterococci and the "Streptococcus-Like" Bacteria; p. 683-713, 2006.

Zubair S, de Villiers EP, Younan M, et al. **Genome Sequences of Two Pathogenic Streptococcus agalactiae Isolates from the One-Humped Camel Camelus dromedarius**. Genome Announcements. 2013;1(4):e00515-13. doi:10.1128/genomeA.00515-13.

Anexos

Currículo Lattes

Anexo 1: Análises suplementares: Funcional com COG, Análise de ProFago e Análise de vias metabólicas.

Anexo 2: Publicação na revista Genome Announcements - Genome Sequence of Lactococcus lactis subsp. lactis NCDO 2118, a GABA-Producing Strain

Anexo 3: Capítulo de livro publicado SMGroup - Bioinformatics

Anexo 4: Artigo submetido a revista Microbial Cell Factory **-** Putative Virulence Factors of *Corynebacterium pseudotuberculosis* FRC 41: Vaccine Potential and Protein Expression Using Multiple Escherichia coli Strains.

Edgar Lacerda de Aguiar

Endereço para acessar este CV: http://lattes.cnpq.br/9444042971968548
Última atualização do currículo em 21/06/2015

Atualmente cursa Mestrado em Bioinformática na Universidade Federal de Minas Gerais. Possui Graduação em Sistemas de Informação pela Faculdade Anhanguera de Belo Horizointe (2012) e Graduação em Análise e Desenvolvimento de Sistemas - Anhanguera Educacional de Belo Horizonte (2012). Possui experiência na área de análise de requisitos e desenvolvimento de sistemas, além de conhecimento dos seguintes temas: bioinformática, genômica comparativa bacteriana, filogenômica , anotação e montagem de genomas bacterianos **(Texto informado pelo autor)**

Identificação

Nome Edgar Lacerda de Aguiar
Nome em citações bibliográficas AGUIAR, E. L.;DE AGUIAR, E. L.

Endereço

Endereço Profissional Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas.
Universidade Federal de Minas Gerais
Pampulha
31270901 - Belo Horizonte, MG - Brasil - Caixa-postal: 31741285
Telefone: (31) 34092982

Formação acadêmica/titulação

2013 Mestrado em andamento em Bioinformática.
Universidade Federal de Minas Gerais, UFMG, Brasil.
Título: Sequenciamento, montagem e anotação do genoma de Streptococcus Agalactiae GBS85147: uma abordagem comparativa,Orientador: Vasco Ariston de Carvalho Azevedo.
Coorientador: Anne Cybelle Pinto.
Grande área: Ciências Biológicas
Grande Área: Ciências Exatas e da Terra / Área: Ciência da Computação / Subárea: Metodologia e Técnicas da Computação / Especialidade: Engenharia de Software.
Grande Área: Ciências Biológicas / Área: Microbiologia / Subárea: Genômica Comparativa Bacteriana.

2007 - 2012 Graduação em Sistemas de Informação.
Faculdade Anhanguera, ANHAN, Brasil.
Título: Comparando o desempenho de bancos de dados NoSQL e relacionais manipulando dados biológicos.
Orientador: Sandro Renato Dias.
Bolsista do(a): Programa Universidade para Todos, PROUNI, Brasil.

2007 - 2012 Graduação em Análise e Desenvolvimento de Sistemas.
ANHANGUERA EDUCACIONAL (BH), ANHAN, Brasil.
Título: cDNA - Cliente.
Orientador: Sandro Renato Dias.
Bolsista do(a): Faculdade Anhanguera, ANHAN, Brasil.

Formação Complementar

2014 - 2014 Docência para Ensino Superior. (Carga horária: 60h).
Universidade Federal de Minas Gerais, UFMG, Brasil.

2014 - 2014 Anotação avançada de Sequências e uso de pipelines. (Carga horária: 15h).
Universidade Federal de Minas Gerais, UFMG, Brasil.

2014 - 2014 Data Base in THOMSON REUTERS INTEGRITY/BIOMARKERS. (Carga horária: 3h).
Universidade Federal de Minas Gerais, UFMG, Brasil.

2014 - 2014 Biocuradoria de Termos Gene Ontology. (Carga horária: 15h).
Universidade Federal de Minas Gerais, UFMG, Brasil.

2014 - 2014 Bacterial Metagenomics. (Carga horária: 3h).
Fundação Oswaldo Cruz (MG), FIOCRUZ, Brasil.

2014 - 2014 PATRIC: recursos integrados para estudo de sistema. (Carga horária: 15h).
Universidade Federal de Minas Gerais, UFMG, Brasil.

2014 - 2014 Phage genomics and metagenomics. (Carga horária: 3h).
Fundação Oswaldo Cruz (MG), FIOCRUZ, Brasil.

2014 - 2014 The Data Scientist?s Toolbox.
Johns Hopkins University, JHU, Estados Unidos.

2013 - 2013 Extensão universitária em Bioinformática Estrutural e Análises de Proteoma. (Carga horária: 90h).
Universidade Federal de Minas Gerais, UFMG, Brasil.

Atuação Profissional

Universidade Federal de Minas Gerais, UFMG, Brasil.

Vínculo institucional

2013 - Atual Vínculo: Bolsista, Enquadramento Funcional: Bolsista, Carga horária: 20, Regime: Dedicação exclusiva.

Faculdade Anhanguera, ANHAN, Brasil.

Vínculo institucional

2011 - 2012 Vínculo: Bolsista, Enquadramento Funcional: Pesquisador, Carga horária: 20

Companhia de Tecnologia da Informação do Estado de Minas Gerais, PRODEMGE, Brasil.

Vínculo institucional

2011 - 2012 Vínculo: Colaborador, Enquadramento Funcional: Análista e Desenvolvedor de Sofware, Carga horária: 20

Banco BMG, BMG, Brasil.

Vínculo institucional

2009 - 2010 Vínculo: Colaborador, Enquadramento Funcional: Análista e Desenvolvedor de Sofware, Carga horária: 30

MCA Serviços, MCA, Brasil.

Vínculo institucional

2010 - 2011 Vínculo: Colaborador, Enquadramento Funcional: Analista e Desenvolvedor de Software, Carga horária: 20

Projetos de pesquisa

2011 - 2012

Comparando o desempenho de Bancos de Dados NoSQL e Relacionais manipulando dados biológicos

Descrição: Apresentar o uso de bancos de dados relacionais e não relacionais do tipo NoSQL para controlar e manipular grandes quantidades de dados, comparando o desempenho de um SGBD (Sistema de Gerenciamento de Bancos de Dados) não relacional, o MongoDB, com um SGBD relacional, o MySQL. Tem como objetivo verificar as vantagens e as desvantagens do uso de bancos de dados não relacionais em comparação aos bancos de dados relacionais, utilizando a base de dados biológicos PDB (Protein Data Bank) para testes de leitura de arquivos em disco e gravação de seu conteúdo nos SGBDs..
Situação: Concluído; Natureza: Pesquisa.
Alunos envolvidos: Graduação: (2) / Doutorado: (1) .

Integrantes: Edgar Lacerda de Aguiar - Integrante / Diego César Batista Mariano - Integrante / Sandro Renato Dias - Coordenador.

2011 - 2012

cDNA - Cliente

Projeto certificado pelo(a) coordenador(a) Sandro Renato Dias em 19/08/2013.

Descrição: Este projeto tem como objetivo minimizar diversos problemas no processamento de tarefas das aplicações de sobreposição de proteínas que possuem um desempenho abaixo do necessário para atual demanda da bioinformática, levando em consideração a dificuldade de se desenvolver uma aplicação que suporte a divisão de tarefas em diversos processos com baixo custo e elevado desempenho. Com o intuito de utilizar múltiplas execuções das sobreposições dos conjuntos de estruturas entre os átomos das proteínas interligando diversas máquinas clientes em uma rede controlada pelo Servidor e integrada com Banco de Dados. O módulo Cell Distributed Network Application - Cliente é responsável por gerenciar o processamento de tarefas distribuídas pelo servidores aonde foram desenvolvidos soluções para compactação e serialização dos dados gerados pelo resultado da sobreposição, com objetivo de minimizar o trafego de dados aumentando o desempenho da aplicação..
Situação: Concluído; Natureza: Pesquisa.
Alunos envolvidos: Graduação: (2) / Doutorado: (1) .

Integrantes: Edgar Lacerda de Aguiar - Integrante / Sandro Renato Dias - Coordenador / FERNANDO CÉSAR PEREIRA MAGIAG - Integrante.

Áreas de atuação

1. Grande área: Ciências Biológicas / Área: Biologia Geral / Subárea: Bioinformática.

2. Grande área: Ciências Biológicas / Área: Microbiologia / Subárea: Genômica Comparativa Bacteriana.

3. Grande área: Ciências Biológicas / Área: Microbiologia / Subárea: Engenharia de Software.

4. Grande área: Ciências Biológicas / Área: Microbiologia / Subárea: Banco de Dados.

5. Grande área: Ciências Biológicas / Área: Microbiologia / Subárea: Linguagens de Programação.

6. Grande área: Ciências Biológicas / Área: Genética / Subárea: Filogenômica.

Idiomas

Inglês Compreende Bem, Fala Pouco, Lê Bem, Escreve Bem.

Espanhol Compreende Razoavelmente, Fala Pouco, Lê Razoavelmente, Escreve Razoavelmente.

Português	Compreende Bem, Fala Bem, Lê Bem, Escreve Bem.

Prêmios e títulos

2014	1º lugar apresentação de pôsteres no ISCB - Latin America X-Meeting on Bioinformatics with BSB and SoiBio 2014. SIMBA: A Web Tool for Complete Assembly of Bacterial Genomes, ISCB - LA / X-Meeting / BSB / SoiBio.
2013	Prêmio Anhanguera de Mérito Científico e Acadêmico - 1º lugar na categoria: Programa de Iniciação Científica - Ciências Exatas, da Terra e Engenharias, Anhanguera Educacional.

Produções

Produção bibliográfica

Artigos completos publicados em periódicos

Ordenar por

Ordem Cronológica

1. OLIVEIRA, L. C. SARAIVA, T. D. L. SOARES, S. C. RAMOS, R. T. J. SA, P. H. C. G. CARNEIRO, A. R. MIRANDA, F. FREIRE, M. RENAN, W. JUNIOR, A. F. O. SANTOS, A. R. PINTO, A. C. SOUZA, B. M. CASTRO, C. P. DINIZ, C. A. A. ROCHA, C. S. MARIANO, D. C. B. **DE AGUIAR, E. L.** FOLADOR, E. L. BARBOSA, E. G. V. ABURJAILE, F. F. GONCALVES, L. A. GUIMARAES, L. C. AZEVEDO, M. AGRESTI, P. C. M. , et al. ; Genome Sequence of Lactococcus lactis subsp. lactis NCDO 2118, a GABA-Producing Strain. Genome Announcements, v. 2, p. e00980-14-e00980-14, 2014.

2. MARIANO, D. C. B. ; **AGUIAR, E. L.** ; DIAS, S. R. . Comparando o desempenho de Bancos de Dados NoSQL e Relacionais manipulando dados biológicos. Anais do Seminário de Produção Acadêmica da Anhanguera, v. 3, p. 1, 2012.

Trabalhos completos publicados em anais de congressos

1. **AGUIAR, E. L.** ; DIAS, S. R. ; MAGIAG, F. C. P. . Cdna-Cliente. In: 12º Congresso de Iniciação Científica CONIC-SEMESP, 2012, São Paulo. 12º Congresso de Iniciação Científica CONIC-SEMESP, 2012.

2. **AGUIAR, E. L.** ; MARIANO, D. C. B. ; DIAS, S. R. . Comparando o desempenho de bancos de dados Nosql e relacionais manipulando dados biológicos. In: 12º Congresso de Iniciação Científica CONIC-SEMESP, 2012, São Paulo. 12º Congresso de Iniciação Científica CONIC-SEMESP, 2012.

Resumos publicados em anais de congressos

1. MARIANO, D. C. B. ; OLIVEIRA, L. C. ; FOLADOR, E. L. ; **AGUIAR, E. L.** ; BENEVIDES, L. ; PEREIRA, F. L. ; VIANA, M. V. C. ; SOUSA, T. J. ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . SIMBA: A Simple Way to Make Complete Assemblies of Bacterial Genomes. In: First ISCB Latin American Student Council Symposium, 2014, Belo Horizonte. Program Booklet - ISCB LA Student Council Symposium, 2014. p. 24-25.

2. **AGUIAR, E. L.** ; MARIANO, D. C. B. ; OLIVEIRA, L. C. ; AMORIM, L. G. ; OLIVEIRA JUNIOR, A. F. ; ROCHA, F. S. ; PEREIRA, F. L. ; SOARES, S. C. ; DORELLA, F. A. ; LEAL, C. ; FIGUEIREDO, H. C. P. ; AZEVEDO, V. A. C. . The complete genome sequence of Streptococcus agalactiae strain GBS85147. In: ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoiBio, 2014, Belo Horizonte. ISCB - LA / X-Meeting / BSB / SoiBio, 2014.

3. **AGUIAR, E. L.** ; MARIANO, D. C. B. ; OLIVEIRA, L. C. ; AMORIM, L. G. ; OLIVEIRA JUNIOR, A. F. ; ROCHA, F. S. ; PEREIRA, F. L. ; SOARES, S. C. ; DORELLA, F. A. ; LEAL, C. A. G. ; FIGUEIREDO, H. C. P. ; AZEVEDO, V. A. C. . The complete genome sequence of Streptococcus agalactiae strain GBS85147. In: First ISCB Latin American Student

Council Symposium, 2014, Belo Horizonte. Program Booklet - ISCB LA Student Council Symposium, 2014. p. 25-26.

4. SANTANA, K. T. O. ; TARTAGLIA, N. R. ; SILVA, R. F. ; MARIUTTI, R. B. ; **AGUIAR, E. L.** ; PORTELA, R. W. D. ; ARNI, R. K. ; MEYER, R. J. ; SILVA, A. ; AZEVEDO, V. A. C. . IN SILICO CHARACTERIZATION, CLONING AND HETEROLOUGOUS EXPRESSION OF FIVE C. PSEUDOTUBERCULOSIS PROTEINS PROBABLY INVOLVED IN VIRULENCE. In: 5º Encontro de Genética de Minas Gerais, 2014, Belo Horizonte. V Encontro de Genética de Minas Gerais, 2014.

5. MARIANO, D. C. B. ; OLIVEIRA, L. C. ; FOLADOR, E. L. ; **AGUIAR, E. L.** ; BENEVIDES, L. ; PEREIRA, F. L. ; RAMOS, R. T. J. ; AZEVEDO, V. A. C. . SIMBA: A Web Tool for Complete Assembly of Bacterial Genomes. In: ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoiBio, 2014, Belo Horizonte. Anais do ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoiBio, 2014.

6. MARIANO, D. C. B. ; **AGUIAR, E. L.** ; DIAS, S. R. . Comparando o desempenho de bancos de dados Nosql e relacionais manipulando dados biológicos. In: XXI Congresso de Pós-graduação da UFLA, 2012, Lavras. Anais - XXI Congresso, 2012.

7. MARIANO, D. C. B. ; **AGUIAR, E. L.** ; DIAS, S. R. . Comparing the performance of databases relational and NoSQL for manipulating biological data. In: 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology, 2012, Campinas. AbstractBook - X-meeting 2012, 2012.

Artigos aceitos para publicação

1. **AGUIAR, E. L.** ; MAGIAG, F. C. P. ; DIAS, S. R. . Cdna-Cliente. Resumos dos Trabalhos de Iniciação Científica (Itatiba), 2013.

2. MARIANO, D. C. B. ; **AGUIAR, E. L.** ; DIAS, S. R. . Comparando o desempenho de bancos de dados NoSQL e relacionais manipulando dados biológicos. Resumos dos Trabalhos de Iniciação Científica (Itatiba), 2013.

Eventos

Participação em eventos, congressos, exposições e feiras

1. ISCB - Latin American X-Meeting on Bioinformatics with BSB & SoiBio. The complete genome sequence of Streptococcus agalactiae strain GBS85147. 2014. (Congresso).

2. 3ª Escola de Verão em Computação do Departamento de Ciência da Computação da UFMG. 2014. (Seminário).

3. Simpósio em Imunidade Antiviral e Dengue. 2014. (Simpósio).

4. Latin American Student Council Symposium.The complete genome sequence of Streptococcus agalactiae strain GBS85147. 2014. (Simpósio).

5. 3º International Workshop on Environmental Microbiology. 2014. (Oficina).

6. Reunião Anual Casadinho - Procad entre Pós Bioinformatica da Ufmg e BCS FIOCRUZ.Streptoccus agalactiae GBS85147 - Genome complete sequencing. 2014. (Encontro).

7. Simpósio Microbiota: the bacteria that govern us. 2013. (Simpósio).

8. Innate Immunity at a Glance. 2013. (Oficina).

9. 8th International Conference of the Brazilian Association for Bioinformatics and Computational Biology. Comparing the performance of databases relational and NoSQL for manipulating biological. 2012. (Congresso).

10. 12º Congresso de Iniciação Científica CONIC-SEMESP. Comparando o desempenho de bancos de dados Nosql e

relacionais manipulando dados biológicos. 2012. (Congresso).

11. 6º Seminário de Produção Acadêmica da Anhanguera. 2012. (Seminário).

Anexo 1

1.1 Análises do COG

No Capítulo 1 foram realizadas análises contendo a distribuição funcional de genes dos genes da linhagem GBS85147, usando BD do COG. As linhagens ILRI005 e 138spar foram submetidas ao COG com objetivo de efetuar uma comparação funcional entre elas. O COG retorna as informações num arquivo gff, os dados são exportados para uma tabela. Após o tratamento das informações foram criados gráficos para facilitar a comparação entre os processos dos genes de cada linhagem.

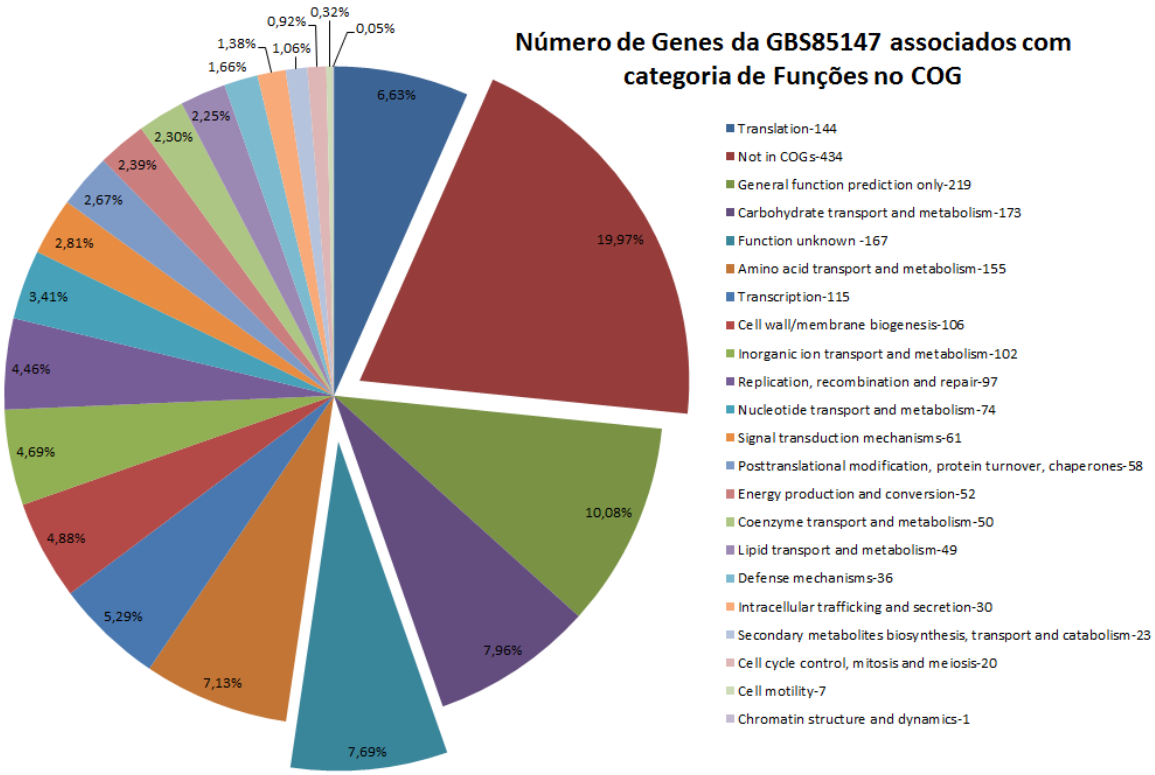


Figura A - Funções COG com GBS85147

Na Figura A os genes desconhecidos (regiões em destaque) e que não foram encontrados na base do COG da linhagem GBS85147 foram de aproximadamente 28%, totalizando 600 genes.

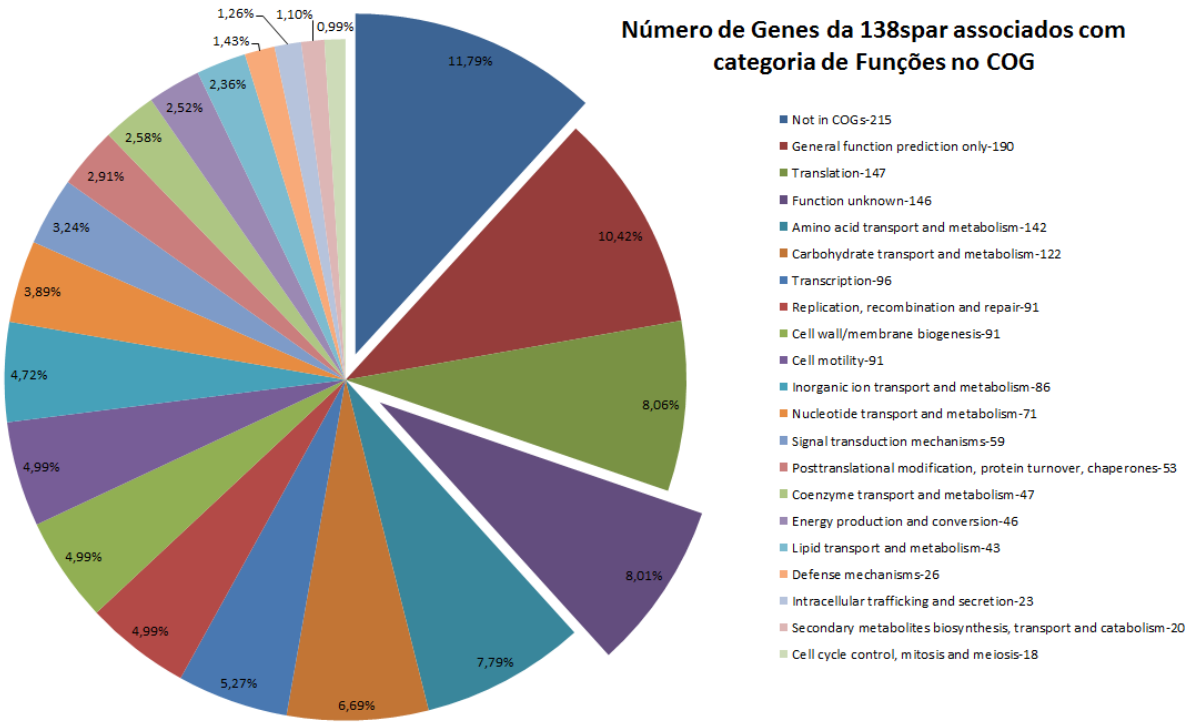


Figura B - Funções COG com 138spar

Já na linhagem 138spar os genes desconhecidos e que não foram encontrados na base do COG foram de aproximadamente 20% totalizando 361 genes (Figura B).

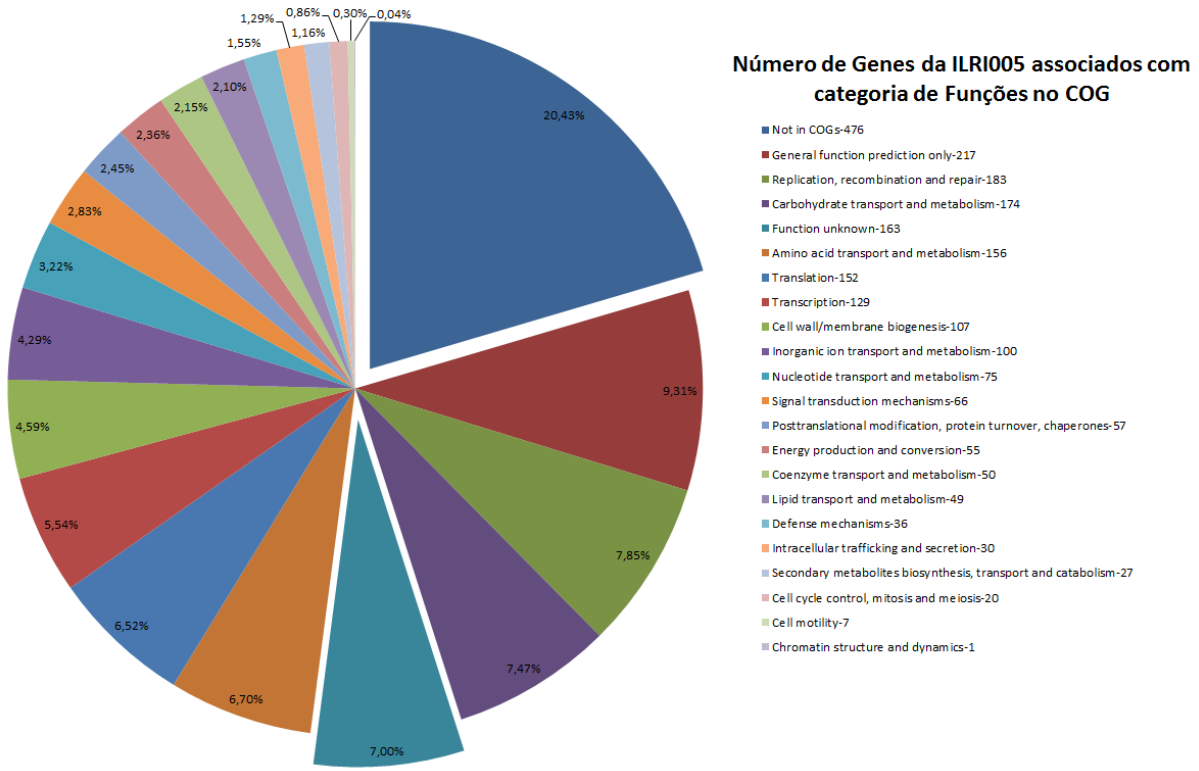


Figura C - Funções COG com ILRI005

Na linhagem ILRI005 os genes desconhecidos e que não foram encontrados na base do COG foram de aproximadamente 27% totalizando 639 genes (Figura C).

A função com maior número de genes em todas as linhagens é “Não está no COG”. Já a “Função desconhecida” nas linhagens ILRI005 e a GBS85147 tiveram total de aproximadamente 20% e 138spar só 8%. Quando são comparados apenas os genes sem função predita entre as três linhagens é possível notar que ainda existe elevado número de genes desconhecidos, isso dificulta os estudos comparativos. Seria necessário novos estudos com core genoma da *S. agalactiae* para facilitar futuros estudos genômicos.

Nas três linhagens, a segunda função mais comum foi “Função geral” que contém as funções essenciais para sobrevivência das linhagens.

Já a terceira função mais comum nas linhagens foi bem variada. Na GBS85147 foi “Transporte de carboidratos e metabolismo”, na 138spar foi “Tradução” e na ILRI005 “Replicação, recombinação e reparação”. Essa variação de funções pode ser um provável indicador da capacidade adaptativa da espécie. Pois é notável a divergência de funções nas linhagens quando comparadas por hospedeiros.

1.2 Predição de Profagos

Para predição de profagos foi utilizado websoftware *Phast,* o mesmo suporta arquivos do tipo gbk e fna. Ele efetua uma busca no arquivo e prediz a localização de profagos, baseado num DB próprio. Ao localizar um profago, retorna os genes envolvidos e uma visualização do trecho no genoma a qual este profago está localizado. O web*software Phast* pode ser acessado no *website http://phast.wishartlab.com/*

Através do software Phast foi possível predizer a existência de um profago na linhagem GBS85147. Porém, o mesmo foi classificado como um profago incompleto, pois faltavam alguns genes essenciais para ser considerado um profago completo.

É necessário novos estudos envolvendo os genes descritos nesse fago, pois diversos genes ainda são desconhecidos. Outro estudo importante seria a busca por fagos nas demais linhagens. Esses estudos poderiam auxiliar a elucidar como ocorre a transferência genica da espécie.

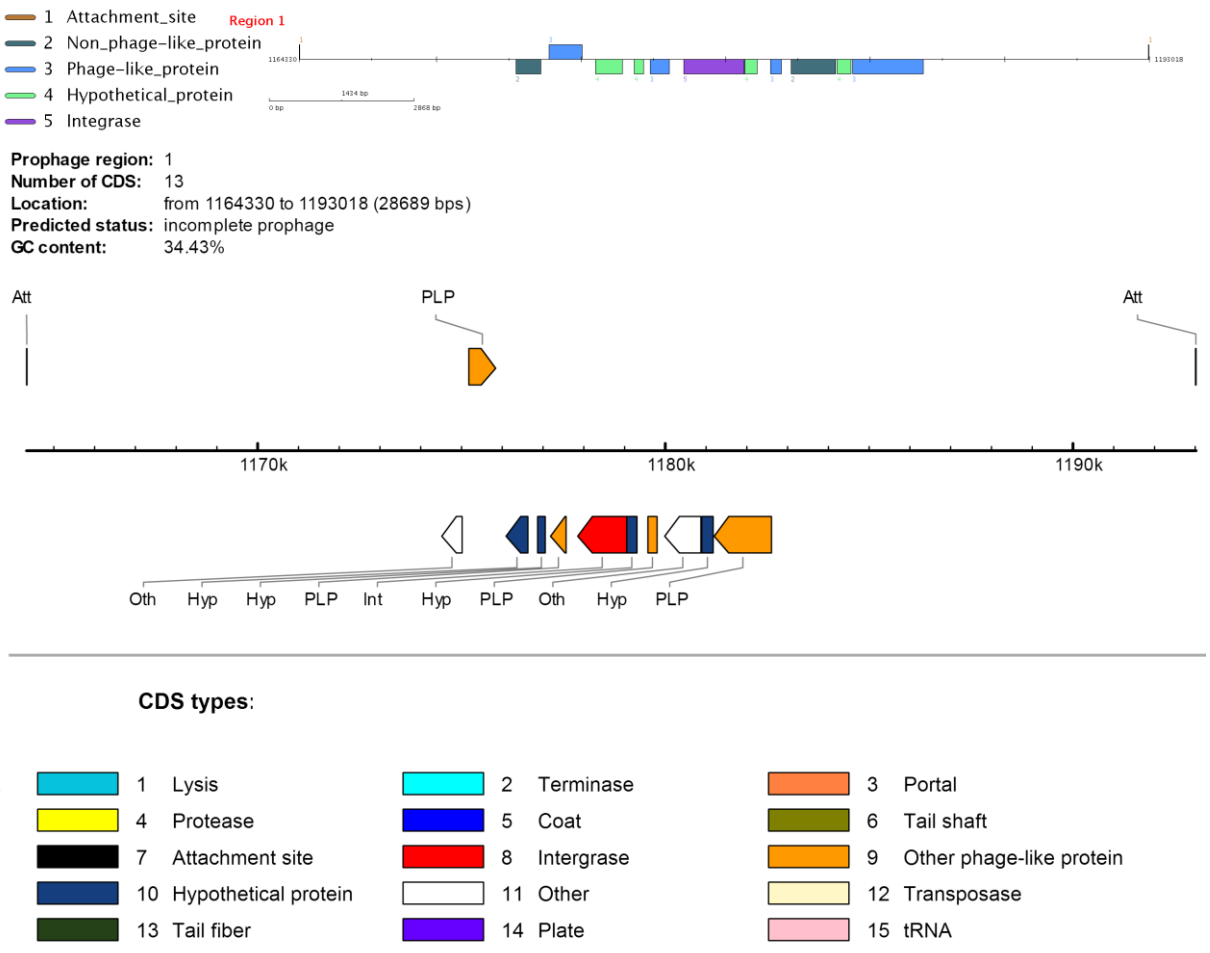


Figura D - Predição de Fago

1.3 Predição de Vias Metabólicas

Pathway Tools é um software de sistemas de biológicos que está associado com a uma extensa coleção de banco de dados genômicos e um dos mais completos de vias metabólicas o BioCyc. O software permite o desenvolvimento de base de dados específicos do organismo que integram variados tipos de dados incluindo genomas, vias metabólicas e regulatórias. Além do BioCyc o Pathway Tools também tem suporte ao BD KEGG (Kyoto Encyclopedia of Genes and Genomes) ele possui diversos esquemas de visualização dos mapas metabólicos e localização dos genes e onde seus produtos atuam (Kanehisa, 2000).

O *software* Pathway Tools pode ser encontrado no *website* *http://bioinformatics.ai.sri.com/ptools/*.

Através do software Pathway Tools, foi gerada uma visão geral do mapa metabólico da GBS85147. Nele foi predito 206 vias metabólicas, 1144 reações enzimáticas, 1919 polipeptídeos. Porém, esse é um resultado parcial, para uma visão mais profunda seria necessária uma curadoria manual das vias e das reações enzimáticas. Estudos de metaboloma são importantes para um entendimento mais abrangente sobre o funcionamento de sistemas biológicos, podendo ser aplicado na elaboração de estratégias terapêuticas e inovações biotecnológicas. Neste sentido, o estudo de enzimas possui grande relevância, uma vez que a descoberta da funcionalidade catalítica e predição de estrutura tridimensional destas biomoléculas são quesitos necessários para a compreensão de processos biológicos e até mesmo elaboração de proteínas sintéticas com importância econômica.

Figura E – Predição do Pathway Tools com GBS85147

Anexo 2: Publicação na revista Genome Announcements - Genome Sequence of Lactococcus lactis subsp. lactis NCDO 2118, a GABA-Producing Strain

genomeA
Journals.ASM.org

Genome Sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-Producing Strain

Letícia C. Oliveira,[a] Tessália D. L. Saraiva,[a] Siomar C. Soares,[a]* Rommel T. J. Ramos,[b] Pablo H. C. G. Sá,[b] Adriana R. Carneiro,[a] Fábio Miranda,[b] Matheus Freire,[b] Wendel Renan,[b] Alberto F. O. Júnior,[a] Anderson R. Santos,[a]* Anne C. Pinto,[a] Bianca M. Souza,[a] Camila P. Castro,[a] Carlos A. A. Diniz,[a] Clarissa S. Rocha,[a] Diego C. B. Mariano,[a] Edgar L. de Aguiar,[a] Edson L. Folador,[a] Eudes G. V. Barbosa,[a] Flavia F. Aburjaile,[a] Lucas A. Gonçalves,[a] Luís C. Guimarães,[a] Marcela Azevedo,[a] Pamela C. M. Agresti,[a] Renata F. Silva,[a] Sandeep Tiwari,[a] Sintia S. Almeida,[a] Syed S. Hassan,[a] Vanessa B. Pereira,[a] Vinicius A. C. Abreu,[a] Ulisses P. Pereira,[a]* Fernanda A. Dorella,[c] Alex F. Carvalho,[c] Felipe L. Pereira,[c] Carlos A. G. Leal,[c] Henrique C. P. Figueiredo,[c] Artur Silva,[b] Anderson Miyoshi,[a] Vasco Azevedo[a]

Laboratory of Cellular and Molecular Genetics, Institute of Biologic Sciences, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil[a]; Institute of Biologic Sciences, Federal University of Pará, Belém, PA, Brazil[b]; AQUACEN, National Reference Laboratory for Aquatic Animal Diseases, Ministry of Fisheries and Aquaculture, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil[c]

* Present address: Siomar C. Soares, AQUACEN, Federal University of Minas Gerais, Belo Horizonte, MG, Brazil; Anderson R. Santos, Federal University of Uberlândia, Minas Gerais, MG, Brazil; Ulisses P. Pereira, Federal University of Uberlândia, Minas Gerais, MG, Brazil.

***Lactococcus lactis* subsp. *lactis* NCDO 2118 is a nondairy lactic acid bacterium, a xylose fermenter, and a gamma-aminobutyric acid (GABA) producer isolated from frozen peas. Here, we report the complete genome sequence of *L. lactis* NCDO 2118, a strain with probiotic potential activity.**

Received 21 August 2014 **Accepted** 26 August 2014 **Published** 2 October 2014

Citation Oliveira LC, Saraiva TDL, Soares SC, Ramos RTJ, Sá PHCG, Carneiro AR, Miranda F, Freire M, Renan W, Júnior AFO, Santos AR, Pinto AC, Souza BM, Castro CP, Diniz CAA, Rocha CS, Mariano DCB, de Aguiar EL, Folador EL, Barbosa EGV, Aburjaile FF, Gonçalves LA, Guimarães LC, Azevedo M, Agresti PCM, Silva RF, Tiwari S, Almeida SS, Hassan SS, Pereira VB, Abreu VAC, Pereira UP, Dorella FA, Carvalho AF, Pereira FL, Leal CAG, Figueiredo HCP, Silva A, Miyoshi A, Azevedo V. 2014. Genome sequence of *Lactococcus lactis* subsp. *lactis* NCDO 2118, a GABA-producing strain. Genome Announc. 2(5):e00980-14. doi:10.1128/genomeA.00980-14.

Address correspondence to Vasco Azevedo, vasco@icb.ufmg.br.

Lactic acid bacteria (LAB), in general, acquire energy from the conversion of sugars into lactic acid (1) and are used for production of many fermented products, such as cheese, yogurt, butter, and wine. Food conservation is due to the medium acidification and production of molecules that inhibit the growth of undesirable microbiota, contributing to the development of desirable organoleptic properties in the final product (2). Moreover, some specific LAB strains produce bioactive molecules such as gamma-aminobutyric acid (GABA) (3), a product of glutamate decarboxylation by the glutamic acid decarboxylase (GAD) enzyme. Usually, GABA acts by modulating the central nervous system, contributing to smooth muscle relaxation and presenting hypotensor activity (4). Also, GABA can immunomodulate the immune system (5). Therefore, GABA-producing bacteria generally present probiotic properties (6). *Lactococcus lactis* NCDO 2118 is a nondairy strain, a xylose fermenter (a common trait of plant-associated strains), and a GABA producer isolated from frozen peas (6, 7).

L. lactis NCDO 2118 was sequenced three times, due to assembling complexity. First, the genome was decoded with the SOLiD 5500 platform with mate-paired libraries, generating a total of 5,133,057,360 bp, (coverage of 2,053 times). The reads were subjected to a Phred 20 quality filter using Quality Assessment software (8) and assembled with the CLC Genomics Workbench, generating a total of 1,641 overlapping sequences. These sequences were removed with the Simplifier (9), ordered and oriented based on the reference *L. lactis* KF147 genome sequence (a plant-associated strain, accession number CP001834). Then manual curation was performed using Artemis (10), and SSPACE (11) and Gapfiller (12) were used to generate the scaffold and resolve gaps, respectively. At the end of curation and sequence assembly, a total of 409 scaffolds (2,874,854 bp) were obtained.

L. lactis NCDO 2118 was then decoded with the Ion PGM platform with fragment libraries generating a total of 187,303,001 bp (coverage of ~71 times). Genome assembly was performed using Mira 3.9 (13), and the assembled genome sequence was reference aligned with CONTIGuator (14). The redundant overlapping sequences were removed with "in-house scripts," closing the remnant gaps. Annotation and frameshifts curation were then performed using Artemis and CLC, reducing the initial 1821 pseudogenes to 480.

Finally, the DNA was sequenced using the Ion Torrent PGM with fragment libraries, yielding a total of ~1,249,154,478 bp (coverage of 474 times). Assembly was performed with Mira 4.0.1 and Newbler 2.9 (15). We used CONTIGuator and FGAP 1.7 (16) to perform the alignment and gap closure steps, respectively. We followed the same previously explained steps for annotation and frameshift curation, reducing the pseudogenes to 52.

The complete genome of *L. lactis* NCDO 2118 consists of a single circular chromosome of 2,554,693 bp, containing 2,386 coding sequences (CDS), which had 52 pseudogenes, 66 tRNA genes, and 6 rRNA operons, with a G+C content of 34.9%. There is one plasmid, pNCDO2118 (37,571 bp), with 48 CDS, from which 4 are pseudogenes with a G+C content of 32.33%.

Nucleotide sequence accession numbers. The *Lactococcus lactis* NCDO 2118 chromosome and the plasmid were deposited at DDBJ/EMBL/GenBank under the accession numbers CP009054 and CP009055, respectively.

ACKNOWLEDGMENTS

This work was supported by Rede Paraense de Genômica e Proteômica, Ministério da Pesca e Aquicultura (MPA), Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). We also acknowledge the support from Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERENCES

1. **Carr FJ, Chill D, Maida N.** 2002. The lactic acid bacteria: a literature survey. Crit. Rev. Microbiol. **28:**281–370. http://dx.doi.org/10.1080/1040-840291046759.
2. **van de Guchte M, Ehrlich SD, Maguin E.** 2001. Production of growth-inhibiting factors by *Lactobacillus delbrueckii*. J. Appl. Microbiol. **91:**147–153. http://dx.doi.org/10.1046/j.1365-2672.2001.01369.x.
3. **Zareian M, Ebrahimpour A, Bakar FA, Mohamed AK, Forghani B, Ab-Kadir MS, Saari N.** 2012. A glutamic acid-producing lactic acid bacteria isolated from Malaysian fermented foods. Int. J. Mol. Sci. **13:**5482–5497. http://dx.doi.org/10.3390/ijms13055482.
4. **Inoue K, Shirai T, Ochiai H, Kasao M, Hayakawa K, Kimura M, Sansawa H.** 2003. Blood-pressure-lowering effect of a novel fermented milk containing gamma-aminobutyric acid (GABA) in mild hypertensives. Eur. J. Clin. Nutr. **57:**490–495. http://dx.doi.org/10.1038/ sj.ejcn.1601555.
5. **Jin Z, Mendu SK, Birnir B.** 2013. GABA is an effective immunomodulatory molecule. Amino Acids **45:**87–94. http://dx.doi.org/10.1007/s00726-011-1193-7.
6. **Mazzoli R, Pessione E, Dufour M, Laroute V, Giuffrida MG, Giunta C, Cocaign-Bousquet M, Loubière P.** 2010. Glutamate-induced metabolic changes in *Lactococcus lactis* NCDO 2118 during GABA production: combined transcriptomic and proteomic analysis. Amino Acids. **39:**727–737. http://dx.doi.org/10.1007/s00726-010-0507-5.
7. **Siezen RJ, Starrenburg MJ, Boekhorst J, Renckens B, Molenaar D, van Hylckama Vlieg JE.** 2008. Genome-scale genotype-phenotype matching of two *Lactococcus lactis* isolates from plants identifies mechanisms of adaptation to the plant niche. Appl. Environ. Microbiol. **74:**424–436. http://dx.doi.org/10.1128/AEM.01850-07.
8. **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 Res. Notes **4:**130. http://dx.doi.org/10.1186/1756-0500-4-130.
9. **Ramos RT, Carneiro AR, Azevedo V, Schneider MP, Barh D, Silva A.** 2012. Simplifier: a web tool to eliminate redundant NGS contigs. Bioinformation **8:**996–999. http://dx.doi.org/10.6026/97320630008996.
10. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B.** 2000. Artemis: sequence visualization and annotation. Bioinformatics **16:**944–945. http://dx.doi.org/10.1093/bioinformatics/16.10.944.
11. **Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W.** 2011. Scaffolding preassembled contigs using SSPACE. Bioinformatics **27:**578–579. http://dx.doi.org/10.1093/bioinformatics/btq683.
12. **Nadalin F, Vezzi F, Policriti A.** 2012. GapFiller: a *de novo* assembly approach to fill the gap within paired reads. BMC Bioinformatics **13:**S8. http://dx.doi.org/10.1186/1471-2105-13-S14-S8.
13. **Chevreux B, Wetter T, Suhai S.** 1999. Genome sequence assembly using trace signals and additional sequence information, p 45–56. Comput. Sci. Biol.: Proc. German Conference on Bioinformatics GCB'99 GCB. Hannover, Germany.
14. **Galardini M, Biondi EG, Bazzicalupo M, Mengoni A.** 2011. CONTIGuator: a bacterial genomes finishing tool for structural insights on draft genomes. Source Code Biol. Med. **6:**11. http://dx.doi.org/10.1186/1751-0473-6-11.
15. **Ohnishi N, Maruyama F, Ogawa H, Kachi H, Yamada S, Fujikura D, Nakagawa I, Hang'ombe MB, Thomas Y, Mweene AS, Higashi H.** 2014. Genome sequence of a *Bacillus anthracis* outbreak strain from Zambia, 2011. Genome Announc. **2**(2):e00116-14. http://dx.doi.org/10.1128/genomeA.00116-14.
16. **Piro VC, Faoro H, Weiss VA, Steffens MB, Pedrosa FO, Souza EM, Raittz RT.** 2014. FGAP: an automated gap closing tool. BMC Res. Notes **7:**371. http://dx.doi.org/10.1186/1756-0500-7-371.

Anexo 3: Capítulo de livro publicado SMGroup - Bioinformatics

SMGroup

Title: A Textbook of Biotechnology

Editor: Zahoorullah S MD

Published by SM Online Publishers LLC

ISBN: 978-0-9962745-3-1

First Published April, 2015

Online Edition available at www.smgebooks.com

SMGroup

Bioinformatics

Alberto Oliveira[1], Leandro Benevides[1], Diego Mariano[1], Edgar Aguiar[1], Thiago Sousa[1], Artur Silva[2] and Vasco Azevedo[1]*

[1]Universidade Federal de Minas Gerais/Instituto de Ciências Biológicas.

[2]Universidade Federal do Pará/ Instituto de Ciências Biológicas.

***Corresponding author:** Vasco Azevedo, Universidade Federal de Minas Gerais/Instituto de Ciências Biológicas, Minas Gerais, Brazil, Email: vasco@icb.ufmg.br

Published Date: April 15, 2015

How the study of nucleotide sequences, amino acids as well as understand the function of these biomolecules, until the three-dimensional structures level, influence in biotechnological processes? Bioinformatics can help.

INTRODUCTION

In recent times has seen a range of innovative studies that has revolutionized the biological sciences as had never seen before. Studies of complete genome sequencing of pathogenic organisms has allowed understanding the virulence mechanisms and helped in the production of new drugs and the development of strategies to contain outbreaks [1-3]; studies of the monitoring of RNA transcription level and regulation of networks has helped to understand different types of cancer [4]; protein folding studies have been correlated with mutations in the DNA, and helped researchers have new insights into the functioning of diseases previously little understood [5]; in vivo experiments with the aid of new technologies have helped to understand damage to the spinal column and allowed to paralyzed limbs can be moved again [6].

These studies were made possible due to the evolution of information technology (IT), evidenced in the improving of software (algorithms and implementation techniques of parallel

processes), and increasing processing power and storage information (improvement of hardware), which in recent times has allowed the storage and processing of huge amounts of data, the "big data" [7]. And with the race to sequencing and mapping of the human genome, new and creative DNA sequencing techniques have contributed to the appearance of robust sequencing platforms, known as Next-Generation Sequencers (NGS), which allowed the sequencing large-scale, with both cost and reduced time. All these factors led to the rise of a new field of study, linking the biological sciences to computer science, called bioinformatics [8].

Bioinformatics is an area of research between biology, mainly molecular biology, and computer science, but also covers other areas such as physics, chemistry, information science, engineering, mathematics and statistics. Bioinformatics can be understood as applying computational techniques for storage, processing and analysis of biological data in large scale. However, it should not be seen only as molecular biology tool, but as a science that has means to elucidate pathways for biological experiments through data analysis, and thus promote prospects for new discoveries.

One of the bioinformatics fundamentals is the modeling of a particular biological problem by representing it so that it can be computationally treated. Thus, beyond the terms *in vivo* and *in vitro*, a new expression arises: *in silico*, representing experiments by computational simulation.

BIOINFORMATICS SOFTWARE

Computational experiments are performed using software, which are built with programming languages following a particular algorithm.

Software development is an important tool for bioinformatics. To know some programming language can be a differentiator for the professionals in this area. Recently, several groups of developers have provided libraries to assist the development of code for bioinformatics in various programming languages. Libraries such as BioPerl, BioJava, BioJS, BioRuby, BioPHP and BioPython have facilitated the development process of new biological software. However, it is not essential the software programming knowledge in this area, after all there are many biological software already developed for various platforms, especially for Linux operating systems [8]. Thus, the field of use of operating systems and specific software to your search area become of fundamental value for bioinformaticians.

Software: logical part of a computer; set of instructions to be interpreted and executed by a processor; the operating system and applications.
Algorithm: computational process that follows a set of actions to perform a specific task in a finite time interval; can be understood as the step-by-step to solving a problem computationally treated.
Hardware: physical part of computers; processors, plates and all electronic components in and out.
Operating System: range of programs responsible for linking the software and the hardware, ensuring operability of the system and allowing the user to get access to computer resources. Currently the most widely used operating systems are Windows, OS X, Linux, Android and IOS, the last two being exclusive both to smartphones and tablets.
Cloud computing: the use of the services by Internet through virtualized resources on remote servers [24].
Web browser: computer program that allows the user to access websites. Examples of web browsers: Firefox, Chrome and Internet Explorer.

However, with the recent strong adoption of cloud computing by the developer community, allowed the improvement in the quality of internet connection, users could to work on collaborative projects most easily through the World Wide Web. This phenomenon positively affected the

manner in which software for bioinformatics are designed. The adoption of the Web has brought greater usability and interoperability for applications of bioinformatics, allowing software could be easily accessed on any operating system through a web browser.

The development of Web tools for bioinformatics has become popular due to low learning curve of Web programming languages (such as PHP, Python and JavaScript), the fast and easy way to access the software without the need to install through web browsers, the ease in performing cooperative work in parallel at distance, and especially the adoption of user-friendly interfaces. Thus, several research groups started to provide software implementation services for the World Wide Web, also allowing easy access to their databases [8]. Therefore, the large amount of information stored in databases could be accessed over the Web through electronic sites.

Note that the concept of database is very important for bioinformatics. But what is a database and what is its importance in representing biological problems?

WHAT IS DATABASE?

Databases are **data** sets organized in order to present comprehensible **information**. Thus, biological databases are data sets derived from biological experiments organized so that you can extract useful information through data mining techniques.

One of the first records of protein storage in databases occurred in the 60s, by Margaret Dayhoff et al. from the National Biomedical Research Foundation (NBRF). They have proposed the use of a substitution matrix, PAM (Point mutations accepted), to calculate the conservation of protein sequence and the mutation probability of a given amino acid. The methods proposed by Dayhoff et al. used a character (letter) to represent each amino acid, allowing a reduction in stored data to describe amino acid sequences [9]. However, biological databases are not just used for representation of proteins (proteomics field), but also in the fields of genomics (databases of genomic sequences) and transcriptomics (transcribed databases).

A biological database example is the PDB (Protein Data Bank). The PDB (http://www.wwpdb.org/) stores information structural proteins, such as the spatial coordinates of atoms. There are databases that relate ontologies of information between genes and proteins as the Gene Ontology (www.geneontology.org/). Another type of database is the KEGG (http://www.genome.jp/kegg/) that stores the biochemical interactions and metabolic pathways, in addition there is another database with information of the metabolic pathways that were cured (manually reviewed and corrected) the MetaCyc (http://www.metacyc.org/). It is possible to cite databases of networks RNA transcription regulation such as CMRegNet (http://lgcm.icb.ufmg.br/cmregnet).

A good computing practice is the management of databases through DBMS (Database Management Systems) that, in general, allow queries and data modification using SQL commands (Structured Query Language). However, many biological databases chose to use simple storage systems, such as text files, in order to facilitate the manipulation and analysis of these data

using word processing tools written in programming languages. One example is the sequences databases.

Nucleic acid sequence databases are stored in different formats by three major institutes: NCBI (National Center for Biotechnology Information of the US National Library of Medicine), DNA Data Bank of Japan (National Institute of Genetics, Japan) and EMBL Data Library (European Bioinformatics Institute, UK). Another example is the UniProtKB, knowledgebase of proteins from the UniProt (Universal Protein Resource), that storing the known protein sequences. The UniProtKB is the union of two protein databases: Swiss-Prot and TrEMBL, and can be accessed through the Website UniProt (http://www.uniprot.org).

The patterns of storing of sequences were developed with the aim of speeding up information retrieval methods. Thus, storage patterns are directly related to the search algorithms of information, which will be presented below, and represent an important area of bioinformatics: sequence analyzes.

SEQUENCE ALIGNMENT

Sequence alignment is an important step toward structural and functional analysis of sequences. This is the method comparison of two (pair-wise alignment) or more (multiple sequence alignment) sequences in the searching for common character patterns and establishing residue–residue correspondence among related sequences.

Input Files (FASTA)

The main input file currently use in most sequence analysis is the fasta format [10]. This is a simple format broadly used when dealing nucleotides and amino acids sequences. The main characteristic which compose this file are the **header** and the **sequence**. Below is showed a example of this file.

>Header

S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S
E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E NC E S E Q U
E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C
E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E Q U E N C E S E
Q U E N C E S E Q U E N C E S E Q U E N C E

>Streptococcus agalactiae - dnaA Chromosomal replication initiator protein DnaA forward

A T G G T A C A A T A T A A C A A T A A T T A T C C A C A A G A C A A T A A G G A A G A A G C T
A T G A C G G A A A A C G A A C A A C T A T T T T G G A A T A G A G T A C T A G A G C T A T C T C G
T T T C A A A T A G C A C C A G C A G C T T A T G A A T T T T T T G T T C T A G A G G C T A G A C T C
C T C A A A A T T G A A C A T C A A A C T G C A G T T A T T A C T T T A G A T A A C A T T G A A A T
G A A A A A G C T A T T C T G G G A A C A A A A T T T G G G G C C T G T T A T C C G T C T A G T C A

C G C G C T T T A G T T G G G G A C T C C C A G T A A A T A T C A C A C C A C C

This file can be represented by many others characters, as:

- **"-"** represent gaps, used when is made alignments in search of evolutionary patterns;
- **"."** can indicate other element in sequence, like heteroatoms in protein structures, for example water or ions;
- **"N"** Any nucleotide or amino acid;
- "\\" can indicates the end of the alignment, but this is not a common symbol;
- **"*"** indicate the translation stop;
- **"Upper and lower"** can be used to discriminate strong and weakly conserved residues respectively.

Sequences are expected to be represented in the standard IUB/IUPAC nucleic acid and amino acid codes (IUPAC), like:

Nucleic Acid Symbol	Meaning
A	Adenine
T	**T**hymine
G	Guanine
C	Cytosine
U	Uracil
R	G or A (pu**R**ine)
Y	C, T or U (pYrimidine)
K	G, T or U (**K**etone)
M	A C (aMino groups)
S	G or C (**S**trong interaction)
W	A, T or U (Weak interaction)
B	not A (i.e. C, G, T or U) - **B** comes after A
D	not C (i.e. A, G, T or U) - D comes after C
H	not G (i.e., A, C, T or U) - **H** comes after G
V	H comes after G - V comes after U

The codes supported for 24 amino acids and three special codes are:

Amino Acid Symbol	Meaning
Z	Glutamic acid (E) or Glutamine (Q)
Y	Tyrosine
X	any
W	Tryptophan
V	Valine
U	Selenocysteine
T	Threonine
S	Serine
R	Arginine
Q	Glutamine
P	Proline
O	Pyrrolysine
N	Asparagine
M	Methionine
L	Leucine
K	Lysine
J	Leucine (L) or Isoleucine (I)
I	Isoleucine
H	Histidine
G	Glycine
F	Phenylalanine
E	Glutamic acid
D	Aspartic acid
C	Cysteine
B	Aspartic acid (D) or Asparagine (N)
A	Alanine
-	gap of indeterminate length
*	translation stop

There is no standard file extension for a text file containing sequences in format FASTA. The table below shows each extension and its respective meaning.

Extension	Meaning	Notes
fasta (.fas)	generic fasta	Any generic fasta file. Other extensions can be fa, seq, fsa
fna	fasta nucleic acid	Used to generically specify nucleic acids.
ffn	FASTA nucleotide coding regions	Contains coding regions for a genome.
faa	fasta amino acid	Contains amino acids. A multiple protein fasta file can have the more specific extension mpfa.
frn	FASTA non-coding RNA	Contains non-coding RNA regions for a genome, in DNA alphabet e.g. tRNA, rRNA.

Alignment Local and Global

The sequence alignment is a tool used in bioinformatics with the objective to identify regions of similarity or identity that may be a consequence of functional, structural or evolutionary relationships between the sequences. There are two ways to do this analysis, using local or global alignments.

The basic difference between local and global alignments [11] is that in a local alignment, you try to match your query with a substring (a portion of your sequence) of your subject (reference sequence). Whereas in global alignment you perform an end to end alignment with the subject.

Local Alignment

5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'

|||||||||||||||||||||||||

5' TACTCACGGATGAGGTACTTTAGAGGC 3'

Global Alignment

5' ACTACTAGACTACTTACGGGTCAGTTACTTTAGAGGCTTACAACCA 3'

||||||||||||||||||||||||||||||||||||||

5' ACTACTAGACT----ACGGGTC--TTACTTTAGAGGCTAACAACCA 3'

Suppose you are searching in a database some information about your sequence. This sequence has, approximately, 300 bp. The database in which you are doing this search has a large reference, around 2000 bp. When is done a global alignment, in the result is demonstrate that you sequence is probably involve in alternative splice - specific event from eucariotic cells. However, when is done a local alignment, i.e. using just some fragments from your sequence, the result is more precision, showing that your sequence align in a specific region featuring one specific organism.

Multiple Sequence Alignment (MSA)

Multiple sequence alignment may be considered as an extension of pairwise alignment, which allow the alignment multiple related sequences to obtain the best matching of the sequences, and thus reveals more biological information. For example, it allows the identification of conserved sequence patterns and motifs in the whole sequence family, which are not obvious to detect by

comparing only two sequences. MSA presents several computational challenges. First, obtain an optimal alignment of several sequences that includes matches, mismatches, and gaps, and also that take into account the degree of variation in all of the sequences at the same time. Due to that, alternative methods had to be developed to speed up the calculations for multiple sequence alignment, including: (1) progressive alignment of the sequences, (2) iterative alignment, (3) alignments based on locally conserved patterns found in the same order in the sequences, and (4) use of statistical methods and probabilistic models of the sequences [12]. The automated alignment often contains misaligned regions that should be edit manually. It is need to check the alignment to correct obvious alignment errors and to remove sections of dubious quality. This involves introducing or removing gaps to maximize biologically meaningful matches. In manual editing, empirical evidence or mere experience is needed to make corrections on an alignment. Manual editing may be performed on the amino acid level for protein coding sequences. In this way, information about protein domain structure, secondary structure, or amino acid physicochemical properties can be taken into consideration [13].

ADVANCED BIOINFORMATICS

In this section we will address the theme "Advanced Bioinformatics" as a specific part, focused in evolutionary studies and structural analyses from biomolecules.

Phylogenetic and Molecular Evolution

The phylogeny arose from the human need to organize and classify the world around them in order to facilitate understanding and communication. Different systems were proposed for classifying organisms, until when, in XIX century, Charles Darwin, inspired in phenotypic variation of finches in the Galapagos Islands, developed his theory of evolution. Based on this theory, the classification of organisms has become not only natural, but also to present a prerequisite for common ancestry. According to this thinking, the organisms derived from each other, since the emergence of life on earth. Darwin represented this pattern through a branching scheme, where the branches represent the time between the ancestral and the new organism, and the nodes represents the organisms themselves. Later, this would be the first phylogenetic tree used to represent evolutionary processes [14]. Nowadays, the evolutionary analysis of ancestry is based on molecular data thanks to the advent of sequencing methods. Similarities and divergence among related biological sequences revealed by sequence alignment often have to be rationalized and visualized in the context of phylogenetic trees. Thus, molecular phylogenetics is a fundamental aspect of bioinformatics [9].

Phylogenetic trees

The evolutionary relationships among genes and organisms can be illustrated using a two-dimensional graph, called phylogeny or phylogenetic tree, which is comparable to a genealogy and shows which genes or organisms are most closely related. The phylogenetic trees receive terms

referring to the various parts of real trees (i.e. root, branch, nodes, and leaf). Terminal (external) nodes or leaves represents the existing taxa and are frequently called operational taxonomic units (OTUs). The trees can be drawn in cladogram or a phylogram. In a phylogram, the branch lengths represent the amount of evolutionary divergence. These kinds of trees are scaled and show the evolutionary relationships and information about the relative divergence time of the branches. In a cladogram, however, the terminal nodes lines up in a row. Their branch lengths are not proportional to the number of evolutionary changes and thus have no phylogenetic meaning. Furthermore, the phylogenetic trees can be rooted or unrooted. The main difference between them is that in the rooted phylogenetic trees exist one outgroup that represents the ancestor of all OTUs and is possible infer the direction of evolutionary process. The unrooted phylogenetic trees only show the positions of the taxa and their relative relationships [15].

Methods for phylogenies inference

The main objective of molecular phylogenetics is to correctly reconstruct the evolutionary history based on the observed sequence divergence between organisms. The first step to constructing phylogenetic trees is alignment of sequences under study. After the alignment has been proposed various methods can be used to estimate the phylogeny of the sequences studied [14]. There is three main methods to find the evolutionary trees: neighbor-joining (NJ), maximum parsimony (MP) and maximum likelihood (ML).

Neighbor-joining

Neighbor-joining is a distance method that employs the number of changes between each pair in a group of sequences to finding pairs of neighbors and thus produces a phylogenetic tree of this group. The NJ algorithm starts with a totally unresolved tree (star topology). The algorithm sequentially, based on a matrix of distances between all pairs of sequences, identifies the pair that presents the shortest distance, links this pair by a node (representing the common ancestor of the pair of sequences) and incorporates into the tree. The distances of each pair of sequence are recalculated for the new node, as well as the distances of all other sequences are recalculated for the new node. This process is then repeated replacing the pair of neighbors united by the new node and using the distances calculated in the previous step [12].

Maximum parsimony

The maximum parsimony is a qualitative method that assumes that within a range of phylogeny, that The maximum parsimony is a qualitative method that assumes that within a range of phylogeny, that phylogeny presents the lower number of evolutionary events (substitutions) should be the most probable to explain the alignment data. For each aligned position, the number of evolutionary changes to produce the observed sequence changes is calculated and the phylogenetic trees that require the smallest number are identified. This analysis continues for every position in the sequence alignment until the tree that requires the minimum number of changes is identified. Although fast, the parsimony methods fail to estimate the evolutionary

relationship between a large number of sequences or sequences with a large amount of variation [14].

Maximum likelihood

Maximum likelihood (ML) is similar to the MP method and combines the alignment information to a statistical model able to better handle the probability of change of a nucleotide to another. Based on a certain evolutionary model, and assuming that each alignment of the site evolves so independent of the others, the likelihood for all possible nucleotide (or amino acid) states in the ancestral (internal) nodes is calculated. Because these likelihoods are very small numbers, their logarithms are usually added to give the logarithm likelihood of each tree. The most likely tree given the data is then identified. The main disadvantage of maximum likelihood methods is that they are computationally intense. However, with faster computers, the maximum likelihood method is seeing wider use and is being used for more complex models of evolution [14].

STRUCTURAL BIOINFORMATICS

The structural bioinformatics is summed up like the study of three-dimensional (3D) structures from biomolecules. In General, this area focuses in understanding the function, interactions and the molecular organization of proteins e some RNAs. The bioinformatics is able to analyze these molecules using a large arsenal of tool and thus understand the prediction of three-dimensional structures, i.e. proteins, that has been characterized with great impact in practical therapeutic applications and biotechnology.

Molecular Modeling

Understanding the physicochemical characteristics of the proteins that interact (for example in a network) and how an interaction is established at the molecular level, can provide a biological understanding of the cells of organisms. The bioinformatics, also known by *In silico* methods, can then be used to investigate the interactions and predict how these molecules might interact at the atomic level. Some computational methods are wieldy to create these structures, referred to as template-based modeling, that includes both the **threading** techniques that return a full 3D description for the target and the **comparative** modeling [16].

Comparative

Currently, much of the three-dimensional structures are obtained by comparative modeling methods, which follow the Protocol held in the study of Sali (1997). This method is based on the use of homologous proteins for the prediction structures of sequences, thus the structure models are constructed from the residuals of the structure template that are aligned to the target sequence in the sequence comparison. Therefore, the quality of this alignment thus is critical for the accuracy achievable. The protocol (figure 1) pointed out four fundamental steps for the creation of the models: (i) identification of homologous structures that could be used as template for modeling; (ii) alignment of the target sequence with the template sequences; (iii) construction

of the model for the target based on the information of the alignment; and (iv) model validation [17].

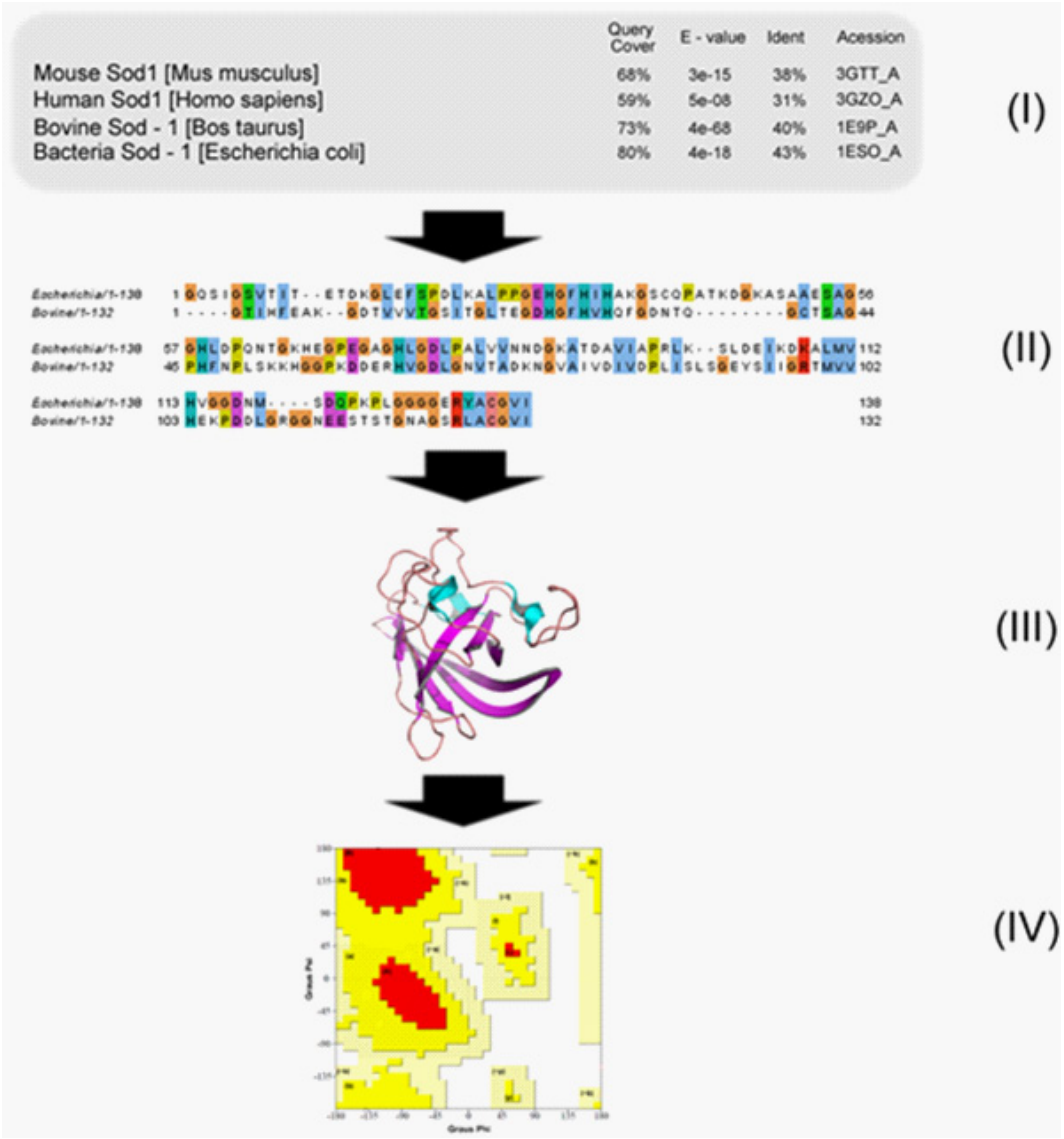


Figure 1: The comparative molecular modeling protocol scheme.

The important of the template-based model comes from template selection and sequence-template alignment, in addition to the difference between the structure and sequence template. It is essential the sequences have good values of identity, wherein it is expected that minimum values are around 30%. The step IV, about the validation, will be explained in the next subtopic.

Threading

When sequence identity is below 30%, it is complex to recognize the better template and generate accurate sequence-template alignments, so the resultant models have a wide range of accuracies. In general, proteins with low values of identity are said to be non-homologous with known proteins. Considering that currently there are several proteins without experimental structures, even a rapid improvement in the accuracy of obtaining the model can have a significant impact on the prediction of large-scale structure and its applications. Given this difficulty, the protein threading method (Figure 2), also known as fold recognition, is a method of protein modeling that analyzes the folding of proteins with others thus is able to create models, i.e. this method uses the secondary structure (or little fragments of your target sequence) searches the same region in database of proteins for the prediction models [18].

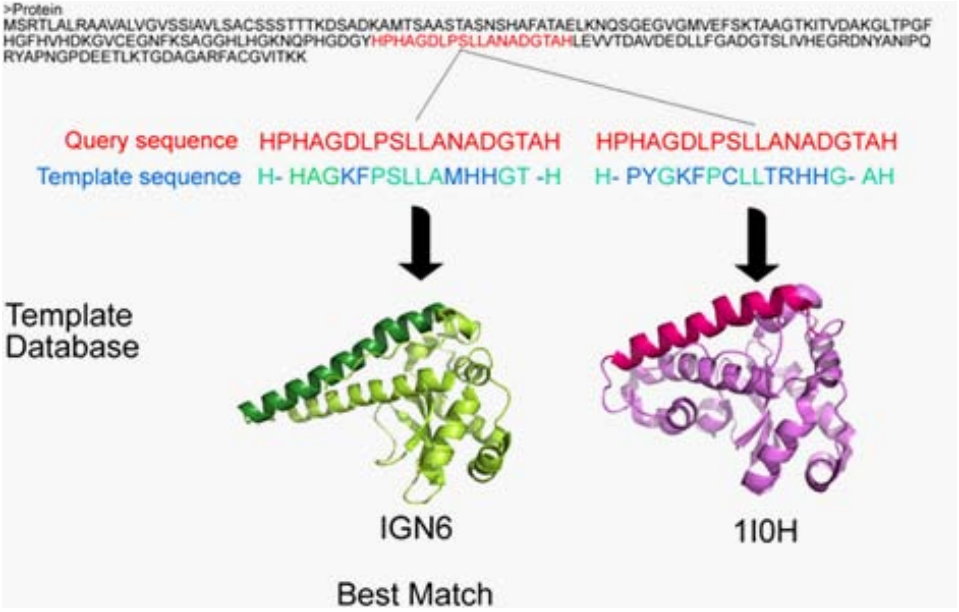


Figure 2: Modeling protocol scheme for threading

Ab Initio

Prediction of the spatial conformation of a protein from its primary structure is the one of the most important open problems in molecular biology, i.e. from its sequence of amino acids. Therefore, when no suitable structure templates can be found, Ab Initio methods can be used to predict the protein structure from the sequence information only. This is essential for a complete solution to the protein structure prediction problem; it can also help us understand the physicochemical principle of how proteins fold in nature. Currently, the accuracy of ab initio modeling (Figure 3) is low and the success is limited to small proteins (<100 residues) [19].

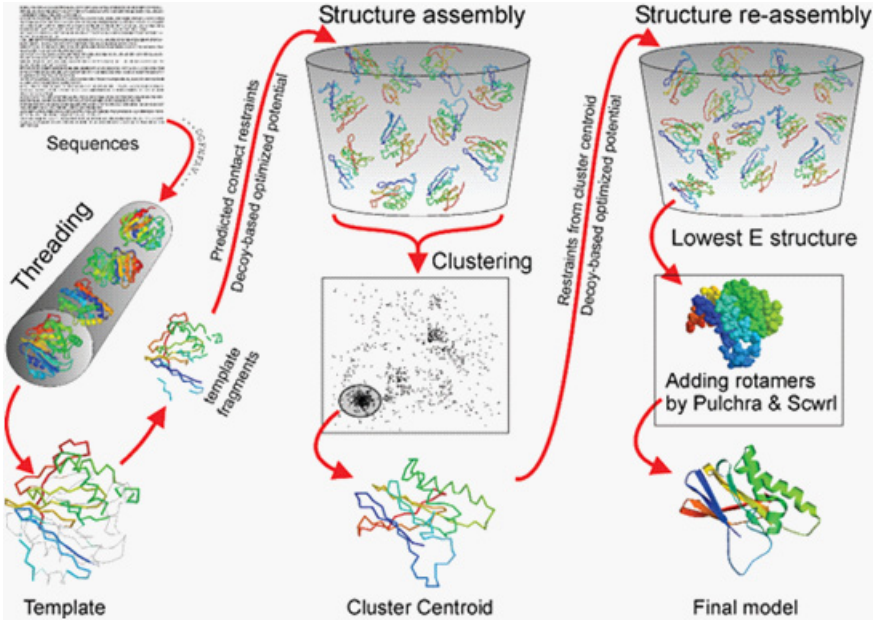


Figure 3: Here is a scheme from I-TASSER server. An internet service for protein structure and function predictions. It allows academic users to automatically generate high-quality predictions of 3D structure and biological function of protein molecules from their amino acid sequences [20].

Validations of Tree-Dimensional Models

Suppose we have two alternative conformations of some any protein, and we want analyze how much similar they are. For example, I have my model structure from molecular modeling, made by comparative modeling methods, and I need validate the conformation of this model. This validation can be made using a distance matrix or known by "RMSD", in which the term mean Root-Mean-Square-Deviation [21]. One protein is typically represented by its virtual Cα atom chain of n residues or points. For a quantitative single-number measure of structural similarity between structures A and B:

$$D^2_{dis}(A,B) = (n(n-1)/2)^{-1} (d_{aij} - d_{bij})^2$$

Where *daij* and *dbij* are the correspondence distances between the ith and jth atoms or the "coordinate RMSD" after optimal rigid body superposition. Currently, many scientists made use of this method to align two or more protein structure to validate your spatial conformation. Thus, the root mean squared errors (deviations) function is defined as follows:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \delta_i^2}$$

Where δ is the distance between N pairs of equivalent atoms (usually $C\alpha$ and sometimes $C,N,O,C\beta$).

Other validation method is Ramachandran plot [22,23] or φ, ψ-plot that has remained nearly unchanged in the ensuing fifty years and continues to be an integral tool for protein structure research. His analysis is focused in the secondary structure of a protein that can be described by the torsion angles Φ (Phi) and Ψ (Psi) on which are repeated for each amino acid along the polypeptide chain. The allowed values of these angles, or acceptable to three-dimensional structures of proteins, are demonstrated at the Ramachandran plot (figure 4). This graph is based on analysis of 118 structures with a resolution of 2 angstroms (Å). For a good model considered, it is expected that more than 90% of the amino acid residues are in the most favorable regions.

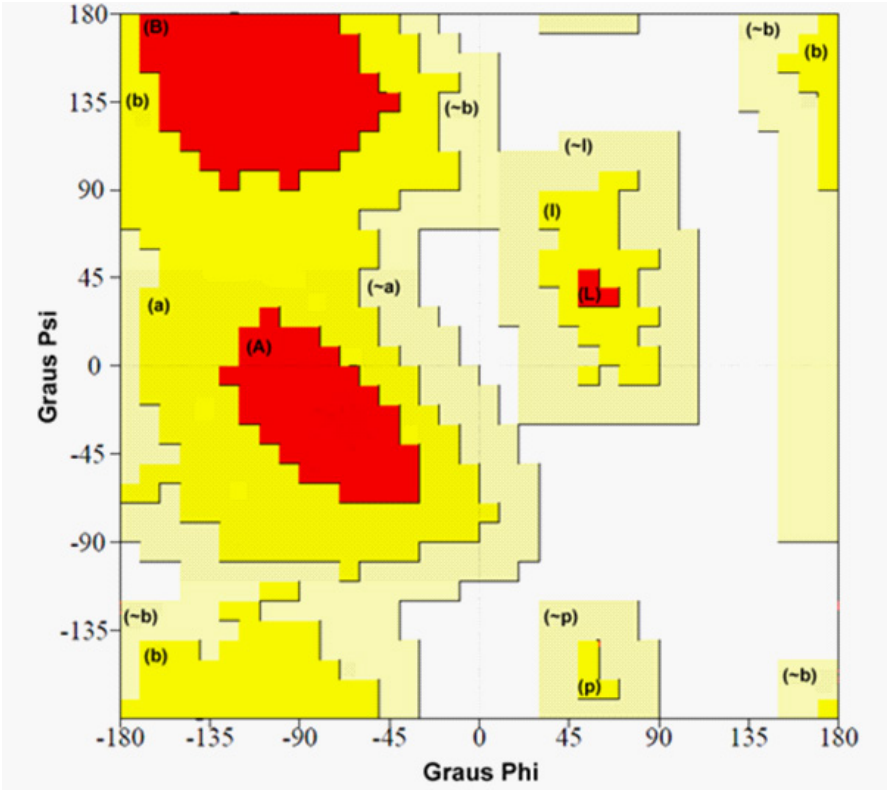


Figure 4: Ramachandran plot representation. The allowed regions are shown for amino acids: (A) more favorable regions of alpha *helix* (*α*-helix); (a) α-helix regions additionally permitted; (~a) α-helix regions generously allowed propellers; (B) β-sheet regions more favorable; (b) β-sheet regions further allowed; (~b) β-sheet generously allowed; (L) α-helix regions of left hand, more favorable; (I) α-helix regions of left hand propeller additionally allowed; (~i) α-helix regions of left hand propellers, generously allowed; (p) glycine regions additionally allowed; (~p) glycine regions generously allowed.

Box: Applications of bioinformatics in Biotechnology

The virtual screening (VS) techniques, through the use of computational methods, also known by *in silico* term, can help on search of organic compounds as promising therapeutic target ligands of interest, whether receptor agonists or antagonists, or enzyme inhibitors. The VS comprises two principal approach: The target structure-based virtual screening (I) and the screening ligand-based (II).

(I) This method considers the three-dimensional structure (3D) the therapeutic target, using the docking calculations as main strategy for selection of potential ligands with chemical characteristics, electronics and structural that promote interactions with the ligand site of the molecular target.

(II) Since the strategies ligand-based using organic molecules with biological activity known, working as molds for the screening in databases of new chemical entities with some level of similarity, sharing with these molds the same biological activity.

Thus, the main objective of a virtual screening is to identify the compounds of a library that are most likely that bind to the biological target of interest.

These targets may be parasites molecules or specific diseases. Since the ligands may be any small molecule found in nature, for example in plants. Thus, in environments with rich biodiversity vegetative can be focused to large production plants collections from biotechnological methods. Biotechnology is inserted in this context in order to study such plants with the aim to extract, purify and identify chemicals (small molecules) by physical-chemical analyzes with bioactive properties from them. Finally, this small molecules can be used in VS to search candidates therapeutic target against diseases or others biological problems.

BIOINFORMATICS LIMITS

Bioinformatics has been highlighting for some years as a powerful tool capable of providing a better and faster understanding of biological problems at the molecular level. However, it is limited by the frontiers of scientific computing, which is unable to show the entire biological setting complex.

To understand the advances in computational field and how they are directly related to the rise of bioinformatics, we can cite Moore's law. In 1965, Gordon Earl Moore, co-founder of Intel, stipulated that the processing power of a microchip would double every eighteen months, while in the same period, the price would drop by half. This prediction has had as a marketing strategy, but stimulated competition among technology companies, and in parallel, also served as a stimulus for the evolution of the computing power of the hardware, which in turn, allowed the emergence of bioinformatics. . However, it is important to understand the fine line between the data processing capacity and obtaining data by biological experiments, being one dependent on the other. Thus, to obtain good results *in silico*, it is necessary that biological experiments have produced good data.

An example of limitation found in bioinformatics is the prediction of protein structures. For these biomolecules, tend to lose their native three-dimensional structure easily to leave their ideal storage conditions. Thus, a structural modeling is intended to ensure the native conformation based on their primary amino acid sequence. First, think to try all possible conformations to get in the native structure of the protein, right? No, it is approach cannot be applied and this is exactly what Cyrus Levinhal affirmed in 1969, where states that if a protein was folding sequentially trying all possible conformations, it would take a period of time equivalent to the age of the universe, and this is at millisecond intervals under natural conditions. Based on this observation, the proteins shall directed by folding and orderly process. Currently, known to that the information necessary for the protein folding is contained in the amino acid sequence itself. Moreover, that

protein folding go through several alternative configurations with intermediate power stages, arriving in a lower level of state that would be the native state, like it was a map funnel-shaped. However, more research are still necessary to fully realize these *in silico* experiments. In short, it is still not possible to predict protein folding altogether, then we come in the next issue, we still lack computing power to simulate these conformations or are using the wrong approach?

The protein interactions are fundamental to almost all maintenance processes of integrity of cell. For example, when is made structural analysis for protein, using *in silico* methods, is lost the environment signalization; is lost the environment biological. Actually, the bioinformatics is not able to reproduce the same biological event. Due this is not possible get a real confirmation about the all structure analyses results.

Another big challenge of bioinformatics is the sequencing field and genome assembly. Sequencing platforms have gone through many evolutions in the last decades have enabled complete sequencing of genomes, reducing the time and expense thereof. However, the genome assembly process is directly impacted by sequencing, and this has computational cost determined by the complexity of the genome sequenced (Genome size and number of repetitive regions), the performance and quality of the sequencing platform used. However, the assembly of complex genomes is possible, but still requires a long time and has an extremely high cost.

The challenges of bioinformatics are still many and only time will tell if it will exceed its limits.

References

1. Gardy JL, Johnston JC, Ho Sui SJ, Cook VJ, Shah L. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med. 2011; 364: 730-739.
2. Lewis T, Loman NJ, Bingle L, Jumaa P, Weinstock GM. High-throughput whole-genome sequencing to dissect the epidemiology of Acinetobacter baumannii isolates from a hospital outbreak. J Hosp Infect. 2010; 75: 37-41.
3. Alexander Mellmann, Dag Harmsen, Craig A Cummings, Emily B Zentz, Shana R Leopold, et al. Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. PLoS ONE. 2011.
4. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009; 10: 57-63.
5. Roth DM, Hutt DM, Tong J, Bouchecareilh M, Wang N. Modulation of the maladaptive stress response to manage diseases of protein folding. PLoS Biol. 2014; 12: e1001998.
6. Shanechi MM, Hu RC, Williams ZM. A cortical-spinal prosthesis for targeted limb movement in paralysed primate avatars. Nat Commun. 2014; 5: 3237.
7. FRANKS Bill. Taming The Big Data Tidal Wave: Finding Opportunities in Huge Data Streams with Advanced Analytics. New Jersey: John Wiley & Sons, Inc. 2012.
8. Arthur M Lesk. Introdução à Bioinformática (Tradução Ardala Elisa Breda Andrade, et al.). Oxford: Oxford University Press. 2008.
9. David W Mount. Bioinformatics: sequence and genome analysis. New York: Cold Spring Harbor Laboratory Press. 2001.
10. Tao Tao. Single Letter Codes for Nucleotides. NCBI Learning Center. National Center for Biotechnology Information. Retrieved 2012-03-15.
11. Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms Mol Biol. 2011; 6: 25.
12. Jin Xiong. Essential Bioinformatics. New York: Cambridge University Press. 2006.
13. Hugo Verli. Bioinformática da Biologia à flexibilidade molecular. Porto Alegre: Bioinfo UFRGS. 2014.

14. Sergio Matioli, Flora Fernandes. Biologia Molecular e Evolução. 2nd edn. Brazil: Holos press. 2012.

15. Philippe Lemey, Marco Salemi, Anne-Mieke Vandamme. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambrigde: Cambridge University Press. 2009.

16. Fiser A. Template-based protein structure modeling. Methods Mol Biol. 2010; 673: 73-94.

17. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D. Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci. 2007; Chapter 2: Unit 2.

18. Peng J, Xu J. Low-homology protein threading. Bioinformatics. 2010; 26: i294-300.

19. Jooyoung Lee, Sitao Wu, Yang Zhang. Ab Initio Protein Structure Prediction. In: Daniel John Rigden, editor. Houten: Springer Netherlands. 2009; 3-25.

20. Yang J, Yan R, Roy A, Xu D, Poisson J. The I-TASSER Suite: protein structure and function prediction. Nat Methods. 2014; 12: 7-8.

21. Maiorov VN, Crippen GM. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. J Mol Biol. 1994; 235: 625-634.

22. ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol. 1963; 7: 95-99.

23. Hovmöller S, Zhou T, Ohlson T. Conformations of amino acids in proteins. Acta Crystallogr D Biol Crystallogr. 2002; 58: 768-776.

24. Borko Furht, Armando Escalante. Handbook of Cloud Computing. New York: Springer Science. 2010.

Anexo 4: Artigo submetido a revista Microbial Cell Factory **-** Putative Virulence Factors of *Corynebacterium pseudotuberculosis* FRC 41: Vaccine Potential and Protein Expression Using Multiple Escherichia coli Strains.

Microbial Cell Factories

Putative Virulence Factors of Corynebacterium pseudotuberculosis FRC 41: Vaccine Potential and Protein Expression Using Multiple Escherichia coli Strains.

--Manuscript Draft--

Manuscript Number:	MICF-D-15-00117
Full Title:	Putative Virulence Factors of Corynebacterium pseudotuberculosis FRC 41: Vaccine Potential and Protein Expression Using Multiple Escherichia coli Strains.
Article Type:	Research
Section/Category:	Recombinant protein production and quality
Funding Information:	Fundação de Amparo à Pesquisa do Estado de Minas Gerais — Prof. Vasco A de C Azevedo CNPq — Prof. Vasco A de C Azevedo Dr. Ricardo B Mariutti Prof. Raghuvir K Arni CAPES — M.SC Karina T. O Santana-Jorge M.Sc Natayme R Tartaglia M.Sc Edgar L Aguiar M.Sc Renata F.S Souza Dr. Túlio M Santos FAPESB — Prof. Roberto Meyer
Abstract:	Abstract Background Corynebacterium pseudotuberculosis, a facultative intracellular bacterial pathogen, is the etiological agent of caseous lymphadenitis (CLA), a contagious disease that primarily infects sheep and goats and is responsible for significant economic loss. The disease is characterized mainly by bacteria-induced caseous necrosis in lymphatic glands. New vaccines are needed for reliable control and management of CLA. Thus, the putative virulence factors SpaC, SodC, NanH, and PknG from C. pseudotuberculosis FRC 41 may represent new target proteins for vaccine development and pathogenicity studies. Results Proteins of microorganisms from the CMNR group (Corynebacterium, Mycobacterium, Nocardia and Rhodococcus) are highly similar to target proteins. However, target proteins presented low similarity with mammalian proteins indicating low possibility of immunological cross-reactions. Ppredicted B and T-cell epitope densities were significant. Target proteins were expressed in E. coli. However, the expression levels of each protein in five different E. coli BL21(DE3) derived strains varied widely, from no apparent expression in some strains to high amounts in others. Conclusions In silico analyses show that the putative virulence factors SpaC, SodC, NanH, and PknG present good potential as targets for vaccine development. The four proteins were successfully over-expressed in E. coli however, fluctuations in protein expression levels observed in different E. coli BL21(DE3) derived strains indicate that the expression experiments performed using multiple host strains, i. e., the expression of a protein in several different strains, under identical experimental conditions of growth medium, temperature, IPTG concentration and expression time-course, can be a quick screening strategy to circumvent problems commonly observed in over-expression experiments usually performed with only one strain. The strategy can be especially useful for expression of large sets of proteins.
Corresponding Author:	Vasco A de C Azevedo, Ph.D Universidade Federal de Minas Gerais Belo Horizonte, Minas Gerais BRAZIL
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Universidade Federal de Minas Gerais

Corresponding Author's Secondary Institution:	
First Author:	Vasco A de C Azevedo, Ph.D
First Author Secondary Information:	
Order of Authors:	Vasco A de C Azevedo, Ph.D
	Karina T. O Santana-Jorge, M.SC
	Natayme R Tartaglia, M.Sc
	Edgar L Aguiar, M.Sc
	Renata F.S Souza, M.Sc
	Ricardo B Mariutti, Ph.D
	Raghuvir K Arni, Ph.D
	Roberto Meyer, Ph.D
	Túlio M Santos, Ph.D
Order of Authors Secondary Information:	

Putative Virulence Factors of *Corynebacterium pseudotuberculosis* FRC 41: Vaccine Potential and Protein Expression Using Multiple *Escherichia coli* Strains.

Karina T. O. Santana-Jorge[1], Natayme R. Tartaglia[1], Edgar L. Aguiar[1], Renata F. S. Souza[1], Ricardo B. Mariutti[2], Raghuvir K. Arni[2], Roberto J.M. Nascimento[3], Túlio M. Santos[1,4], Vasco A. C. Azevedo[1§]

[1]Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

[2]Multiuser Center for Biomolecular Innovation, Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista "Júlio de Mesquita Filho", São José do Rio Preto, SP, Brazil

[3]Departamento de Bio-Interação, Instituto de Ciências da Saúde, Universidade Federal da Bahia, Salvador, BA, Brazil

[4]Uniclon Biotecnologia, Belo Horizonte, MG, Brazil

[§]Corresponding author

Vasco Azevedo, Departamento de Biologia Geral, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais. Avenida Antonio Carlos, 6627, Pampulha, Belo Horizonte, Brazil, 31270-901, Tel: 55 31 3409-2610; E-mail: vasco@icb.ufmg.br

Email addresses:

KTOSJ: karinasantana.bqi@gmail.com

NRT: nrtbiomed@gmail.com

ELA: edgarlaguiar@gmail.com

RFSS: refasi@hotmail.com

RBM: ricardomariutti@yahoo.com.br

RKA: arni@sjrp.unesp.br

RJMN: rmeyer@ufba.br

TMS: uniclon@uniclon.com

VACA: vasco@icb.ufmg.br

Abstract

Background

Corynebacterium pseudotuberculosis, a facultative intracellular bacterial pathogen, is the etiological agent of caseous lymphadenitis (CLA), a contagious disease that primarily infects sheep and goats and is responsible for significant economic loss. The disease is characterized mainly by bacteria-induced caseous necrosis in lymphatic glands. New vaccines are needed for reliable control and management of CLA. Thus, the putative virulence factors SpaC, SodC, NanH, and PknG from *C. pseudotuberculosis* FRC 41 may represent new target proteins for vaccine development and pathogenicity studies.

Results

Proteins of microorganisms from the CMNR group (*Corynebacterium*, *Mycobacterium*, *Nocardia* and *Rhodococcus*) are highly similar to target proteins. However, target proteins presented low similarity with mammalian proteins indicating low possibility of immunological cross-reactions. Ppredicted B and T-cell epitope densities were significant. Target proteins were expressed in *E. coli*. However, the expression levels of each protein in five different *E. coli* BL21(DE3) derived strains varied widely, from no apparent expression in some strains to high amounts in others.

Conclusions

In silico analyses show that the putative virulence factors SpaC, SodC, NanH, and PknG present good potential as targets for vaccine development. The four proteins were successfully over-expressed in *E. coli* however, fluctuations in protein expression levels observed in different *E. coli* BL21(DE3) derived strains indicate that the expression experiments performed using multiple host strains, i. e., the expression of a protein in several different strains, under identical experimental conditions of

growth medium, temperature, IPTG concentration and expression time-course, can be a quick screening strategy to circumvent problems commonly observed in over-expression experiments usually performed with only one strain. The strategy can be especially useful for expression of large sets of proteins.

Keywords

Corynebacterium pseudotuberculosis, virulence factor, vaccine potential, protein expression, *E. coli* strain.

Background

Caseous lymphadenitis (CLA) is a chronic, pyogenic, contagious disease of sheep and goat that imposes considerable economic losses for farmers in many countries [1 - 4]. The disease is caused by *Corynebacterium pseudotuberculosis* (*C. pseudotuberculosis*): a facultative intracellular, non-capsulated, non-motile, fimbriated, Gram-positive pleomorphic bacterium [1, 2, 3].

CLA is expressed in external and visceral forms, either separately or together [3 - 6]. External CLA lesions appear initially as abscesses that convert later on to pyogranulomas ranging in size from millimeters to centimeters. These external lesions are mostly located within superficial lymph nodes, but infrequently in subcutaneous tissues. Wool or hair over CLA lesions may be lost due to the weak dermonecrotic action of *C. pseudotuberculosis* exotoxins and the pressure atrophy of overlying skin by the lesions. Visceral lesions are not detectable clinically but express themselves according to their number, site and effect on the involved organ. Progressive weight

loss, respiratory disorders and chronic recurrent ruminal tympany are the most prominent signs that may accompany visceral CLA lesions.

Identification/removal of infected animals is a key factor for success of disease control measures [1, 2, 3]. Although signs of CLA are considered characteristic and highly suggestive, particularly if several animals within the same flock are affected, clinical signs are not always confirmatory. Reliability of clinical diagnosis of CLA is negatively impacted by the occurrence of subclinical cases, too small undetectable lesions, and the accidental evacuation of some superficial lesions that reappear after a period during which the animal may appear clinically normal [6]. Clinical diagnosis is only suggestive as there are many bacterial organisms besides *C. pseudotuberculosis* that are able to induce superficial abscesses in small ruminants [2, 3]. However, abscesses which are caused by other pyogenic bacteria mostly occur sporadically in contrast to CLA lesions which usually involve the flock [2].

The most powerful confirmatory diagnostic method for CLA is the isolation and identification of *C. pseudotuberculosis*, but isolation failure and inaccessibility of visceral lesions to be sampled are limitations for such diagnostic method [6]. In practical terms, identification/removal of infected animals usually requires serologic testing to detect humoral responses to PLD exotoxin, enabling the culling of seropositive animals [4].

Among several serological tests being used to detect humoral response against *C. pseudotuberculosis* (e.g., agar gel immunodiffusion test - AGID, microagglutination assay, hemolysis inhibition test, synergistic hemolysis inhibition test - SHI, indirect

haemagglutination test, Westernblotting, tube agglutination assay, and dot-blot), ELISA is considered the gold standard due to its capacity, speed, acceptable sensitivity and specificity [3, 6]. However, polymerase chain reaction may represent a promising alternative to traditional methods for diagnosis of CLA primarily due to its specificity and sensitivity in detecting as little as 1 pg of pathogen genomic material and up to 103 CFU in clinical samples [3, 6, 7].

Vaccination of healthy animals is another strategy broadly recommended for disease control. In fact, control of CLA depends on vaccination in most countries [3, 4]. Although bacterin, toxoid, combined, and live vaccines are available, the disease has persisted even after prolonged vaccination, indicating the suppressive nature of CLA vaccination [4]. Because of its zoonotic potential, *C. pseudotuberculosis* infection of animals can contaminate meat and milk, putting consumers at risk [8]. The ability of *C. pseudotuberculosis* to infect both animals and humans makes studies on prevention and diagnosis of this pathogen important. Therefore, more specific and sensitive diagnostic tests for precocious identification/removal of infected animals as well as new drugs and vaccines are still necessary for a reliable control and management of CLA [8].

The study of *C. pseudotuberculosis* virulence factors involved in CLA pathogenesis can provide new approaches for diagnostic, treatment, and disease control. The complete genome sequence of *C. pseudotuberculois* FRC41 allowed the identification of *spaC* (a pili tip protein), *sodC* (a copper, zinc-dependent superoxide dismutase), and *nanH* (a neuraminidase H) as genes putatively involved in pathogen virulence [9].

As part of important cell signaling mechanisms, eukaryotic-like serine/threonine protein kinases encountered in bacteria are a class of molecules that also deserves attention since they are part of complex signaling pathways and play a diversity of physiological roles in developmental processes, secondary metabolism, cell division, cell wall synthesis, essential processes, central metabolism, and virulence [10, 11]. *M. tuberculosis* genome encodes 11 eukaryotic-like serine/threonine protein kinases (PknA to PknL, except for PknC). PknG is well conserved in *Mycobacterium* and *Corynebacterium*. It promotes survival within macrophages in *Mycobacterium* and is linked to cellular glutamate/glutamine metabolism [12 - 15]. However, the function of PknG in *C. pseudotuberculosis* still needs to be investigated.

In the present work, an evaluation of PknG, SpaC, SodC, and NanH as potential vaccine target was performed using bioinformatic tools. The heterologous expression of these *C. pseudotuberculosis* putative virulence factors in *E. coli* using a host strain changing strategy is also described. The production of these proteins in large amounts represents an important step for future pathogenicity and vaccine development studies.

Results and Discussion

C. pseudotuberculosis pathogenicity involves the potent exotoxin phospholipase D (PLD), a permeability factor that promotes the hydrolysis of sphingomyelin ester bonds in mammalian cell membranes causing dermal necrosis, possibly contributing to the spread of the bacteria from the initial site of infection to secondary sites within the host, as well as toxic cell wall lipids which are macrophage necrotizing [1, 4, 8].

After *C*. *pseudotuberculosis* FRC41 genome sequencing, SpaC, SodC, NanH proteins have been proposed as novel virulence factors of the pathogen [9]. SpaC is an adhesive pili tip protein [16]. The pilus structure can probably make the initial contact with host cell receptors to enable additional ligand-receptor interactions and to facilitate the efficient delivery of virulence factors and intracellular invasion.

The copper,zinc-dependent superoxide dismutase SodC is probably involved in detoxification reactions against peroxynitrite and other reactive nitrogen and oxygen intermediates produced by macrophages as part of their antimicrobial response [9, 17] SOD converts superoxide anions into molecular oxygen and H_2O_2, the latter being broken in turn to H_2O by the enzymatic activity of catalase. The protective activity of Cu,Zn-SODs has been associated with virulence in many bacteria, such as *Mycobacterium tuberculosis*, *Neisseria meningitidis* and *Hemophylus ducreyi* [9, 17].

NanH, by its turn, is an extracellular neuraminidase [9]. Neuraminidases, or sialidases, belong to a class of glycosyl hydrolases that catalyze the removal of terminal sialic acid residues from a variety of glycoconjugates and can contribute to the recognition of sialic acids exposed on host cell surfaces [18, 19]. A homologous counterpart of NanH was recently characterized in *C*. *diphtheriae* KCTC3075 and shown to be a protein containing neuraminidase and trans-sialidase activities [19].

Protein kinaseG (PknG) gained particular interest because it affects the intracellular traffic of *M*. *tuberculosis* in macrophages. Most microbes and nonpathogenic mycobacteria quickly find themselves in lysosomes, where they are killed. By contrast, *M*. *tuberculosis* stays within phagosomes; the bacterium releases PknG to

block phagosome-lysosome fusion. Bacteria lacking *pknG* gene are rapidly transferred to lysosomes and eliminated [12, 13]. Other studies of PknG in *M*. *tuberculosis* [14] and the related actinomycete *Corynebacterium glutamicum* [15] have implicated this kinase in the regulation of glutamine metabolism and in the regulation of 2-oxoglutarate dehydrogenase, an enzyme of the tricarboxylic acid cycle.

Therefore, *C*. *pseudotuberculosis* SpaC, SodC, NanH and PknG proteins, may be good candidate antigens for vaccine development once they may play important role in pathogen survival against host defenses and pathogenicity.

Membrane and secreted proteins are considered potential vaccine targets once they are at the host-pathogen interface. These proteins may interact more directly with host molecules for cell adhesion, invasion, multiplication, immune response evasion, damage generation to the host, and survive to host cell defenses. PSORTb program [20] was used to predict the subcellular localization of *C*. *pseudotuberculosis* putative virulence proteins. PknG was predict as cytoplasmic protein, SpaC in pathogen cell wall while SodC and NanH as extracellular proteins (**Tab. 2**). PknG of *M*. *tuberculosi*s is predominantly found soluble in cytoplasm [11]. The presence of signal peptide was predicted by SignalP 4.1 program [21] only in SodC and NanH sequences (data not shown) as already observed in their homologs [9]. SpaC by its turn has a LPXTG motif for cell wall-anchorage as observed in its *C*. *diphtheriae* homolog [22].

The conservation degree among the putative virulence factors and proteins of the CMNR microorganism group (*Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*) was evaluated by BLAST searches in UniprotKB database [23] and

CLUSTALW2 [24] alignments. This kind of analysis is important for the development of new drugs and vaccines once they can be used not only for *C*. *pseudotuberculosis* FRC41 but also for other pathogen strains and pathogens of other genera. The results show (**Tabs. 3, 4,** and **5**) that *C*. *pseudotuberculosis* FRC41 SpaC, SodC, NanH, and PknG sequences share high identity degree with sequences from other *C*. *pseudotuberculosis* strains (1002, C231, and 258). Well conserved homologs of the target proteins are also found in some microorganisms of the CMNR group. High conservation regions identified by CLUSTALW2 alignments between target proteins and homologs are ideal for vaccine preparations (data not shown).

The conservation degree among *C*. *pseudotuberculosis* FRC41 putative virulence factors and mammalian (*Ovis*, *Bos*, and *Equs* genera, *Mus musculus*, and *Homo sapiens*) proteins was also evaluated by BLAST searches. The analysis was important to reveal the conservation degree among pathogen proteins and host proteins and so the possibility of undesirable immunological cross-reactions. The results (**Tabs. 3, 4,** and **5**) show that *C*. *pseudotuberculosis* FRC41 SpaC, SodC, NanH, and PknG sequences share 30% identity in average with mammalian sequences. CLUSTALW2 alignments show that some of the high conservation regions are in functional domains like the kinase domain in PknG (data not shown). Thus, the lowest conservation regions revealed by CLUSTALW2 alignments between target proteins and mammalian homologs are ideal targets for vaccine development (data not shown).

The identification of epitopes that invoke effective antibody responses from B-cells and strong and long lasting memory from T-cells is one of the key steps in designing effective vaccines against pathogens [25, 26]. Bioinformatic analyses were performed

to predict T and B-cell epitopes in SpaC, SodC, NanH, and PknG proteins. The analyses using MED [27] and NetMHCII [28] programs for T-cell epitope prediction show PknG as the protein with the highest MHCI and MHCII epitope density while the others show average densities. Bcepred program [29] for B-cell epitope prediction shows a homogenous epitope distribution in SpaC, SodC, NanH, and PknG sequences and a small number of epitopes in SodC (**Tab. 6**).

Large amounts of SpaC, SodC, NanH, and PknG proteins are necessary for future studies on their role in *C. pseudotuberculosis* patogenicity and to develop new vaccines. *Escherichia coli* remains as one of the most attractive hosts among many systems available for heterologous protein production [30, 31]. Its well-characterized genetics, fast growth at a high density in an inexpensive medium, and the availability of a large number of cloning vectors and mutant host strains have enable *E. coli* a first choice host for rapid, high yield, and economical production of recombinant proteins. However, in spite of the extensive knowledge on the genetics and molecular biology of *E. coli*, not every gene can be expressed efficiently and high-level production of functional eukaryotic proteins in *E. coli* may not be a routine matter, and sometimes it is quite challenging. Many factors contribute to this challenge, including the unique and subtle structural features of the gene sequence to be expressed, the stability and translational efficiency of mRNA, the difficulty of protein folding, degradation of protein by host cell proteases, and codon usage toxicity of the protein to the host [32].

Methods to optimize heterologous protein overproduction in *E. coli* have been developed to significantly enhance the yield of foreign proteins [30 – 34, 36]. Five strategies can be used to increase the expression and solubility of over-expressed

protein: (1) changing the vector, (2) changing the host, (3) changing the culture parameters of the recombinant host strain, (4) co-expression of other genes, and (5) changing the gene sequences, which may help increase expression and the proper folding of desired protein [30].

Nowadays so many biotech companies provide different types of genetically altered *E. coli* strain for suitable expression of foreign genes [30, 37]. This makes the strategy of changing the host strain a very quick and easy starting point towards a successful production of heterologous proteins in *E. coli* since it requires only well established cell transformation and screening protocols.

In this way, *pknG*, *spacC*, *sodC*, and *nanH* codon-optimized ORFs cloned into the same expression vector system were individually transformed into five different BL21(DE3) derived *E. coli* strains (**Tab. 1**). The SDS-PAGE analyses show that the expression levels of each recombinant protein varied widely from strain to strain, from no apparent expression to large amounts under the same experimental conditions of growth medium, temperature, IPTG concentration, and expression time-course (**Fig. 1**). The best expression performances were obtained in OverExpress™ C41(DE3) for SpaC and NanH, OverExpress™ C43(DE3) pLysS for SodC, and BL21 Star™ (DE3) for PknG.

BL21 strain is the preferred choice for expression because of the absence of two main proteases [30]. *E. coli* strain BL21(DE3) has *Lon* and *OmpT* protease deleted in its genome. *Lon* and *OmpT* protease deficient strain of *E. coli* are unable to degrade foreign protein. Leaky expression of the desired gene in BL21(DE3) cells is possible.

To minimise leaky expression of toxic genes, the BL21 host strain has been improved. The improved strains are BL21(DE3)pLys S and BL21(DE3)pLys E [30, 37]. Both strains have lysozyme coding plasmid. Lysozyme is an inhibitor of T7 polymerases which inhibits residual T7 polymerase and thus prevents leaky expression. Leaky expression of a toxic gene is detrimental to the host cell—the host cell will not survive or it will change its mechanism so that even in presence of inducer it will not allow production of toxic protein.

Many times expression does not occur in *E. coli* due to differences in the codons present in a gene of interest and preferential codon usage by the host *E. coli* strain [30, 36]. To solve this codon bias in host cell machinery, a few companies provide modified *E. coli* hosts that have extra tRNA coding genes (AUA, AGG, AGA, CUA, CCC and GGA) compensating for the scarcity of rare tRNA. BL21(DE3) and derived strains were designed to enhance the expression of heterologous proteins by containing codons rarely used in *E. coli*.

The enumerated features of BL21(DE3) and its derived strains used in this study as well as the codon usage optimization of *pknG*, *spacC*, *sodC*, and *nanH* ORFs did not prevent a wide fluctuation of their expression levels in each strain as observed (**Fig. 1**). Consequently, the expression of a target ORF in one strain but not in another may be credited to physiological peculiarities of each *E. coli* strain. Silent mutations that accumulate over generations of growth can be selected by the advantage they confer to *E. coli* in the presence of a detrimental recombinant gene [36].

Conclusions

All together, the bioinformatic analyses show that SpaC, SodC, NanH, and PknG present good potential as targets for vaccine development. The four putative virulence factors of *C. pseudotuberculosis* FRC 41 were successfully expressed in *E. coli*. However, fluctuations in the protein expression levels observed in each *E. coli* BL21(DE3) derived strains indicate that the expression in multiple host strains, i. e., the expression of a protein in several different strains, can be a quick and easy screening strategy to circumvent problems frequently observed in expression experiments that often use only one strain. Performing protein expression experiments using multiple *E. coli* host strains enhance the chances of having a successful expression by quickly finding an *E. coli* strain that does not show problems when dealing with synthesis, folding, and toxicity of a particular exogenous protein that is being over-expressed. This expression strategy can be particularly useful for, e. g., structural and protein-based pharmaceutical studies which collectively have to target and purify tens of thousands of different proteins.

Methods

Bioinformatic analyses

PSORTb program [20] was used to predict the subcellular localization of *pknG* [GeneID# 9449800], *spacC* [GeneID# 9449841], *sodC* [GeneID# 9448545], and *nanH* [GeneID# 9448316] open reading frames (ORFs). SignalP 4.1 program [21] was used to predict the presence of signal peptides. BLAST searches in UniProtKB database [23] were performed to identify homologs of the target proteins (PknG, SpaC, SodC, and NanH) in the CMNR (*Corynebacterium*, *Mycobacterium*, *Nocardia* and *Rhodococcus*) group of microorganisms. BLAST searches in UniprotKB

Eucariota database were performed in order to identify homologs of the target proteins in mammalian species of the genera *Ovis*, *Bos*, and *Equs* as well as in *Mus musculus* and *Homo sapiens*. ClustalW2 [24] alignments were performed among target proteins and their homologs identified by BLAST searches. MED [27], NetMHCII [28], and Bcepred [29] programs were used to predict MHC class I, MHC class II, and B-cell epitope densities, respectively, in the target proteins.

Cloning Procedures

Miniprep plasmid purifications, agarose gel electrophoresis, and *Escherichia coli* media were as described [35]. Predicted signal peptide signatures were removed from the ORF sequences before cloning procedures. ORF codons of target proteins were replaced by *E. coli* preferential codons. Afterwards, the ORF sequences were synthesized and individually cloned into pD444-NH expression vector by DNA2.0 (Menlo Park, CA). Each ORF-containing plasmid (pD444-NH;*pknG*, pD444-NH;*spacC*, pD444-NH;*sodC*, and pD444-NH;*nanH*) was transformed into five different BL21(DE3) *E. coli* strains (**Tab. 1**) according to the OverExpress™ Electrocompetent Cells kit (Lucigen, Middleton) instructions.

Protein Expression in *E. coli*

Protein expression protocol was according to OverExpress™ Electrocompetent Cells kit (Lucigen, Middleton) instructions. Transformed cell cultures at OD 0.5-0.7 were induced with 1 mM IPTG for 5 hours at 37ºC. SDS-PAGE of non-induced and induced cell culture samples and coomassie blue staining were as described [35].

List of abbreviations used

CLA, caseous lymphadenitis; CMNR microorganism group, *Corynebacterium, Mycobacterium, Nocardia, Rhodococcus*; PLD, phospholipase D; ELISA, enzyme-linked immunosorbent assay; CFU, colony-forming unit; SOD, superoxide dismutase; ORF, open reading frame; SDS-PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

KTOSJ, NRT, RFSS and RBM carried out the experiments. KTOSJ and ELA performed the bionformatic analyzes. KTOSJ and TMS drafted the manuscript. VACA, TMS, RJMN and RKA participated in the design and coordination of the study. All authors have read and approved the manuscript.

Acknowledgements

This study was supported by CAPES, CNPq, FAPEMIG and FAPESB.

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**(2):201-218
2. Baird GJ, Fontaine MC: ***Corynebacterium pseudotuberculosis* and its role in ovine caseous lymphadenitis**. *J Comp Pathol* 2007, **137**(4):179-210

3. Guimarães AS, Carmo FB, Pauletti RB, Seyffert N, Ribeiro D, Lage AP, ... Gouveia AMG: **Caseous lymphadenitis: epidemiology, diagnosis, and control**. *IIOAB J* 2011, **2**(2)
4. Windsor, PA: **Control of caseous lymphadenitis**. *Vet Clin N Am - Food Animal Practice* 2011, **27**(1):193-202
5. Fontaine MC, Baird GJ: **Caseous lymphadenitis**. *Small Ruminant Res* 2008, **76**(1):42-48
6. Oreiby AF: **Diagnosis of caseous lymphadenitis in sheep and goat**. *Small Ruminant Res* 2015, **123**(1):160-166
7. Pacheco LG, Pena RR, Castro TL, Dorella FA, Bahia RC, Carminati R, 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**(4):480-486
8. Bastos BL, Dias Portela RW, Dorella FA, Ribeiro D, Seyffert N, et al (2012) ***Corynebacterium pseudotuberculosis*: Immunological Responses in Animal Models and Zoonotic Potential**. *J Clin Cell Immunol* S4:005.
9. Trost E, Ott L, Schneider J, Schröder J, Jaenicke S, Goesmann A, 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**(1):728
10. Pereira SF, Goss L, & Dworkin J: **Eukaryote-like serine/threonine kinases and phosphatases in bacteria**. *Microbiol Mol Biol R* 2011, **75**(1):192-212

11. Forrellad MA, Klepp LI, Gioffré A, Sabio y Garcia J, Morbidoni HR, Santangelo MDLP, Bigi F: **Virulence factors of the *Mycobacterium tuberculosis* complex**. *Virulence* 2013, **4**(1):3-66

12. Walburger A, Koul A, Ferrari G, Nguyen L, Prescianotto-Baschong C, Huygen K, Pieters J: **Protein kinase G from pathogenic mycobacteria promotes survival within macrophages**. *Science* 2004, **304**(5678):1800-1804

13. Warner DF, Mizrahi V: **The survival kit of *Mycobacterium tuberculosis***. *Nat Med* 2007, **13**(3):282-284

14. Cowley S, Ko M, Pick N, Chow R, Downing KJ, Gordhan BG, Av‐Gay Y: **The *Mycobacterium tuberculosis* protein serine/threonine kinase PknG is linked to cellular glutamate/glutamine levels and is important for growth in vivo**. *Mol Microbiol* 2004, **52**(6):1691-1702

15. Niebisch A, Kabus A, Schultz C, Weil B, Bott M: **Corynebacterial protein kinase G controls 2-oxoglutarate dehydrogenase activity via the phosphorylation status of the OdhI protein**. *J Biol Chem* 2006, **281**(18):12300-12307

16. Rogers EA, Das A, Ton-That H: **Adhesion by pathogenic corynebacteria**. *Adv Exp Med Biol* 2011, **715**:91-103

17. Gengenbacher M, Kaufmann, SH: ***Mycobacterium tuberculosis*: success through dormancy**. FEMS microbiology reviews 2012, **36**(3):514-532

18. Vimr ER, Kalivoda KA, Deszo EL, Steenbergen SM: **Diversity of microbial sialic acid metabolism**. *Microbiol Mol Biol R* 2004, **68**(1):132-153

19. Kim S, Oh DB, Kwon O, Kang, HA: **Identification and functional characterization of the NanH extracellular sialidase from *Corynebacterium diphtheriae***. *J Biochem* 2010, **147**(4):523-533

20. Nancy YY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Brinkman FS: **PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes.** *Bioinformatics* 2010, **26**(13):1608-1615

21. Petersen TN, Brunak S, von Heijne G, Nielsen H: **SignalP 4.0: discriminating signal peptides from transmembrane regions**. *Nat Methods* 2011, **8**(10):785-786

22. Chang C, Mandlik A, Das A, Ton‐That H: **Cell surface display of minor pilin adhesins in the form of a simple heterodimeric assembly in *Corynebacterium diphtheriae***. *Mol Microbiol* 2011, **79**(5):1236-1247

23. UniProt Consortium: **UniProt: a hub for protein information**. *Nucleic Acids Res* 2014, gku989.

24. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Higgins DG: **Clustal W and Clustal X version 2.0**. *Bioinformatics* 2007, **23**(21):2947-2948

25. Yasser EM, Honavar V: **Recent advances in B-cell epitope prediction methods**. *Immunome Res* 2010, **6**(Suppl 2), S2.

26. Lundegaard C, Hoof I, Lund O, Nielsen M: **State of the art and challenges in sequence based T-cell epitope prediction**. *Immunome Res* 2010, **6**(Suppl 2), S3.

27. Santos AR, Pereira VB, Barbosa E, Baumbach J, Pauling J, Röttger R, Azevedo, V: **Mature Epitope Density-A strategy for target selection based**

on immunoinformatics and exported prokaryotic proteins. *BMC Genomics* 2013, **14**(Suppl 6), S4.

28. Nielsen M, Lundegaard C, Lund O: **Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method.** *BMC Bioinformatics* 2007, **8**(1):238

29. Saha S, Raghava GPS: **BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties**. *Lec Notes Comput Sc* 2004, **3239**:197-204

30. Gopal GJ, Kumar A: **Strategies for the production of recombinant protein in *Escherichia coli*.** *Protein J* 2013, **32**(6):419-425

31. Sugiki T, Fujiwara T, Kojima C: **Latest approaches for efficient protein production in drug discovery.** *Expert Opin Drug Dis* 2014, **9**(10):1189-1204

32. Sivashanmugam A, Murray V, Cui C, Zhang Y, Wang J, Li Q: **Practical protocols for production of very high yields of recombinant proteins using *Escherichia coli***. *Protein Sci* 2009, **18**(5):936-948

33. Sahdev S, Khattar SK, Saini KS: **Production of active eukaryotic proteins through bacterial expression systems: a review of the existing biotechnology strategies**. *Mol Cell Biochem* 2008, **307**(1-2):249-264

34. Gräslund S, Nordlund P, Weigelt J, Bray J, Gileadi O, Knapp S, Zhang F: **Protein production and purification**. *Nat Methods* 2008, **5**(2):135-146

35. . Sambrook J, Russell DW: Molecular Cloning: A Laboratory Manual, 3rd ed. New York, Cold Spring Harbor Laboratory Press 2001

36. Saida F, Uzan M, Odaert B, Bontems F. **Expression of highly toxic genes in *E. coli*: special strategies and genetic tools**. *Curr Protein Pept Sc* 2006, **7**(1):47-56.

37. Miroux B, Walker JE. **Over-production of proteins in *Escherichia coli*: mutant hosts that allow synthesis of some membrane proteins and globular proteins at high levels.** *J Mol Biol* 1996, **260**(3):289-298.

Figure legend

Figure 1. Heterologous expression of the *C. pseudotuberculosis* FRC41 putative virulence factors in different *E. coli* strains.

Coomassie blue-stained SDS-PAGE gel analyses of the protein expression experiments. **PE1**, rPknG expression (83 kDa, 10% gel); **PE2**, rSpaC expression (86 kDa, 10% gel); **PE3**, rSodC expression (18 kDa, 15% gel); **PE4**, rNanH expression (71.5 kDa, 10% gel). **A1**, **B1**, **C1**, pre-stained protein ladder; **A2**, **A3**, *E. coli* strain C41 (DE3); **A4**, **A5**, *E. coli* strain C41 (DE3) pLysS; **B2**, **B3**, *E. coli* strain C43 (DE3); **B4**, **B5** *E. coli* strain C43 (DE3) pLysS; **C2**, **C3**, *E. coli* strain BL21 Star (DE3); **NI**, non-induced time 0 and **I**, induced with 1 mM IPTG for 5 hours at 37ºC. Arrows indicate the recombinant protein position in the gels.

Tables

Table 1 - *E. coli* strains used for cloning and expression of *C. pseudotuberculosis* FRC41 putative virulence factors.

Strain*	Genotype
OverExpress™ C41(DE3)	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3)
OverExpress™ C41 (DE3) pLysS	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3) pLysS (CmR)
OverExpress™ C43(DE3)	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3)
OverExpress™ C43(DE3) pLysS	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3) pLysS (CmR)
BL21 Star™ (DE3)	F- *ompT hsdSB* (*rB-mB-*) *gal dcm rne131* (DE3)

*All from Lucigen, Middleton.

Table 2 - Subcellular localization of *C. pseudotuberculosis* FRC41 putative virulence factors.

Subcellular Localization Predictions by PSORTb[1]					
Protein	Score				Final Prediction
	Cytoplasmic	Cytoplasmic membrane	Cell Wall	Extracellular	
PknG	9.89	0.09	0.01	0.02	Cytoplasmic
SpaC	0.01	0.01	9.97	0.01	Cell Wall
SodC	0.24	0.05	0.80	8.91	Extracellular
NanH	0.00	0.14	0.16	9.70	Extracellular

[1]PSORTb [20].

Table 3 - CMNR microorganism and mammalian proteins identified by BLAST as homologs of *C. pseudotuberculosis* FRC41 putative virulence factor PknG.

Homolog proteins in CMNR[1] microorganisms				Homolog proteins in Mammalians[2]			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	PknG	100	0.0	*Ovis aries*	Uncharacterized Protein PRKX[4]	30	$4.0x10^{-12}$
C. pseudotuberculosis 258	PknG	99	0.0	*Ovis aries*	Uncharacterized Protein CAMK1[5]	29	$4.0x10^{-12}$
C. pseudotuberculosis C231	PknG	100	0.0	*Bos Taurus*	Uncharacterized Protein PRKX[4]	30	$3.0x10^{-12}$
C. diphteriae HC02	PknG	74	0.0	*Bos Taurus*	CAMK1[5]	29	$1.0x10^{-12}$
C. glutamicum ATCC 14067	S/T PK[3]	61	0.0	*Equs caballus*	Uncharacterized Protein CAMK1[5]	31	$9.0x10^{-13}$
M. tuberculosis SUMu007	PknG	46	$1.0x10^{-169}$	*Mus musculus*	CAMKI[5]	29	$3x10^{-13}$
N. farcinica IFM10152	Putative S/T PK[3]	41	$4.0x10^{-173}$	*Mus musculus*	Smok2b[6]	31	$1x10^{-13}$
R. pyridinivorans SB3094	S/T PK[3]	46	$1.0x10^{-179}$	*Homo sapiens*	CAMK1[5]	29	$2x10^{-12}$

BLAST searches were performed against UniprotKB database [23] to find CMNR microorganism homologs and against UniprotKB Eucariota to find mammalian homologs. [1]*Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*; [2]*Ovis*, *Bos*, and *Equs* genera, *Mus musculus*, and *Homo sapiens*; [3]S/T PK, serine/threonine protein kinase; [4]PRKX, protein kinase X-linked; [5]CAMK1, Calcium/calmodulin-dependent protein kinase 1; [6]Smok2b, sperm motility kinase 2B.

Table 4 - CMNR microorganism and mammalian proteins identified by BLAST as homologs of *C. pseudotuberculosis* FRC41 putative virulence factors SpaC and NanH.

SpaC homolog proteins in CMNR[1] microorganisms				SpaC homolog proteins in mammalians[2]			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	Uncharacterized protein	99	0.0				
C. pseudotuberculosis C231	Uncharacterized protein	100	0.0		No match		
C. diphteriae HC02	Putative fimbrial subunit	27	$5x10^{-15}$				
NanH homolog proteins in CMNR[1] microorganisms				**NanH homolog proteins in mammalians[2]**			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	NanH	100	0.0	*Bos taurus*	Plectin	30%	$1x10^{-8}$
C. pseudotuberculosis C231	NanH	100	0.0	*Mus musculus*	Plectin	30%	$2x10^{-8}$
C. diphteriae HC02	NanH	50	0.0	*Homo sapiens*	Plectin	30%	$2x10^{-9}$
C. glutamicum ATCC 14067	Uncharacterized protein	36	$1.0x10^{-56}$	*Homo sapiens*	NEU[3]	27%	$2x10^{-7}$

BLAST searches were performed against UniprotKB database [23] to find CMNR microorganism homologs and against UniprotKB Eucariota to find mammalian homologs. [1]*Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*; [2]*Ovis*, *Bos*, and *Equs* genera, *Mus musculus*, and *Homo sapiens*; [3]Sialidase 1 (lysosomal sialidase).

Table 5 - CMNR microorganism and mammalian proteins identified by BLAST as homologs of *C. pseudotuberculosis* FRC41 putative virulence factor SodC.

Homolog proteins in CMNR[1] microorganisms				Homolog proteins in mammalians[2]			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	SodC	100	$3x10^{-122}$	*Ovis aries*	Sod	36%	$3x10^{-15}$
C. pseudotuberculosis 258	SodC	100	$3x\ 10^{-122}$	*Bos taurus*	Sod1	34%	$2x10^{-15}$
C. pseudotuberculosis C231	SodC	100	$3x10^{-122}$	*Equs caballus*	Sod1	32%	$8x10^{-15}$
C. diphteriae HC02	SodC	68	$2x10^{-75}$	*Mus musculus*	Sod1	34%	$2x10^{-16}$
M. tuberculosis SUMu007	SodC	43	$9x10^{-28}$	*Homo sapiens*	Sod1	34%	$3x10^{-16}$
N. farcinica IFM 10152	SodC	44	$4x10^{-35}$				
Rhodococcus sp. RHA1	SodC	49	$2x10^{-42}$				

BLAST searches were performed against UniprotKB database [23] to find CMNR microorganism homologs and against UniprotKB Eucariota to find mammalian homologs. [1]*Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*; [2]*Ovis*, *Bos*, and *Equs* genera, *Mus musculus*, and *Homo sapiens*.

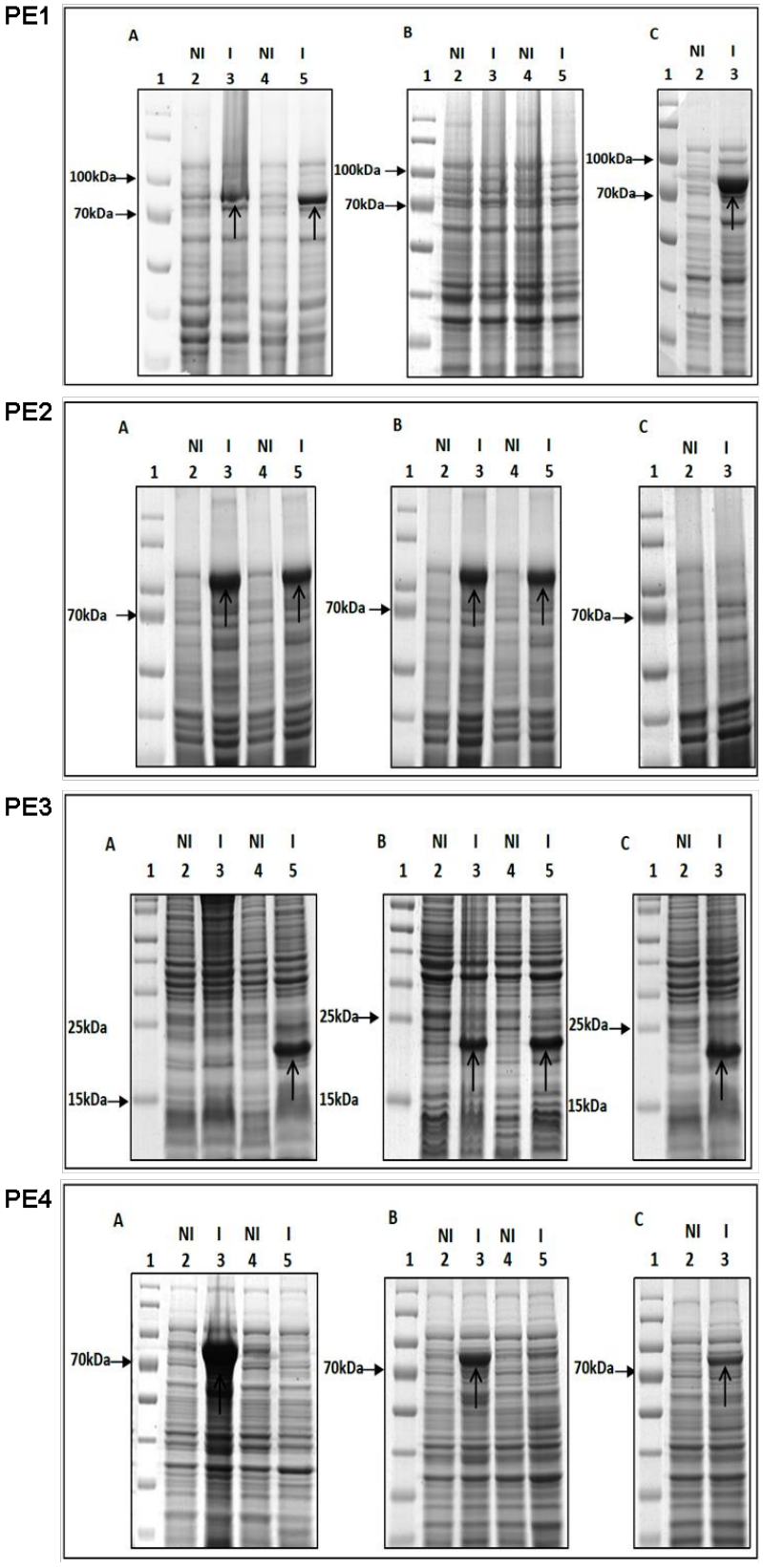
Table 6 - B cell and T cell epitope density in *C. pseudotuberculosis* FRC41 putative virulence factors.

MHC Class I Epitope Density Prediction by MED[1]	
Protein	**Score**
PknG	12,67
SpaC	7,14
SodC	6,38
NanH	7,84
MHC Class II Epitope Prediction by NetMHCII[2]	
Protein	**High Binders (%)**
PknG	5
SpaC	2
SodC	2
NanH	3
B Cell Linear Epitope Prediction by BcePred[3]	
Protein	**Number of B-cell linear epitopes**
PknG	27
SpaC	27
SodC	4
NanH	23

[1]MED, Mature Epitope Density [27]; [2]NetMHCII [28]; [3]BcePred [29].

Figure
Click here to download Figure: Figures.docx

Figure 1 - Heterologous expression of the *C. pseudotuberculosis* FRC41 putative virulence factors in different *E. coli* strains.



Tables

Table 1 - *E. coli* strains used for cloning and expression of *C. pseudotuberculosis* FRC41 putative virulence factors.

Strain*	Genotype
OverExpress™ C41(DE3)	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3)
OverExpress™ C41 (DE3) pLysS	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3) pLysS (CmR)
OverExpress™ C43(DE3)	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3)
OverExpress™ C43(DE3) pLysS	F – *ompT hsdSB* (*rB- mB-*) *gal dcm* (DE3) pLysS (CmR)
BL21 Star™ (DE3)	F- *ompT hsdSB* (*rB-mB-*) *gal dcm rne131* (DE3)

*All from Lucigen, Middleton.

Table 2 - Subcellular localization of *C. pseudotuberculosis* FRC41 putative virulence factors.

Subcellular Localization Predictions by PSORTb[1]					
Protein	Score				Final Prediction
	Cytoplasmic	Cytoplasmic membrane	Cell Wall	Extracellular	
PknG	9.89	0.09	0.01	0.02	Cytoplasmic
SpaC	0.01	0.01	9.97	0.01	Cell Wall
SodC	0.24	0.05	0.80	8.91	Extracellular
NanH	0.00	0.14	0.16	9.70	Extracellular

[1]PSORTb [20].

Table 3 - CMNR microorganism and mammalian proteins identified by BLAST as homologs of *C. pseudotuberculosis* FRC41 putative virulence factor PknG.

Homolog proteins in CMNR[1] microorganisms				Homolog proteins in Mammalians[2]			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	PknG	100	0.0	*Ovis aries*	Uncharacterized Protein PRKX[4]	30	$4.0x10^{-12}$
C. pseudotuberculosis 258	PknG	99	0.0	*Ovis aries*	Uncharacterized Protein CAMK1[5]	29	$4.0x10^{-12}$
C. pseudotuberculosis C231	PknG	100	0.0	*Bos Taurus*	Uncharacterized Protein PRKX[4]	30	$3.0x10^{-12}$
C. diphteriae HC02	PknG	74	0.0	*Bos Taurus*	CAMK1[5]	29	$1.0x10^{-12}$
C. glutamicum ATCC 14067	S/T PK[3]	61	0.0	*Equs caballus*	Uncharacterized Protein CAMK1[5]	31	$9.0x10^{-13}$
M. tuberculosis SUMu007	PknG	46	$1.0x10^{-169}$	*Mus musculus*	CAMKI[5]	29	$3x10^{-13}$
N. farcinica IFM10152	Putative S/T PK[3]	41	$4.0x10^{-173}$	*Mus musculus*	Smok2b[6]	31	$1x10^{-13}$
R. pyridinivorans SB3094	S/T PK[3]	46	$1.0x10^{-179}$	*Homo sapiens*	CAMK1[5]	29	$2x10^{-12}$

BLAST searches were performed against UniprotKB database [23] to find CMNR microorganism homologs and against UniprotKB Eucariota to find mammalian homologs. [1]*Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*; [2]*Ovis*, *Bos*, and *Equs* genera, *Mus musculus*, and *Homo sapiens*; [3]S/T PK, serine/threonine protein kinase; [4]PRKX, protein kinase X-linked; [5]CAMK1, Calcium/calmodulin-dependent protein kinase 1; [6]Smok2b, sperm motility kinase 2B.

Table 4 - CMNR microorganism and mammalian proteins identified by BLAST as homologs of *C. pseudotuberculosis* FRC41 putative virulence factors SpaC and NanH.

SpaC homolog proteins in CMNR[1] microorganisms				SpaC homolog proteins in mammalians[2]			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	Uncharacterized protein	99	0.0				
C. pseudotuberculosis C231	Uncharacterized protein	100	0.0		No match		
C. diphteriae HC02	Putative fimbrial subunit	27	$5x10^{-15}$				
NanH homolog proteins in CMNR[1] microorganisms				**NanH homolog proteins in mammalians[2]**			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	NanH	100	0.0	*Bos taurus*	Plectin	30%	$1x10^{-8}$
C. pseudotuberculosis C231	NanH	100	0.0	*Mus musculus*	Plectin	30%	$2x10^{-8}$
C. diphteriae HC02	NanH	50	0.0	*Homo sapiens*	Plectin	30%	$2x10^{-9}$
C. glutamicum ATCC 14067	Uncharacterized protein	36	$1.0x10^{-56}$	*Homo sapiens*	NEU[3]	27%	$2x10^{-7}$

BLAST searches were performed against UniprotKB database [23] to find CMNR microorganism homologs and against UniprotKB Eucariota to find mammalian homologs. [1]*Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*; [2]*Ovis*, *Bos*, and *Equs* genera, *Mus musculus*, and *Homo sapiens*; [3]Sialidase 1 (lysosomal sialidase).

Table 5 - CMNR microorganism and mammalian proteins identified by BLAST as homologs of *C. pseudotuberculosis* FRC41 putative virulence factor SodC.

Homolog proteins in CMNR[1] microorganisms				Homolog proteins in mammalians[2]			
Organism	**Protein**	**Identity (%)**	**E-value**	**Organism**	**Protein**	**Identity (%)**	**E-value**
C. pseudotuberculosis 1002	SodC	100	$3x10^{-122}$	*Ovis aries*	Sod	36%	$3x10^{-15}$
C. pseudotuberculosis 258	SodC	100	$3x\ 10^{-122}$	*Bos taurus*	Sod1	34%	$2x10^{-15}$
C. pseudotuberculosis C231	SodC	100	$3x10^{-122}$	*Equs caballus*	Sod1	32%	$8x10^{-15}$
C. diphteriae HC02	SodC	68	$2x10^{-75}$	*Mus musculus*	Sod1	34%	$2x10^{-16}$
M. tuberculosis SUMu007	SodC	43	$9x10^{-28}$	*Homo sapiens*	Sod1	34%	$3x10^{-16}$
N. farcinica IFM 10152	SodC	44	$4x10^{-35}$				
Rhodococcus sp. RHA1	SodC	49	$2x10^{-42}$				

BLAST searches were performed against UniprotKB database [23] to find CMNR microorganism homologs and against UniprotKB Eucariota to find mammalian homologs. [1]*Corynebacterium*, *Mycobacterium*, *Nocardia*, *Rhodococcus*; [2]*Ovis*, *Bos*, and *Equs* genera, *Mus musculus*, and *Homo sapiens*.

Table 6 - B cell and T cell epitope density in *C. pseudotuberculosis* FRC41 putative virulence factors.

MHC Class I Epitope Density Prediction by MED[1]	
Protein	**Score**
PknG	12,67
SpaC	7,14
SodC	6,38
NanH	7,84
MHC Class II Epitope Prediction by NetMHCII[2]	
Protein	**High Binders (%)**
PknG	5
SpaC	2
SodC	2
NanH	3
B Cell Linear Epitope Prediction by BcePred[3]	
Protein	**Number of B-cell linear epitopes**
PknG	27
SpaC	27
SodC	4
NanH	23

[1]MED, Mature Epitope Density [27]; [2]NetMHCII [28]; [3]BcePred [29].