Universidade Federal de Minas Gerais

Instituto de Ciências Biológicas

Programa de Pós Graduação em Bioinformática

"Predição Computacional de Interações de Proteína-Proteína em Proteomas Preditos de *Leishmania*"

por

Antonio Mauro Rezende

Belo Horizonte

Dezembro/2012

Agradecimentos

Primeiramente gostaria de agradecer à Deus por ter me iluminado, me inspirado, me encorajado em tantas horas nas quais os obstáculos eram tão grandes que meu único pensamento era desistir. Ele me provou que era possível e colocou no meu caminho todas as pessoas que aqui agradeço

Agradeço à minha família, em especial ao meus pais Ronaldo Rezende Silva e Ana Maria Capozoli Rezende por todo amor, carinho e confiança. Meu maior orgulho é saber que estas duas pessoas são meu pais. Exemplos essenciais na minha vida, sem eles provavelmente estaria perdido nesta minha jornada.

Agradeço à Tatiana M. Teodoro Rezende, minha esposa, que me acompanha à 8 anos. Muito obrigado por todo amor, carinho, paciência e companheirismo. Estamos juntos e cada vez mais fortes nesta caminhada. Te amo muito minha Linda!

Agradeço também à minha Tia Rozana (Rozana Rezende Silva) que sempre esteve ao meu lado me fazendo perceber que era possível alcançar meus sonhos. Muito obrigado por tudo Tia.

Agradeço aos meus irmãos Alberto Henrique Rezende (Beto) e Aline Capozoli Rezende por todo amor e carinho.

Agradeço ao meu orientador Dr. Jeronimo C. Ruiz pela amizade, orientação, confiança, paciência e pelos ensinamentos.

Agradeço à dois grandes amigos, Raul Torrieri e Patrícia C. Rui. Sem eles o meu início na Bioinformática teria sido bem mais doloroso.

Agradeço ao amigos, principalmente Armando Menezes, Edson L. Folador, Daniela M. Resende, Luciana de Oliveira, Marco Aurélio Soares, Nesley J. D. Oliveira por toda a ajuda e por todos os bons momentos vividos fora e dentro do ambiente de trabalho. Nos divertimos muito. Obrigado.

Agradeço ao Laboratório de Parasitologia Celular e Molecular e ao Centro de Pesquisas René Rachou por toda estrutura que possibilitou a realização deste projeto.

Agradeço à Universidade Federal de Minas Gerais e o Programa de Pós-Graduação em Bioinformática pelo apoio.

Agradeço à Fundação de Amparo à Pesquisa de Minas Gerais pela bolsa de doutorado concedida a mim.

Agradeço ao CNPq pelo suporte financeiro ao projeto.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Anexos

Resumo

Os parasitos Tripanosomatídeos *Leishmania braziliensis*, *Leishmania infantum* e *Leishmania major* são importantes patógenos humanos. Apesar de anos de estudo e da disponibilidade de seus genomas, nenhuma vacina eficaz foi desenvolvida até o presente momento, e os tratamentos disponíveis em geral são altamente tóxicos. Portanto, está claro que apenas estudos integrados com uma abordagem interdisciplinar terão sucesso na tentativa de buscar novos alvos para desenvolvimento de drogas e vacinas.

Uma parte essencial deste racional está relacionada ao estudo de redes de interações de proteínas as quais podem fornecer um melhor entendimento de interações proteicas complexas em sistemas biológicos.

Assim, na presente tese de doutorado modelamos redes de interações de proteínas para as três espécies de *Leishmania* citadas acima através de métodos computacionais utilizando comparação de sequência (*Interolog Mapping*), e desenvolvemos um sistema de pontuação combinado para avaliar a robustez das predições.

A avaliação de desempenho da abordagem de predição de redes foi realizada utilizando o conjunto de dados de interação de proteínas de *Escherichia coli* como padrão ouro positivo e negativo, e o valor de AUC obtido foi 0,94.

Como resultado, 39.420, 45.325 e 43.531 interações foram preditas para *L. braziliensis*, *L. Infantum* e *L. major*, respectivamente. Para cada rede predita, as 20 proteínas melhor ranqueadas pelo índice topológico MCC ("*Maximal Clique Centrality*") foram selecionadas. Além disso, informações relacionadas ao grau de conservação da sequência proteica entre os ortólogos, grau de identidade comparado às proteínas de hospedeiros potenciais, e potencial imunológico foi integrado.

Aqui vale a pena ressaltar que os algoritmos utilizados para predição de epítopos foram previamente avaliados em relação ao seus desempenhos. A avaliação ocorreu utilizando dados da base IEDB como padrão ouro. Deste modo, os programas com melhor desempenho foram então empregados.

Retomando, esta integração fornece um melhor entendimento e usabilidade das redes preditas o que pode ver valioso para seleção de novos alvos biológicos potenciais para desenvolvimento de drogas e vacinas.

Outro ponto que mereceu atenção neste estudo esta vinculado à modularidade das redes, em especial os módulos conservados, característica chave quando se está interessado em desestabilizar a rede de interação de proteína com propósitos de droga e vacina. Estas análises revelaram um padrão associado com renovação do repertório proteico.

Além disso, aproximadamente 50% das proteínas descritas como hipotéticas presentes nas redes de interação receberam algum grau de anotação funcional, o que representa uma contribuição importante uma vez que aproximadamente 60% do proteoma predito das espécies do gênero *Leishmania* não possui nenhuma predição de função.

Abstract

The Trypanosomatid parasites *Leishmania braziliensis*, *Leishmania infantum* and *Leishmania major* are important human pathogens. Despite years of study and the availability of their genomes, no effective vaccine was developed until the present moment, and the available treatments are in general highly toxic. Therefore, it is clear that only integrated studies with an interdisciplinary approach will be succeeded in trying to search new targets for drug and vaccine development.

One essential part of this rational is related to protein-protein interaction network (PPI) study which can provide a better understanding of complex protein interactions in biological systems.

Thus, in the present doctorate thesis, we modeled PPI for the three above cited species of *Leishmania* by computational methods using sequence comparison approach (*Interolog Mapping*), and developed a system of combined score to evaluate the robustness of the predictions.

The performance evaluation of the PPI prediction approach was performed using a set of protein interaction data of *Escherichia coli* as gold standard, and the value of AUC found was 0.94.

As result, 39,420, 45,325, and 43,531 interactions were predicted for *L. braziliensis*, *L. infantum* and *L. major*, respectively. For each PPI predicted, the top 20 ranked proteins in according to topological index MCC ("*Maximal Clique Centrality*") were selected. Furthermore, information related to the conservation of protein sequence among orthologs, level of identity compared to potential host proteins, and immunological potential were integrated.

Here, it is worth highlighting that the algorithms used to epitope prediction had their performances previously evaluated. This was performed utilizing data from IEDB as gold standard. Hence, the programs with the best performance were employed.

Rescuing, this integration provides a better understanding and usability of the PPIs predicted which can be valuable for selection of new biological targets for drug and vaccine development.

Other point that deserved attention in this study is linked to network modularity, focusing on conserved modules, key feature when one is interested in destabilizing the PPI for drug and vaccine purpose. These analyses revealed a pattern associated with protein turnover.

In addition, nearly 50% of the proteins describes as hypothetical present in the PPIs received some level of functional annotation, which represent an important contribution since approximately 60% of predicted proteome of species from *Leishmania* genus does not have any functional prediction.

1 – Introdução

1.1– Tripanosomatideos

A família Trypanosomatidae pertence ao reino Protista, sub-reino Protozoa, filo Sarcomastigophora, sub-filo Mastigophora, classe Zoomastigophorae e a ordem Kinetoplastida.

Esta família representa um grupo que inclui parasitas obrigatórios unicelulares, flagelados, e que alberga importantes patógenos de humanos e animais. Os principais gêneros incluem: *Leptomonas*, *Leishmania*, *Phytomonas*, *Crithidia*, *Blastocrithidia*, *Herpetomonas* e *Trypanosoma*, sendo que os gêneros *Leishmania* e *Trypanosoma* possuem importância médica. Além disso, ambos os gêneros podem ser considerados grupos monofiléticos segundo os dados da literatura (Lukeš, Jirků *et al.*, 1997; Haag, O'huigin *et al.*, 1998; Hannaert, Opperdoes *et al.*, 1998; Stevens, Noyes *et al.*, 2001; Simpson, Gill *et al.*, 2004; Simpson, Stevens *et al.*, 2006).

As espécies do gênero *Trypanosoma* causam a doença do sono e a doença de Chagas, enquanto as espécies do gênero *Leishmania* causam as leishmanioses que matam e debilitam centenas de milhares de pessoas todo ano no mundo (Simpson, Stevens *et al.*, 2006).

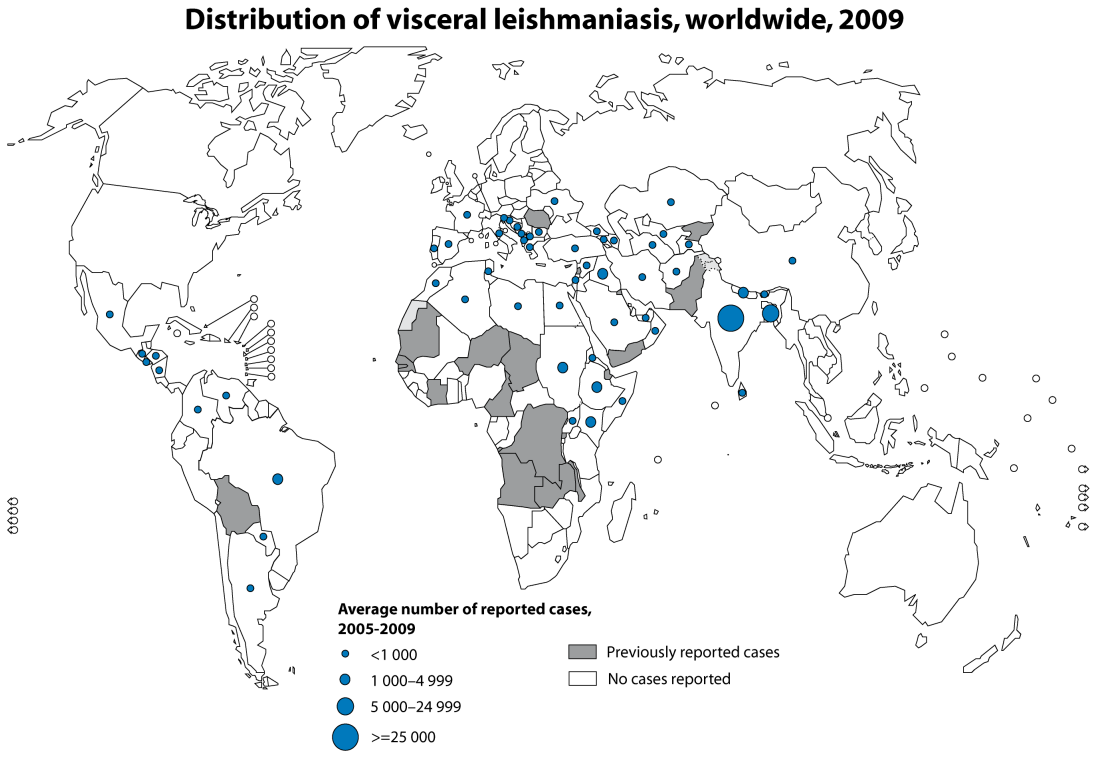
Além disso, as espécies desta família podem ser caracterizadas pela presença de um único flagelo e uma única mitocôndria contendo uma organela chamada de kinetoplasto, a qual possui o DNA mitocondrial (Simpson, Stevens *et al.*, 2006; Teixeira, De Paiva *et al.*, 2012). Os parasitos desta família, em geral, possuem ciclos de vida complexos que envolvem o homem como um de seus vários hospedeiros. Outras características peculiares aos organismos da família Trypanosomatidae são:

edição de RNA mitocondrial complexa e com alto consumo de energia, RNA polimerase I utilizada para transcrever genes codificadores de proteínas, "*trans-splicing*" de todos mRNA transcritos, organização dos genes em grande grupos policistrônicos, modificação de nucleotídeos sem precedentes, compartimentalização da glicólise, evasão da resposta do sistema imune do hospedeiro utilizando uma cobertura protéica e/ou glicídica de superfície variável, e a habilidade de escapar da destruição através da emigração do vacúolo fagocítico (Simpson, Stevens *et al.*, 2006; Teixeira, De Paiva *et al.*, 2012). Estas características únicas fazem destes organismos modelos interessantes para estudo de evolução genômica e outros aspectos da genômica funcional.

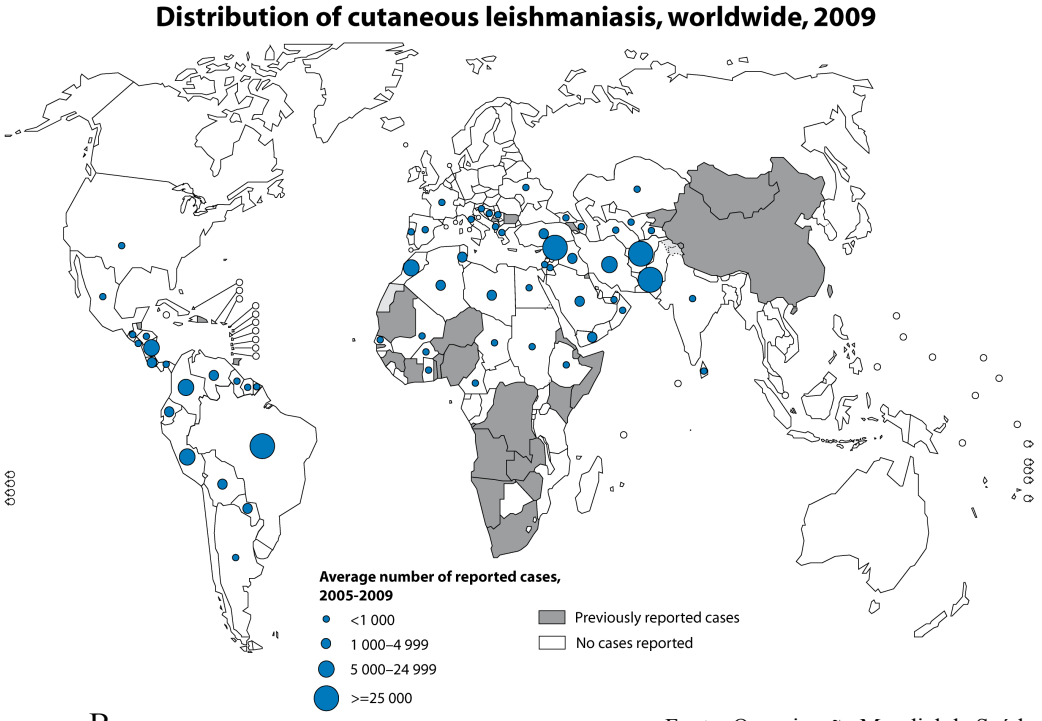
1.2– Gênero *Leishmania*

Mais especificamente sobre os organismos do gênero *Leishmania*, os quais são os alvos de estudo deste trabalho, estes são protozoários parasitos transmitidos pela picada de insetos do grupo dos flebotomíneos, que são endêmicos em regiões tropicais e subtropicais do globo. Mais de 20 espécies de leishmania são responsáveis por um amplo espectro de doenças conhecidas como leishmanioses (Herwaldt, 1999; Piscopo e Mallia, 2006).

De acordo com a Organização Mundial de Saúde (WHO – "*World Health Organization*"), estima-se que ocorram mais de 2 milhões de novos casos de leishmaniose por ano, com mais de 360 milhões de pessoas com risco de contrair a doença em 88 países dos 5 continentes (Figura 1A e 1B).



A

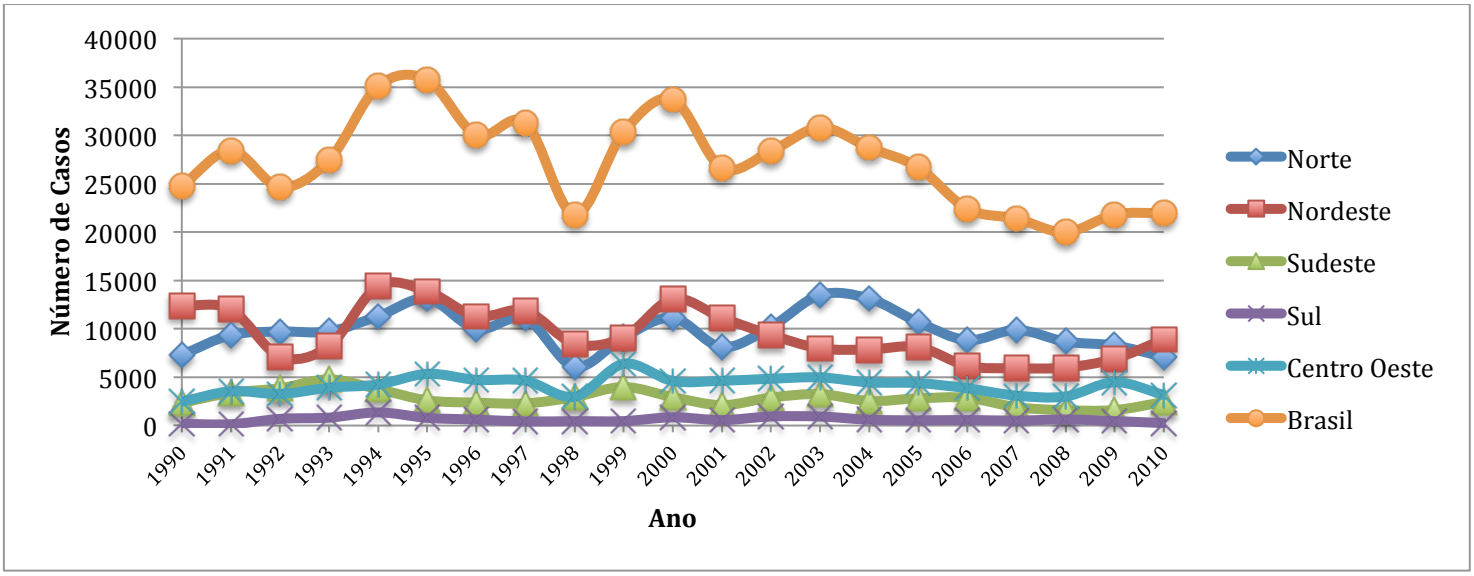


B

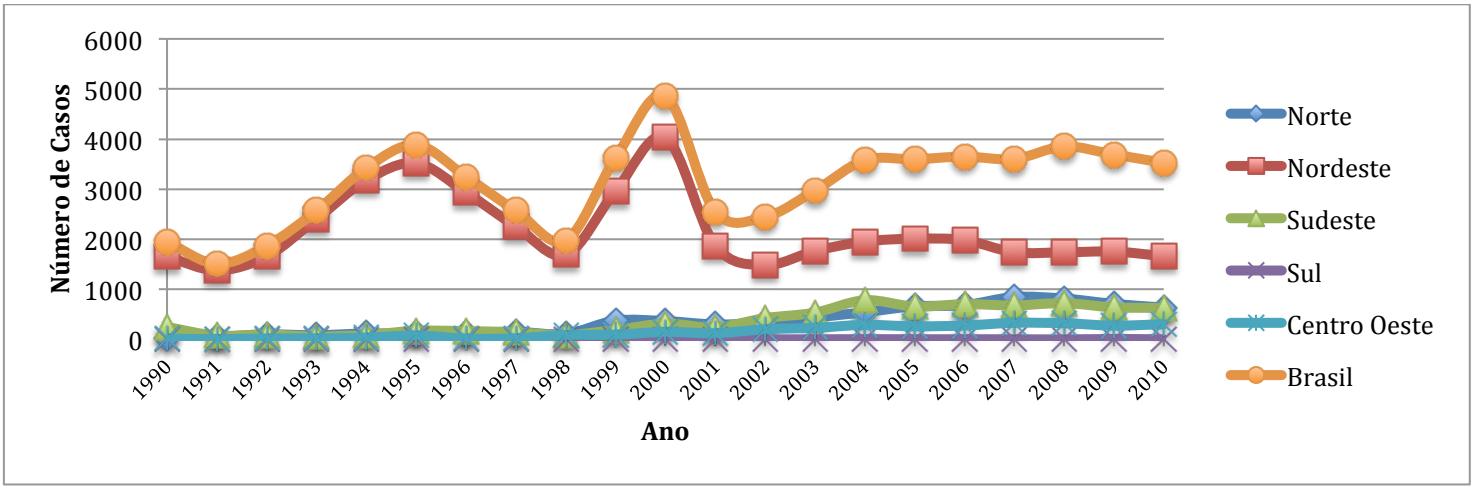
Fonte: Organização Mundial de Saúde

Figura 1 – Distribuição mundial da leishmaniose visceral (A) e da leishmaniose cutânea (B).

Especificamente no Brasil, a transmissão das leishmanioses vem sendo descrita em vários munícipios de todas as unidades federadas (Figura 2A e 2B). Além disso, as análises epidemiológicas têm sugerido mudanças no padrão de transmissão da doença, inicialmente considerada zoonose de animais silvestres, que acometia pessoas em contato com as florestas. Posteriormente, a doença começou a ocorrer em zonas rurais, já praticamente desmatadas, e em regiões periurbanas (Secretatia de Vigilância em Saúde - http://portalsaude.saude.gov.br/portalsaude/index.cfm?portal=pagina.visualizarArea&codArea=376).



A



B

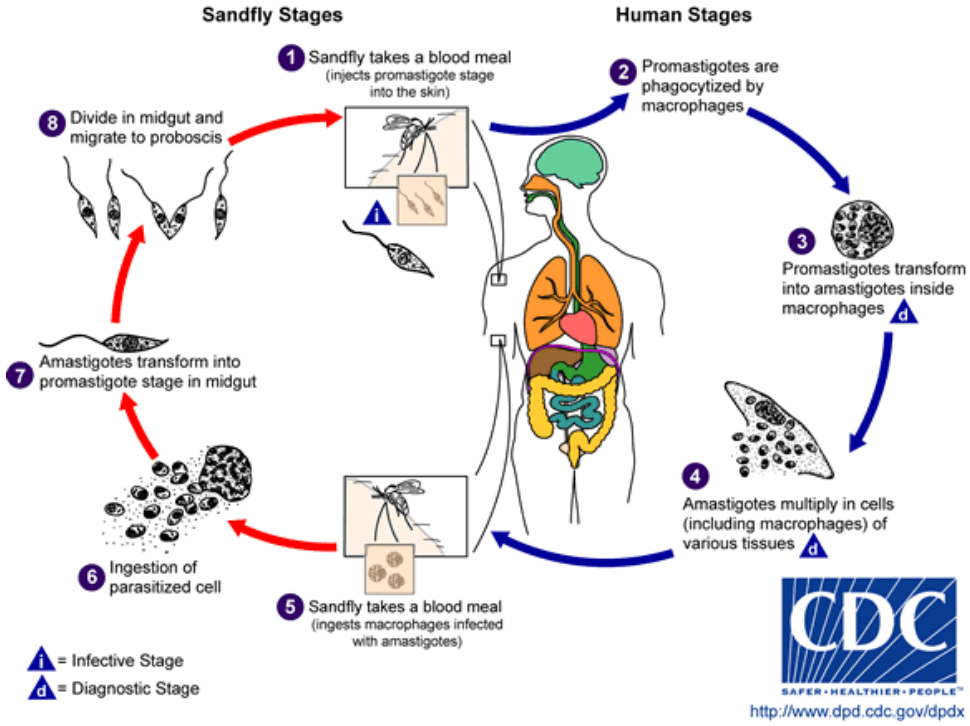
Fonte: Secretaria de Vigilância em Saúde

Figura 2 – (A) Número de casos de leishmaniose cutânea por região no Brasil de 1990 até 2010 e (B) número de casos de leishmaniose visceral por região no Brasil no mesmo período.

Outro aspecto importante da biologia destes parasitos é que os organismos pertencentes a este gênero podem ser classificados em dois subgêneros distintos de acordo com a parte do intestino do inseto flebótomo onde a colonização e o desenvolvimento do parasito ocorrem. O subgênero *Leishmania (Viannia)* consiste de parasitos que se desenvolvem na parte posterior do intestino. Por outro lado, o subgênero *Leishmania (Leishmania)* compreende parasitos que possuem como local

de desenvolvimento a porção anterior e média do intestino (Lainson, Ward *et al.*, 1977; Bates, 2007).

O ciclo de vida dos organismos do gênero *Leishmania* é alternado entre o trato alimentar do inseto vetor, onde o parasito desenvolve-se como promastigota flagelar extracelular e diferencia-se na forma infectiva não multiplicativa que é a forma promastigota metacíclica, e o fagolisossomo do hospedeiro vertebrado, onde o parasito diferencia-se na forma replicativa e aflagelada chamada amastigota (Figura 3).



Fonte: Centers for Disease Control and Prevention (CDC)

Figura 3 – Ciclo de vida dos parasitos do gênero *Leishmania* que infectam o homem

Como foi dito anteriormente, as leishmanioses compõe um grupo de doenças com amplo espectro de manifestações patológicas. De maneira geral elas podem ser divididas em três categorias diferentes: 1) Leishmaniose visceral – Esta é a forma mais grave da doença. Possui um período de incubação de 3 a 8 meses e tem como sintomas febre, perda de peso, hepatoesplenomegalia (usualmente baço muito maior

que o fígado), linfadenopatia, pancitopenia e hipergamaglobulinemia. Pigmentação na pele pode ser encontrada. Este tipo de leishmaniose pode ser causada pelas seguintes espécies de leishmania: *Leishmania infantum* e *Leishmania donovani*. 2) Leishmaniose Cutânea – Inicia-se como um pápula no local da picada do flebótomo que então aumenta o tamanho, formando um crosta, e eventualmente com ulcerações. Em 90% dos casos, são necessários de 3 a 18 meses para a cura. O período de incubação dura em torno de 2 semanas a vários meses. Espécies que podem causar este tipo de leishmanioses são: *Leishamania major*, *Leishmania tropica* e *Leishmania mexicana*. 3) Leishmaniose Muco-cutânea – O período de incubação é de 1 a 3 meses, porém este tipo de leishmaniose pode ocorrer muitos anos após o início da cura de uma leishmaniose cutânea. Geralmente esta patologia afeta as mucosas do nariz, cavidade oral e faringe. Esta tipo de manifestação é principalmente vinculada à espécie *Leishmania braziliensis* (Piscopo e Mallia, 2006).

1.2.1 – Genoma das espécies do gênero *Leishamania*

A primeira espécie do gênero *Leishamania* a ter o genoma sequenciado foi a *Leishamania (Leishmania) major* (Ivens, Peacock *et al.*, 2005). O seu genoma haploide possui aproximadamente 32,8 milhões de pares de bases (32,8Mb) distribuídos entre 36 cromossomos que variam de 0,28 a 2,8Mb (Wincker, Ravel *et al.*, 1996; Ivens, Peacock *et al.*, 2005). Dois anos mais tarde, os genomas das espécie *Leishmania (Leishmania) infantum* e *Leishmania (Viannia) braziliensis* também foram completamente sequenciados (Peacock, Seeger *et al.*, 2007). O primeiro também possui 36 cromossomos, enquanto o último possui apenas 35 cromossomos, número que é resultado da fusão entre os cromossomos 20 e 34 (referentes ao genoma

de *L.major*) (Britto, Ravel *et al.*, 1998). O genoma de *L. major* foi obtido através do sequenciamento de pequenos fragmentos aleatórios de grandes insertos clonados e purificados a partir do DNA cromossomal (Ivens, Peacock *et al.*, 2005). Já os genomas de *L. infantum* e *L. braziliensis* foram sequenciados a partir de uma abordagem de fragmentação de todo o genoma ("*whole-genome shotgun*") com uma cobertura de 5 e 6 vezes, respectivamente (Peacock, Seeger *et al.*, 2007). Segundo a base de dados TriTrypDB versão 4.1 e os trabalhos na literatura (Peacock, Seeger *et al.*, 2007), os genomas destas três espécies do gênero *Leishmania* possuem as características gerais descritas na Tabela 1.

Tabela 1 – Resumo dos genomas das três espécies alvos do estudo

	L. major	***L. infantum***	***L. braziliensis***
Número de Cromossomos	36	36	35
Tamanho	32,86Mb	32,13Mb	32,09Mb
Conteúdo G+C	59,7%	59,13%	57,76%
Genes Codificantes	8412	8241	8357
Psudogenes	93	54	156
Conteúdo de G+C Região Codificante	62,5%	62,45%	60.38%

É possível perceber através da Tabela 1, como os três genomas são semelhantes entre si. O tamanho, o conteúdo G+C tanto no genoma total quanto somente em regiões codificantes, e o número de genes são equivalentes. Estudos de genômica comparativa entre as três espécies surpreendentemente mostraram que existe muito pouca divergência entre os genomas de *L. major*, *L. infantum* e *L.*

braziliensis em termos de sequência e organização (Peacock, Seeger *et al.*, 2007; Lynn e Mcmaster, 2008), apesar de uma divergência dentro do gênero de 20 a 100 milhões de anos (Lukeš, Jirků *et al.*, 1997). Em relação à organização dos genomas, 99% dos genes dos três organismos são sintênicos (Peacock, Seeger *et al.*, 2007), e a maioria deles estão anotados como de função desconhecida. Esta alta conservação global das sequências genômicas e da sintenia indica que o genoma das Leishmanias é altamente estável e não sofreu ou está sofrendo grandes rearranjos genômicos durante a especiação (Teixeira, De Paiva *et al.*, 2012). Frente a esta alta conservação, é possível que as manifestações clínicas distintas observadas nas diferentes patologias causadas por estes parasitos tenham como explicação a expressão gênica diferencial que ocorre em vários estágios do ciclo de vida de cada um dos três parasitos (Teixeira, De Paiva *et al.*, 2012). Ainda sobre a conservação genômica entre as três espécies, a grande quantidade de proteínas comuns entre elas poderia ser utilizada para estudos de potenciais alvos para desenvolvimento de drogas e vacinas.

Apesar da alta conservação existente entre os três organismos alvos do estudo, existem alguns genes espécie específicos. Estes ocorrem geralmente em regiões onde não há sintenia, e consistem de membros de grandes famílias de antígenos de superfície. RNAs estruturais, retroelementos e famílias gênicas em expansão também estão associadas com quebras na conservação da sintenia gênica (El-Sayed, Myler, Blandin *et al.*, 2005).

Em relação a expressão gênica nas Leishmanias, é preciso primeiro citar que os genes nestes organismos estão em geral organizados em grupos gênicos direcionais que se assemelham às unidade de transcrição policistrônicas de procariotos (Martínez-Calvillo, Nguyen *et al.*, 2004). Este tipo de organização gênica e de transcrição policistrônica possuem implicações severas na regulação gênica, que deve então ser

baseada em mecanismos de controle pós-transcricional. Uma vez que a dependência de mecanismos de iniciação da transcrição baseados em promotores para o controle dos níveis de mRNA é muito reduzido, uma maior ênfase é dada para os mecanismos regulatórios pós-transcricionais por meio do controle da estabilidade e da tradução do mRNA, bem como da renovação do repertório protéico do organismo (Clayton, 2002). Um exemplo de mecanismo de regulação pós-transcricional importante no controle da expressão gênica das Leishmanias é o processamento do RNA policistrônico (pre-mRNA). Este RNA contém muitos genes transcritos, assim ele primeiramente é individualizado em mRNA monocistrônicos. Após este evento, sequências idênticas de 39 nucleotídeos, conhecida como "*splice leader*" (SL) são colocadas nas extremidades 5' de todos os mRNAs através de uma reação conhecida como "*trans-splicing*" (Liang, Haritan *et al.*, 2003). Posteriormente, ocorre uma reação de poliadenilação nas extremidades 3' dos mRNAs. Estas reações de processamento são guiadas através de regiões de polipirimidinas que estão presentes em cada região intergênica. Os mRNAs maduros são então exportados para o citoplasma, onde a estabilidade e eficiência de tradução do mesmos estão altamente vinculadas aos elementos presentes em suas regiões não traduzidas ("*untranslated region*" – UTR). A presença ou ausência de determinados elementos nestas regiões é resultado de "*splicing*" e sítios de poliadenilação alternativos, o que pode portanto alterar a expressão de determinado gene (Kolev, Franklin *et al.*, 2010; Nilsson, Gunasekera *et al.*, 2010; Siegel, Hekstra *et al.*, 2010).

1.2.2 – Tratamento e Controle das Leishmanioses

As leishmanioses representam um grupo de doenças com diversos sintomas e peculiaridades. Esta diversidade acaba por se refletir no diversos métodos e substâncias utilizadas nos tratamentos empregados atualmente.

A leishmaniose cutânea causada por espécies do gênero *Leishmania* do velho mundo geralmente apresentam auto cura dentro 2 a 4 meses (*L. major*) ou de 6 a 15 meses *(L. tropica*). Contudo, a doença causada por Leishmanias do novo mundo, apresentam uma auto cura de aproximadamente 75% para *L. mexicana*, 10% para *L. braziliensis* e 35% para *L. panamensis* (Murray, Berman *et al.*, 2005). Estas últimas geralmente são tratadas principalmente para evitar a disseminação, por exemplo o desenvolvimento para forma muco-cutânea, a recaída, e ainda para diminuir as possíveis cicatrizes em regiões do corpo que podem causar impacto psicossocial no paciente (Herwaldt, 1999). O tratamento geralmente é realizado ministrando compostos antimoniais via parenteral. Porém estes apresentam grande toxicidade e desencadeiam diversos efeitos colaterais tais como disfunção gastrointestinal, dores musculares difusas, enrijecimento das articulações, dor no local da injeção e aumento da diurese por perda transitória da capacidade de concentração urinária, até mesmo reações adversas graves como arritmias cardíacas e pancreatite (Herwaldt, 1999; Murray, Berman *et al.*, 2005).

Quanto à leishmaniose visceral, que é a forma mais grave, os compostos antimoniais também são utilizados como a primeira linha de tratamento para esta doença. Outros compostos mais eficientes com menor toxicidade também estão sendo utilizados. Um destes é a anfotericina B lipossomal, que é a droga mais potente disponível comercialmente, atuando nas formas promastigotas e amastigotas do parasito. Contudo, a mesma possui um elevado custo, o que a torna inviável para o

tratamento em países com alto índice de pobreza onde a doença é endêmica (Murray, Berman *et al.*, 2005). A miltefosina também é uma outra opção no tratamento da leishmaniose visceral. Ela possui como grande vantagem a administração via-oral, porém seu alto custo e a preocupação com seu uso inadequado levando assim ao aparecimento de resistência são suas maiores desvantagens (Murray, Berman *et al.*, 2005).

Deste modo, é possível perceber que grande parte das substâncias atualmente disponíveis para o tratamento das leishmanioses apresentam sérias desvantagens como alta toxicidade, grande espectro de efeitos colaterais e/ou alto custo. Outro sério problema está relacionado com o surgimento de resistência por parte dos parasitos contra estas substâncias. Vários mecanismos de resistência às drogas têm sido identificados em espécies do gênero *Leishmania*. Na Índia aproximadamente 60% dos pacientes com leishmaniose visceral não respondem ao tratamento com compostos antimoniais devido à resistência do parasito a essas drogas (Lira, Sundar *et al.*, 1999; Sundar, 2002). Assim, fica evidente a necessidade de estudos para a busca de novos fármacos e novos alvos dos parasitos.

Esta necessidade de novos estudos também pode ser percebida em relação ao desenvolvimento de vacinas para as leishmanioses. Atualmente, nenhuma vacina eficaz contra qualquer espécie do gênero *Leishmania* foi desenvolvida. Contudo, a possibilidade de uma vacina é atestada pelo fato de que muitos indivíduos que já se infectaram uma vez e foram curados são resistentes às manifestações clínicas da doença em uma reinfecção (Costa, Peters *et al.*, 2011). Além disso, uma antiga prática realizada no Oriente Médio chamada de "leishmanização", que consiste em inocular parasitas da espécie *L. major* vivos e virulentos em regiões do corpo que não são expostas, confere proteção contra leishmaniose cutânea (Nadim, Javadian *et al.*, 1983).

De acordo com Noazin e colaboradores (Noazin, Modabber *et al.*, 2008) boa parte das vacinas testadas em todo o globo apresentam como agente o parasito atenuado, vivo, geneticamente modificado ou morto por métodos físicos e/ou químicos. Porém, nenhuma vacina com eficácia suficiente para a utilização em humanos foi desenvolvida. Parte deste problema pode ser explicado pela complexidade da resposta imune desenvolvida pelo hospedeiro contra o parasito. Existe um balanço complexo de citocinas envolvidas na resposta imune celular que vão nortear o resultado do processo imune (Herwaldt, 1999; Murray, Berman *et al.*, 2005; Piscopo e Mallia, 2006; Costa, Peters *et al.*, 2011). Além disso, estudos contraditórios existem em relação ao papel dos anticorpos no envolvimento da resposta imune contra as Leishmanias (Costa, Peters *et al.*, 2011).

De acordo com Shane Crotty (Heller, 2009), a maioria das 25 vacina licenciadas para uso em humanos contra diferentes organismos são efetivas porque elas dirigem respostas imunes a diversos alvos de um patógeno. Assim, a capacidade de responder à múltiplos antígenos pode ser um requisito essencial de uma vacina eficaz para as leishmanioses (Costa, Peters *et al.*, 2011). Dada a complexidade dos parasitos do gênero *Leishmania*, o fato de que a resposta observada após a inoculação ou infecção seguida de cura (auto cura ou medicamentosa) em humanos é poli específica, e que a imunização com combinações de antígenos aparentemente confere um maior grau de proteção até então em primatas não humanos sugerem que um aumento no repertório de antígenos para estes parasitas é necessário (Costa, Peters *et al.*, 2011).

Assim, aplicações utilizando uma abordagem sistêmica para a descoberta de novos antígenos e alvos para fármacos se fazem necessárias. Estas aplicações podem ser alcançadas através da integração de diferentes conjuntos de dados, tais como

genoma, transcriptoma e proteoma. Atualmente esta abordagem biológica baseada na interdisciplinaridade focando no estudo de interações complexas em sistemas biológicos é chamada de Biologia de Sistemas ("*Systems Biology*"). A Biologia de Sistemas juntamente com a abordagem de Vacinologia Reversa, item a ser tratado posteriormente neste texto, possuem um grande potencial na sugestão de novos alvos para desenvolvimento de vacinas.

1.3– Biologia de Sistemas

Como resultado do grande avanço nos sequenciamentos genômicos, existe atualmente uma grande quantidade de dados dos mais diversos organismos pertencentes a vários ambientes. Contudo, com o objetivo de aproveitar todo potencial destes dados, nós temos que estar aptos a converter sequência genômica em conhecimento biológico. O primeiro passo envolve a predição de genes, o qual permite certo nível de anotação funcional utilizando predição de domínios, homologia ou ontologia. No entanto isto produz apenas uma lista de genes, e não expõe quais e como estes genes funcionam juntos. Isto é essencial para entender a complexidade codificada em um genoma (Harrington, Jensen *et al.*, 2008).

Além disso, estudos genômicos têm possibilitado, em um primeiro nível, o melhor entendimento sobre a composição e a evolução dos genomas, sintenia, manutenção e diversidade das famílias multigênicas e em um nível posterior, a melhor compreensão sobre parasitismo, virulência, epidemiologia e evolução do organismo (Harrington, Jensen *et al.*, 2008).

Assim, um grande desafio da biologia moderna é revelar a organização e as interações das redes celulares que possibilitam a existência de processos complexos

como a bioquímica do crescimento ou a divisão celular. A complexidade destes processos origina das interações dinâmicas entre uma grande quantidade de constituintes celulares, tais como genes, proteínas e metabólitos. Além disso, as interações entre estes componentes variam em natureza (regulatória, estrutural, e catalítica), efeito e força (Sauer, Heinemann *et al.*, 2007). Desta maneira, para sobrepor este desafio, existe a necessidade de uma abordagem mais ampla do que o paradigma de pesquisa biológica atual (reducionista).

Esta abordagem reducionista identificou com sucesso a maioria dos componentes e muitas interações, mas, infelizmente, não oferece nenhum conceito ou método para compreender como as propriedades do sistema emergem. Assim, ao contrário da visão reducionista, o pluralismo das causas e efeitos nas redes biológicas é mais bem gerenciado pela observação, através de medidas quantitativas, dos múltiplos componentes simultaneamente, e através da integração dos dados com modelos matemáticos. Esta perspectiva, então denominada Biologia de Sistemas, é necessária para que as propriedades das redes biológicas, tais como um estado funcional particular ou robustez, possam ser quantitativamente entendidas e racionalmente manipuladas (Sauer, Heinemann *et al.*, 2007).

Dentro deste contexto, a Biologia de Sistemas tem como objetivo integrar toda informação biológica disponível a fim de obter uma visão sistemática do organismo a ser estudado. Além disso, a Biologia de Sistemas possui como meta permitir a simulação de como as moléculas funcionam em coordenação para alcançar um resultado particular, e possibilitar alto poder de predição para resultados de perturbações ainda não estudadas (Pujol, Mosca *et al.*, 2010).

Um dos métodos utilizados para alcançar este objetivo na Biologia de Sistemas é o estudo de redes biológicas celulares, por exemplo, rede interação de

proteínas – proteínas ("*ppi networks*"). Estes estudos fornecem informação a respeito de quais proteínas de um genoma que se interagem e como elas o fazem (Harrington, Jensen *et al.*, 2008).

1.3.1 – Redes Biológicas Celulares

A célula de um organismo é um sistema complexo cujas características são definidas através da atividade de muitos componentes, que interagem uns com os outros através de interações par a par. Este componentes então podem ser representados por uma série de nós, os quais são conectados por ligações, que representam as interações. Nós e ligações juntos formam uma rede (Barabasi e Oltvai, 2004).

Existem diversos tipos de redes biológicas, desde redes representando vias metabólicas até redes que representam interações entre genes e fatores de transcrição que os regulam.

Estas redes ainda podem ser redes direcionais, isto é, aquelas na quais a informação representada possui uma direção (Figura 4A). Por exemplo, as redes regulatórias, em que fatores de transcrição regulam genes. Enquanto existem as redes não-direcionais, as quais a informação representada não possui nenhuma direção (Figura 4B). Por exemplo, as redes de interação proteínas, em que se representam as interações físicas entre proteínas.

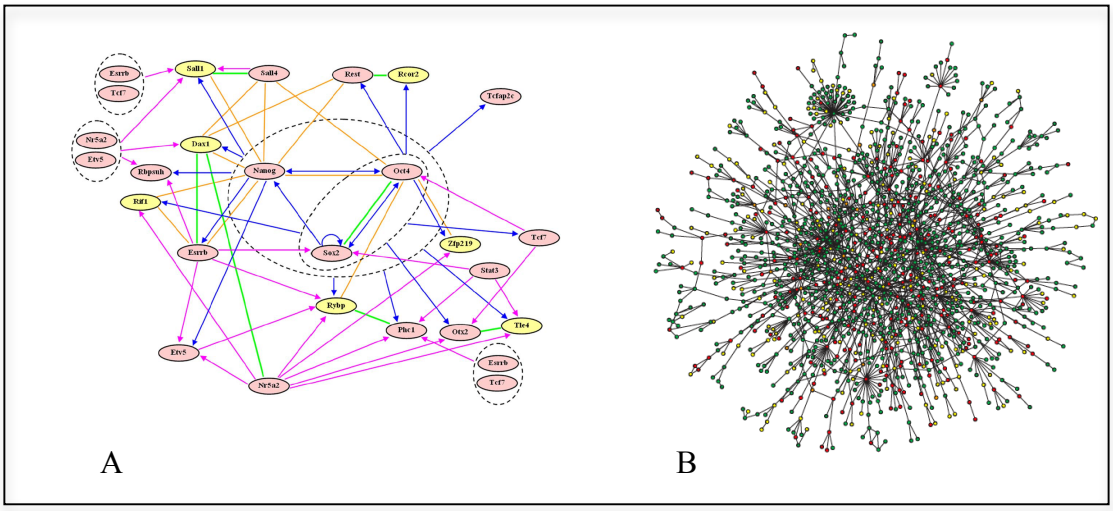
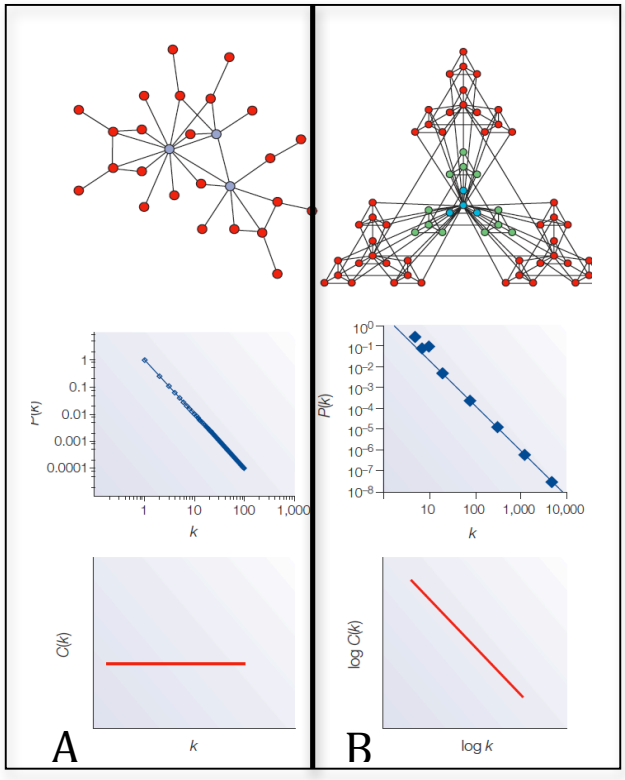


Figura 4 – (A) Rede regulatória de células tronco embrionárias de camundongo (Zhou, Chipperfield *et al.*, 2007). A interações entre os nós possuem direção. (B) Rede de interação de proteínas de levedura *S. Cerevisae* (Jeong, Mason et al., 2001). Não existe direção entre a interações dos nós

Além disso, as redes biológicas celulares apresentam algumas características topológicas em comum. A primeira delas é que estas redes possuem uma distribuição na probabilidade de encontrar um nó com um número *K* de interações denominada lei de potência ($P(K) \sim K^{-\gamma}$). Assim, existe um grande número de nós com poucas interações, e um pequeno número de nós com muitas interações. Esta distribuição dará origem a uma topologia de rede conhecida como livre de escala ("*Scale-free*") (Figura 5A) (Barabasi e Albert, 1999) .Esta topologia possui dois mecanismos básicos para sua geração. O primeiro deles é chamado de crescimento; durante este processo, novos nós são adicionados ao sistema. O segundo mecanismo é denominado agregação preferencial ("*Preferential attachment*"), no qual os novos nós possuem a preferência de se ligarem a nós que já possuem muitas ligações. Interessantemente, estes mecanismos já foram mapeados em fenômenos biológicos para redes biológicas celulares, mais especificamente para redes de interação de proteínas. Eles estariam ligados ao fenômeno de duplicação gênica, o qual produz duas proteínas iguais inicialmente e que interagem com o mesmo nó (crescimento). Portanto cada proteína

que está em contato com uma proteína duplicada recebe uma ligação extra. Consequentemente, proteínas altamente conectadas possuem uma vantagem natural. Elas possuem uma maior probabilidade de ter uma ligação a uma proteína duplicada do que as proteínas com um pequeno número de ligações, portanto as proteínas altamente conectadas são mais aptas a receber novas ligações (agregação preferencial) (Pastor-Satorras, Smith *et al.*, 2003).

Outra característica das redes biológicas celulares é referente à capacidade de cada nó ser agrupado; esta capacidade recebe o nome de Coeficiente de Agrupamento (C) ("*Clustering Coefficient*"). Assim, a distribuição do valor C(*K*), que é o valor médio de C para os nós com *K* interações, nas redes celulares segue também uma lei de potência ($C(K) \sim K^{-1}$). Esta propriedade é a assinatura mais importante de modularidade hierárquica existente nas redes celulares (Figura 5B). O conceito de modularidade assume que as funcionalidades celulares podem ser particionadas em uma coleção de módulos. Cada módulo é uma entidade discreta de muitos componentes elementares e que realiza uma tarefa identificável, separável de outras funções de outros módulos (Ravasz, Somera *et al.*, 2002). Esta arquitetura hierárquica pode ser definida por uma rede em que nós conectados de maneira esparsa são parte de áreas altamente agrupadas, com a comunicação entre os diferentes grupos sendo mantida por nós com alto número de interações chamados de *hubs* (Barabasi e Oltvai, 2004).



Fonte: (Barabasi e Oltvai, 2004)

Figura 5 – (A) Modelo de rede livre de escala, a distribuição de *P(k)* segue um lei de potência, no entanto *C(k)* é uniforme. (B) Modelo hierárquico, tanto a distribuição de *P(k)* quanto a de *C(k)* seguem uma lei de potência.

Uma terceira característica importante das redes biológicas celulares é denominada de "efeito de pequeno mundo" ("*Small-world effect*"), em que dois nós quaisquer podem ser conectados por um caminho de apenas poucas ligações (Barabasi e Oltvai, 2004). Isto indica que perturbações locais podem alcançar níveis superiores em uma rede celular rapidamente. Além disso, as redes celulares possuem uma propriedade denominada "*disassortative*", isto é, nós altamente conectados evitam se ligarem uns aos outros (Maslov e Sneppen, 2002).

Interessantemente, a soma destas características descritas acima, resulta em outra denominada robustez. Esta refere à capacidade que as redes biológicas celulares possuem para responder a mudanças na organização interna enquanto mantém um comportamento relativamente normal. Portanto, se em média 80% de nós selecionados aleatoriamente falharem, os 20% restantes ainda formam um grupo

compacto com caminhos conectando quaisquer dois nós. Isto é porque a seleção aleatória afeta na maioria das vezes nós com pouco número de interações. Porém os nós com muitas interações induzem a uma vulnerabilidade ao ataque (Barabasi e Oltvai, 2004).

Por fim, como todas essas características estão presentes nas redes biológicas, e as redes de interações de proteína fazem parte deste grupo, logo este tipo de rede irá apresentar as características topológicas mencionados acima.

1.3.2 – Redes de Interação de Proteínas

Retomando o que foi dito anteriormente, as redes de interação de proteína possuem em geral as mesmas características topológicas de outros tipos de redes biológicas celulares. Além disso, as redes de interação de proteína são redes do tipo não-direcional, ou seja, não existe direção no fluxo de informação representado no grafo, visto que o mesmo apresenta uma abstração de interações físicas entre proteínas.

Existem atualmente diversas metodologias experimentais para realizar a predição de redes interação de proteínas. Entre elas, esta a técnica de duplo híbrido utilizada para detectar interações físicas entre proteínas, e que atualmente com advento da miniaturização e robótica passou a ser empregada em larga escala. Simultaneamente, métodos de purificação por afinidade acoplados a espectrometria de massa de alta produção estão sendo utilizados na determinação de complexos proteicos (Harrington, Jensen *et al.*, 2008).

No entanto, estes métodos podem não ser geralmente aplicáveis para todas as proteínas em todos os organismos, e podem também ser propícios a erros sistemáticos.

Portanto, um grande número de abordagens computacionais tem sido desenvolvido para predição em larga escala de interações de proteína-proteína baseadas na sequência protéica ou nucleotídica (Skrabanek, Saini *et al.*, 2008). As abordagens mais conhecidas envolvendo sequências são: 1) Perfil filogenético ("Phylogenetic Profile"), 2) Vizinhança genômica ("Genome Neighbourhood"), 3) Fusão gênica ("Gene Fusion"), 4) Co-evolução de sequência ("Sequence Co-evolution"), 5) Homologia de sequência contra banco de interações, conhecida também como "*Interolog Mapping*".

Para iniciar uma breve descrição destas metodologias, iniciaremos pelo Perfil filogenético em que é utilizada a correlação entre a distribuição filogenética de duas proteínas. O raciocínio para este método é que se duas proteínas estão funcionalmente relacionadas, deve existir uma tendência para que as duas proteínas sejam co-herdadas, uma vez que a perda de uma das duas seria prejudicial para uma determinada função (Huynen e Bork, 1998; Pellegrini, Marcotte *et al.*, 1999).

A segunda abordagem aqui citada, vizinhança genômica, pode ser vista como uma extensão da abordagem de Perfil filogenético, porém além dela procurar pela tendência de certo genes serem co-herdados, ela também busca pela tendência destes estarem agrupados no genoma. Assim, deve existir uma pressão para manter estes genes próximos no genoma indicando uma relação funcional entre as proteínas codificadas por eles. É importante salientar que esta abordagem além de predizer interações entre as proteínas, ela também prediz relação funcional entre as mesmas (Harrington, Jensen *et al.*, 2008).

A próxima abordagem citada é Fusão gênica que é quando dois genes se fundem em uma única fase aberta de leitura. Além de permitir uma forte co-regulação da expressão, a fusão gênica pode levar ao aumento na eficiência de vias de

sinalização e vias bioquímicas através da acoplagem de passos sucessivos (Enright, Iliopoulos *et al.*, 1999; Marcotte, Pellegrini *et al.*, 1999).

A quarta metodologia baseada em sequência é chamada de Co-evolução de sequência a qual assume que proteínas que interagem possuem árvores filogenéticas moleculares similares porque existe uma co-evolução mantida pela interação (Sato, Yamanishi *et al.*, 2005). Assim, as taxas evolutivas das proteínas são comparadas para selecionar os pares de interação.

Por fim, a última metodologia citada é baseada em homologia de sequência contra banco de interações ("*Interolog Mapping*") (Matthews, Vaglio *et al.*, 2001; Yu, Luscombe *et al.*, 2004; Kim, Park *et al.*, 2008; Florez, Park *et al.*, 2010). Assim as sequências das proteínas de interesse são comparadas contra as sequências das proteínas dos bancos de interações, e posteriormente as interações entre as proteínas de interesse são mapeadas utilizando o resultado das comparações.

No presente projeto, a abordagem conhecida como "*Interolog Mapping*" foi utilizada para realizar a predição e a modelagem das redes de interação de proteínas para os organismos alvos de nosso estudo (*L. braziliensis*, *L. infantum* e *L. major*).

1.3.3 - Aplicação dos Estudos de Redes de Interação de Proteínas

Existem atualmente diversas aplicações para o estudo das redes biológicas celulares. Muito mais do que conhecer a biologia complexa de um organismo, as redes celulares estão sendo utilizadas para busca de alvos terapêuticos para drogas e vacinas.

Uma das aplicações das redes celulares neste sentido é o seu uso na busca de alvos e ou marcadores para diferentes tipos de câncer. Pujana e colaboradores (Pujana,

Han *et al.*, 2007) identificaram novos genes associados com risco maior de câncer de mama. Além disso, Chuang e colaboradores (Chuang, Lee *et al.*, 2007) extraíram propriedades funcionais de proteínas diretamente do estudo topológico de redes de interação para identificar marcadores de metástase de câncer de mama.

Outro estudo importante utilizando redes de interação está vinculado ao trabalho sobre o interatoma de linfócito B humano, que identificou interações desreguladas em fenótipos patológicos específicos (Mani, Lefebvre *et al.*, 2008).

Dentro da parasitologia, foram identificados um grupo de interação de proteínas no interactoma de *Plasmodium falciparum*, um dos agentes patológicos da malária, relacionado a invasão celular (Lacount, Vignali *et al.*, 2005). As proteínas deste grupo podem ser potenciais alvos para vacina. Outro trabalho na área de parasitologia recentemente publicou uma lista de candidatos alvos de droga em *L. major* utilizando uma rede de interação para este organismo (Florez, Park *et al.*, 2010).

Até mesmo na fitoparasitologia, o estudo de redes celulares tem sido aplicado, como por exemplo, o estudo da rede de interação do fungo *Magnaporthe grisea*, que é responsável pela perda de 10 a 30 por cento da produção anual de arroz (He, Zhang *et al.*, 2008). Neste estudo, foi possível identificar características topológicas das proteínas que estão relacionadas com a patogenicidade do fungo.

Portanto é possível notar o grau de importância para estudos que envolvam redes biológicas celulares. Mais especificamente aqueles cujo objetivo é a busca de alvos para drogas e vacinas. Assim, dentro deste contexto, o principal meio de identificar estes alvos é a utilização de índices topológicos, além das análises de agrupamento das redes.

1.3.4 - Índices Topológicos

Um dos principais questionamentos realizados frente a uma rede de interação de proteína é como rastrear os diversos efeitos que decorrem da neutralização de alguma proteína, e assim determinar sua essencialidade no sistema. Algumas pessoas preferem focar principalmente na análise de interações par-a-par diretas e consideram os efeitos indiretos menos importantes. Outras tendem a dar ênfase às consequências dos efeitos indiretos visto que praticamente todas as proteínas estão conectadas com todas. Vários efeitos de perturbação se espalham em uma rede de interação de proteína, e ambas propriedades globais e a posição local de uma proteína específica fortemente influenciarão no que acontecerá na rede. Obviamente, vizinhos diretos de uma proteína canalizarão a dispersão dos efeitos de sua neutralização. Portanto a estrutura de uma rede limita quem afetará quem e qual será a extensão da perturbação. Para tentar entender assim os efeitos que podem ser causados em uma rede a partir da neutralização de uma proteína, os índices topológicos são utilizados.

Índices topológicos são medidas atribuídas para cada nó que levam em consideração a posição e o papel que um determinado nó ocupa em uma rede. Existem diversos índices topológicos tais como: Degree, Bottle-neck ou Betweenness (Freeman, 1977), Trophic Index (Jordán, Liu *et al.*, 2003), Trophic Field Overlap (Ferenc, Liu *et al.*, 2009), etc.

Neste trabalho foram utilizados os índice Degree e MCC ("*Maximal Clique Centrality*"). O Degree é a própria contagem do número de vizinhos diretos que um nó possui. Um nó com alto valor de Degree é denominado *hub*.

Já o MCC é a medida relacionada ao quão central um nó é para vários cliques (módulos ou subgrafos), além disso, o tamanho dos cliques influencia diretamente o

seu valor. Assim, se um nó com um alto valor de MCC for neutralizado provavelmente vários módulos de uma rede serão afetados substancialmente. A Figura 6 ilustra como o MCC é calculado. Assumindo que uma rede possua *K* cliques, onde o nó *v* está inserido, e que Z_i corresponda ao tamanho do clique *i* incluindo o nó *v* para qual o MCC está sendo calculado. Assim, o *MCC(v)* será encontrado utilizando a seguinte fórmula:

$$MCC(v) = \sum_{i=1}^{K} (Z_i - 1)!$$

Assim, a utilização dos índices acima citados é uma maneira de tentar correlacionar características topológicas de um determinado nó com suas características funcionais em um sistema. No caso deste projeto, existe a tentativa de correlacionar o Degree e o MCC com a potencialidade de uma proteína ser alvo para novas drogas e vacinas.

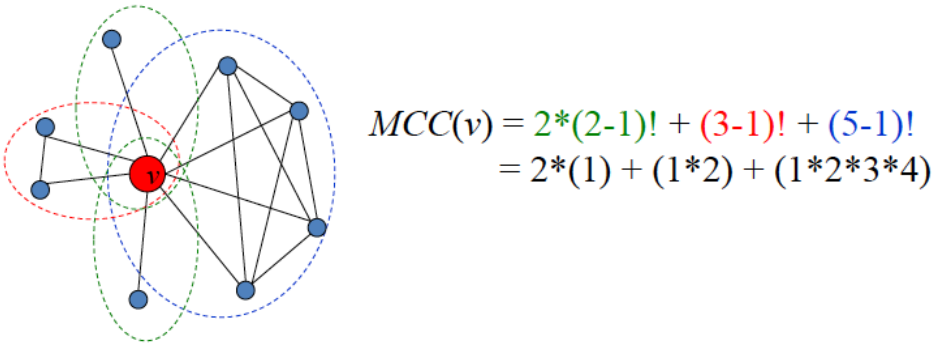


Figura 6 – Ilustração do cálculo do índice MCC para o nó *v* (vermelho)

1.3.5 – Contexto de Uma Proteína na Rede de Interação e Sua Diversidade

Como foi dito na seção 1.3.1, as redes de interação de proteína são redes que se encaixam em um modelo livre de escala. Assim, a probabilidade *P(k)* que uma proteína na rede interaja com *k* outras proteínas decai como uma lei de potência,

seguindo $P(k) \sim k^{-\gamma}$. Foi dito também que são necessários dois ingredientes para que a rede se desenvolva obedecendo o modelo livre de escala, são eles: o crescimento e a agregação preferencial. Além disso, mencionamos que estes fenômenos já foram mapeados no contexto biológico e que estariam relacionados ao fenômeno de duplicação gênica. Este modelo de desenvolvimento das redes de proteínas sugerem então que proteínas mais ancestrais e conservadas são as mais conectadas (*hubs*) (Fraser, Hirsh *et al.*, 2002; Wuchty, 2004; 2006). Uma vez estas proteínas sendo altamente conectadas, elas exerceram um papel importante nas redes de interação. Assim, um pressão seletiva existe sobre elas.

A conservação dos *hubs* e assim a importância dos mesmos tem sido corroborada na literatura. Foi visto que proteínas *hubs* em *Saccharomyces cerevisae* possuem uma menor distância evolutiva em relação a seus ortólogos em *Caenorhabditis elegans* e possuem uma maior probabilidade de terem ortólogos em outros organismos (Fraser, Hirsh *et al.*, 2002; Krylov, Wolf *et al.*, 2003). Assim, as correlações observadas entre a essencialidade e a conectividade de uma proteína sugerem que *hubs* são provavelmente conservados na evolução, implicando na emergência de um correlação entre o número de interações de um proteína e sua conservação evolutiva.

1.3.6 - Análise de Modularidade

Como também foi mencionado na seção 1.3.1, as redes de interação de proteína são redes hierárquicas. Logo, possuem uma estrutura determinada em módulos. Alguns pesquisadores argumentam que os módulos podem ser considerados como um nível crítico na organização biológica (Hartwell, Hopfield *et al.*, 1999). Os

módulos podem ser compostos de muitos tipos de moléculas, quando existem tipos diferentes de rede interagindo. Eles possuem funções discretas que originam da interação entre seus componentes (Hartwell, Hopfield *et al.*, 1999).

Esta organização modular permite que a função principal de um determinado módulo seja robusta à alterações. Porém, permite mudanças em propriedades e funções de uma célula através de alterações nas conexões entre os diferentes módulos (Hartwell, Hopfield *et al.*, 1999; Barabasi e Oltvai, 2004).

Contudo, as funções de cada módulo não podem ser preditas a partir dos estudos das propriedades isoladas de seus componentes. Deve se tentar entender o módulo como um todo. Assim, para realizar a análise de modularidade nas redes de interação de proteína, diversos algoritmos de agrupamento podem ser utilizados (Blatt, Wiseman *et al.*, 1996; Enright, Van Dongen *et al.*, 2002; Bader e Hogue, 2003; King, Przulj *et al.*, 2004; Sharan, Suthram *et al.*, 2005). A aplicação destes algoritmos em primeira instância leva à identificação dos módulos presentes nas redes. Posteriormente, é necessário a caracterização dos módulos identificados. Para isso, é desejável um vocabulário controlado sobre processos biológicos. Felizmente, atualmente a comunidade científica já dispõe deste tipo de recurso, e ele é denominado Gene Ontology (Ashburner, Ball *et al.*, 2000). Deste modo, utilizando ferramentas para mapear com um determinado rigor estatístico os termos referentes aos processos biológicos de cada módulo, é possível caracterizá-lo, e sobrepor esta informação com as informações topológicas referentes a cada nó presente no módulo.

1.4 - Anotação de Proteínas Hipotéticas

De uma maneira geral, as proteínas hipotéticas são definidas como proteínas preditas computacionalmente a partir de sequências de nucleotídeos, geralmente dos genomas, porém não existem evidências experimentais de que elas existam. Além disso, não existem proteínas com função determinada que sejam similares às proteínas hipotéticas.

O termo "proteína hipotética conservada" é também amplamente empregado e descreve uma fração de produtos gênicos em genomas sequenciados que são encontrados em organismos de diversas linhagens filogenéticas porém novamente não foram funcionalmente caracterizados e descritos ao nível químico de aminoácidos.

Como foi dito anteriormente, as proteínas hipotéticas não possuem proteínas similares a elas com função determinada. Assim, o método clássico de atribuição de função a uma proteína, que é através de similaridade de sequência, não é eficaz quando aplicado neste conjunto de dados.

Deste modo, as redes de interação de proteína surgem como mais uma opção para realizar a predição de função das proteínas hipotéticas. O racional por de trás desta abordagem é que proteínas que interagem provavelmente compartilham de funções similares (Hishigaki, Nakai *et al.*, 2001). Assim, se uma proteína hipotética interage com uma proteína de função conhecida, a primeira pode ter sua função predita pela segunda. Este raciocínio é razoável uma vez que foi observado que de 70 a 80% das proteínas compartilham um função com seus vizinhos de interação (Titz, Schlesner *et al.*, 2004).

No entanto, utilizar somente os vizinhos diretos limita a predição de função às proteínas que tem pelo menos um interação com um parceiro com função conhecida

(Chua, Sung *et al.*, 2006). Portanto, a utilização de interações indiretas também é explorada na predição de função de proteínas hipotéticas utilizando redes de interação. Em muitos casos, é observado que uma proteína não compartilha nenhuma função com seus vizinhos diretos, porém exibe um alto grau de similaridade de função com os vizinhos indiretos (Chua, Sung *et al.*, 2006).

Deste modo, este tipo de análise é muito útil no auxílio do processo de predição funcional de proteínas em genomas que possuem alta taxas de proteínas hipotéticas, como é o caso dos genomas das três espécies do gênero *Leishmania* deste estudo.

1.5 - Análise de Predição de Epítopos

A vacinologia reversa utiliza as sequências genômicas de vírus, bactérias ou protozoários parasitos de interesse médico ao invés das células como material inicial para a identificação de novos antígenos, cuja atividade deve ser subsequentemente confirmada por abordagens experimentais (Bambini e Rappuoli, 2009). Uma maneira mais eficiente de aplicar os princípios da vacinologia reversa é a aplicação da mesma em proteínas que se destacaram nas análise topológicas e de modularidade nas redes de interação de proteínas. Deste modo, o universo de busca fica reduzido, potencializando as chances de sucesso.

Pizza e colaboradores em colaboração com O Instituto de Pesquisa Genômica (TIGR – "*The Institute for Genomic Research*") relataram o primeiro exemplo de sucesso da aplicação da abordagem da vacinologia reversa (Pizza, Scarlato *et al.*, 2000). Eles descrevem que a identificação *in silico* de candidatos a vacina contra *Neisseria meningitides* grupo sorológico B, que é a principal causa de meningite em

crianças e adultos, foi possível enquanto abordagens convencionais para obter a vacina falharam por décadas.

Deste modo, a Imunoinformática é uma área emergente das técnicas de bioinformática que tem como foco a estrutura, função e interação de moléculas envolvidas na imunidade. Um dos seus principais objetivos é a predição *in silico* de epítopos imunogênicos. Assim, ferramentas de predição *in silico* e bases de dados voltadas para imunologia recentemente desenvolvidas podem ser utilizadas para identificar, caracterizar ou predizer epítopos antigênicos reconhecidos por células T e B, as quais possuem papéis significativos na infecção e na imunidade protetora (Korber, Labute *et al.*, 2006).

Epítopos são unidades essenciais mínimas de informação derivados de proteínas do próprio organismo ou não, e que estimulam respostas imunológicas celulares (célula T) e/ou humorais (célula B). A célula T reconhece epítopos que são derivados de proteínas endógenas e exógenas, que foram processadas, e que estão presentes nas fendas das moléculas de MHC classe I ou MHC classe II (MHC - "*Major Histocompatibility Complex*") na superfície das células apresentadoras de antígenos para receptores de células T. Após a ativação de células T CD8+ ou células T CD4+, respectivamente, eventos celulares ocorrem tais como ativação da citotoxicidade e secreção de citocinas. As células B também reconhecem epítopos, porém este reconhecimento é feito em proteínas intactas. Os epítopos reconhecidos por células B podem ser lineares, aminoácidos contíguos, ou discontíguos, isto é, aminoácidos que estão juntos devido a estrutura tridimensional da proteína. Após a ativação, as células B diferenciam-se em plasmócitos, e iniciam a secreção de anticorpos. As respostas imunológicas de células B e T são denominadas humoral e

resposta imune celular adaptativa, respectivamente (Borja-Cabrera, Cruz Mendes *et al.*, 2004).

Um grande variedade de técnicas de aprendizado de máquina estão sendo comumente utilizadas na bioinformática para predição de epítopos, algumas delas são: Redes Neurais Artificiais (ANN – "*Artificial Neural Networks*") (Baldi e Atiya, 1994), Modelos Ocultos de Markov (HMM – "*Hidden Markov Models*") (Hughey e Krogh, 1996) e Vetor de Suporte de Máquinas (SVM – "*Support Vector Machine*") (Vapnik e Vashist, 2009). ANN e SVM são idealmente aptos em reconhecer padrões não lineares, o que pode contribuir para o reconhecimento de epítopos de célula T (Lundegaard, Lund *et al.*, 2007). Por outro lado, métodos empregando HMM são aptos em caracterizar motivos biológicos com um composição estrutural inerente, e estes têm sido utilizados no campo da imunologia para predizer peptídeos com afinidade de ligação para molécula de MHC classe I (Mamitsuka, 1998).

1.6 - Banco de Dados Relacional

O conceito de banco de dados relacional surge na década de 70 (Codd, 1970). Dentro deste paradigma, tabelas representam entidades, ou objetos do mundo real que se relacionam entre si. Em 1976, Chen (Chen, 1976) propõe um modelo gráfico para auxiliar os profissionais a lidarem com o novo paradigma de armazenamento de dados. Este modelo é chamado de modelo Entidade-Relacionamento ou MER. Uma das principais dificuldades em trabalhar com este modelo é o estabelecimento correto da relação entre as entidades. As entidades e os relacionamentos juntos podem ser então convertidos em um Modelo Relacional (MR). A simplicidade do MER na representação de dados, e a possibilidade de converter um MER em um MR

propiciaram o surgimento de um conjunto de programas, os Sistemas Gerenciadores de Banco de Dados, ou SGDB (DBMS – "*Data Base System Management* "). Um SGDB é um conjunto de programas, que trabalha de forma integrada e sincronizada para viabilizar o armazenamento, a manipulação e a recuperação de dados. Atualmente os SGDB mais conhecidos são: MySQL, PostGre e OracleDB.

Para se ter acesso ao dados armazenados em um banco de dados relacional, os SGDB geralmente utilizam de uma linguagem específicas para consultas. Esta é conhecida como SQL ("*Structured Query Language*").

Atualmente com o crescimento da quantidade de dados nos mais diversos ramos de atuação, os banco de dados relacionais tem sido ferramentas essenciais para armazenamento e extração de conhecimento. Dentro da biologia, os uso deste modelo de armazenamento tem sido constante. Conceitualmente, o banco de dados dever ser apto a fornecer de modo simples acesso aos resultados experimentais, evitando redundância de dados de pesquisa. Um banco de dados cujo o MER foi bem desenhado irá fornecer apoio ao pesquisadores, facilitando buscas por novas correlações entre os dados armazenados. Por outro lado, um MER mau concebido torna o processo de mineração de dados difícil, e a integração de novos dados inviável (Shekhar e Chawla, 2003). Portanto, dentro desta perspectiva, os processos de reconstrução e de redesenho são frequentes (Shekhar e Chawla, 2003). El-Tabakh e colaboradores (Eltabakh, Ouzzani *et al.*, 2006) mencionam que a falta de uma ferramenta de banco de dados que possibilite trabalhar de maneira mais alinhada às necessidades da área biológica tem sido um fator negativo no progresso científico deste campo de pesquisa. Assim, a utilização de banco de dados relacional modelado adequadamente é um pré-requisito essencial para a construção de uma base

tecnológica mais viável ao tratamento e interpretação das predições biológicas realizadas em *in silico* ou através de abordagens experimentais.

O atual desafio da biologia moderna é revelar e entender os sistemas complexos de organizações biológicas, e sinalizar em todos seus detalhes a um nível molecular. Uma porção essencial deste processo é realizado através da bioinformática, particularmente através da integração dos mais diversos tipos de dados, utilizando como recursos a mais variadas linguagem de programação (PERL, Java, Python, C , R) e SGDBs. Os dados biológicos residem em banco de dados especializados que representam estágios de interpretação dos dados ou diferentes facetas de um fenômeno biológico. Além disso, os banco de dados biológicos apresentam uma peculiaridade: Eles são altamente complexos quando comparados com os dados da grande maioria das aplicações. Portanto, as definições de tais bancos biológicos devem ser aptas a representar um subestrutura complexa de dados, bem como as relações entre eles, e também assegurar que nenhuma informação seja perdida durante a modelagem dos dados. O modelo de dados deve ser capaz de representar qualquer nível de complexidade em qualquer esquema, relacionamento, ou esquema de subestrutura e não somente em formato de dados hierárquico, binário ou tabular.

2 – Justificativa

Os organismos *L. braziliensis*, *L. infantum* e *L. major* são patógenos intracelulares, que causam um grande impacto na saúde pública da América Latina e em outras regiões do globo. Além disso, a ampla ocorrência e o impacto econômico devido à morbidade causada pelas doenças que os mesmos provocam têm impulsionado os estudos de suas patogêneses.

Evidências muito sólidas têm levado à conclusão que somente abordagens integradas de estudo poderão levar à um controle eficaz das enfermidades causadas por esses organismos, assim sendo, é de fundamental importância uma abordagem de Biologia de Sistemas, a qual tem como objetivo integrar dados genômicos, transcriptômicos e proteômicos, para alcançar um entendimento amplo da biologia dos parasitos. Isto possibilitará a elaboração de abordagens mais eficientes para busca de novos alvos para desenvolvimento de drogas e vacinas.

3 – Objetivos

3.1 – Objetivo Geral

O objetivo geral deste projeto é modelar e analisar redes de interação de proteínas de três espécies de parasito do gênero *Leishmania* (*L. braziliensis*, *L. infantum* e *L. major*) à partir de seus proteomas preditos.

3.2 – Objetivos Específicos

Como objetivos específicos, este projeto possui:

1) Avaliar o método de predição de interações de proteínas, que é baseado em informação de interação de proteínas de outros organismos presentes em bases de dados públicas conhecido como *Interolog Mapping*.
2) Predizer as interações protéicas para os proteomas preditos dos três organismos alvos deste estudo, utilizando a abordagem avaliada.
3) Análisar as redes de proteínas preditas frente a diversos modelos de rede tais como: "*Scale-free*", hierárquico e aleatório.
4) Obter a anotação funcional das proteínas presentes nas redes modeladas utilizando como vocabulário controlado a ontologia do consórcio Gene Ontology.
5) Realizar análise de agrupamento e alinhamento para identificação de módulos funcionais conservados nas redes modelas.
6) Executar análises topológicas nas proteínas presentes nas redes preditas utilizando índices topológicos tais como: *Degree* e MCC.
7) Verificar relação entre a diversidade das proteínas presentes nas redes e o número de interações que estas fazem.
8) Melhorar o nível de anotação funcional das proteínas sem anotação predita (proteínas hipotéticas) a partir do contexto das mesma nas redes de interações de proteínas modeladas.
9) Avaliar diferentes métodos de predição de epítopos com afinidade de ligação para receptores de célula B (anticorpos) e de célula T (MHC classes I e II).
10) Aplicação dos preditores de epítopos avaliados nas sequências de aminoácidos das proteínas mais bem ranqueadas nas análises topológicas.

11) Construção de um banco de dados relacional para integração do dados e extração de conhecimento.
12) Produzir uma lista de potenciais alvos para desenvolvimento de drogas e vacinas baseada na integração dos dados gerados.

4 – Materiais e Métodos

A descrição da metodologia utilizada para a realização deste projeto está exposta nesta seção. Um fluxograma representando toda a abordagem empregada pode ser visualizado na Figura 7.

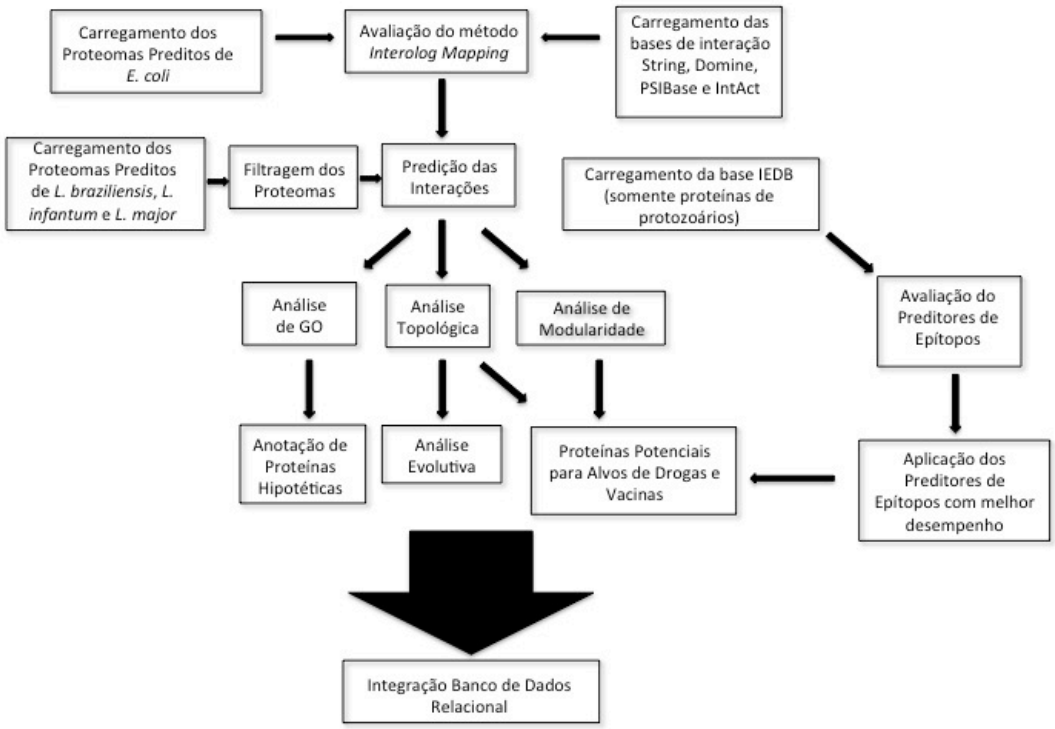


Figura 7 – Fluxograma representando a metodologia utilizada no desenvolvimento do projeto.

4.1 – Avaliação do Método de Predição das Redes de Interação

Com o objetivo de avaliar a eficiência do método empregado para predição das redes de interação de proteínas deste trabalho, uma avaliação de desempenho do mesmo foi conduzida.

Para a realização de uma análise de desempenho de qualquer metodologia, são necessários o conjunto de dados que representa o controle positivo e o conjunto de dados que representa o controle negativo. Especificamente para este projeto, o controle positivo é formado por um conjunto de interações de proteínas descritas experimentalmente. Este conjunto de dados foi extraído do DIP ("*Database of Interacting Proteins*" - http://dip.doe-mbi.ucla.edu/dip/Main.cgi) (Xenarios, Rice *et al.*, 2000). A base de dados DIP possui interações entre proteínas determinadas experimentalmente, integra informações de diversas fontes, e é manualmente curada por especialistas da área de interações protéicas.

Além disso, devido a alta consistência da informação presente nesta bases de dados, e levando em conta a quantidade de informação relativa à rede de interação de proteínas de *Escherichia coli*, a rede de proteínas deste organismo foi selecionada para compor o conjunto de dados pertencente ao controle positivo. Especificamente sobre a formação do controle positivo, dois pontos presentes no trabalho de Muley e Ranjan (Muley e Ranjan, 2012) foram considerados para a seleção das interações que compuseram o controle positivo. Entre eles: somente foram consideradas interações protéicas descritas como interações físicas; e as interações aqui selecionadas foram descritas pelo menos por um experimento de pequena escala ("*low throughput*"). Assim, foram selecionadas 702 interações para compor o controle positivo, nestas estavam presentes 530 proteínas do proteomas de *E. coli*.

Como foi mencionado anteriormente, para a realização de uma análise de desempenho, além do controle positivo, também é necessário o controle negativo. Este conjunto de dados foi construído por meio de algumas orientações presentes em trabalhos da literatura científica (Gomez, Noble *et al.*, 2003; Jansen, Yu *et al.*, 2003; Jansen e Gerstein, 2004; Zhang, Wong *et al.*, 2004; Qi, Klein-Seetharaman *et al.*,

2005). Uma seleção aleatória de interações protéicas foi realizada com base em todas as interações protéicas possíveis a partir do proteoma predito de *E. coli* subtraindo as interações experimentalmente validadas no DIP, independente do tipo de experimento. Além disso, somente pares de proteínas contendo ambas proteínas localizadas em diferentes localizações sub-celulares foram mantidos. Uma proporção de 1:5 entre pares de interações positivos e negativos foi utilizada. Assim, o conjunto de dados que compuseram o controle negativo continha 3.510 interações de proteínas.

Finalmente, com as interações protéicas representando o controle positivo e o controle negativo e o proteoma do organismo modelo, neste caso a bactéria *E. coli*, foi possível identificar, a partir das predições realizadas pelo método de predição de interação de proteína empregado neste trabalho, as predições verdadeiro positivas (TP – "*true positive*") e as verdadeiro negativas (TN – "*true negative*"). Além disso, foram mapeadas as interações falso positivas (FP – "*false positive*") e as falso negativas (FN – "*false negative*"). Com estes valores em mãos, a análise de desempenho foi realizada através da metodologia da curva ROC ("*Receiver Operating Characteristic*") utilizando o pacote ROCR para o ambiente de programação R (http://www.r-project.org/) (Sing, Sander *et al.*, 2005). A curva ROC é um gráfico da taxa de falso positivo (FPR – "*False Positive Rate*") contra a taxa de verdadeiro positivo (TPR – "*True Positive Rate*"), que é também conhecida como sensibilidade, para a predição de uma determinada metodologia. O valor de FPR possui relação com a especificidade do método a ser avaliado, uma vez que subtraindo este valor de 1 teremos então o valor de especificidade. Para o cálculo do FPR e do TPR, as seguinte fórmulas são aplicadas:

$$FPR = FP/(FP+TN)$$

$$TPR = TP/(TP+FN)$$

Uma metodologia cuja a predição seja aleatória terá um valor de área sob a curva ROC de 0.5. Por outro lado, uma metodologia com um predição perfeita terá um valor de 1 (Muley e Ranjan, 2012). Esta medida de área é chamada de AUC ("*Area Under Curve*"), e serve de referência para a avaliação do desempenho de um método através da curva ROC.

4.2 – Filtragem dos Dados

Anteriormente ao início da predição das redes de interação de proteínas nos três organismos alvos, uma filtragem foi realizada em seus proteomas preditos. Esta filtragem teve como objetivo remover possíveis erros de anotação presentes nos genomas dos organismos.

As versões dos genomas aqui utilizadas foram versão 2, versão 3 e versão final para *L. braziliensis*, *L. infantum* e *L. major*, respectivamente. Estas foram carregadas para os nossos servidores a partir da base de dados TriTrypDB (http://tritrypdb.org/tritrypdb/) (Aslett, Aurrecoechea *et al.*, 2010). Mais detalhadamente, três critérios foram utilizados para esta filtragem. Primeiro, as sequências de proteínas deveriam iniciar com o aminoácido Metionina, isto assegurou um grau maior de confiança no início correto dos modelos gênicos. Segundo, as sequências não deveriam possuir caracteres ilegais tais como: X, B, Z, U e "*", uma vez que estes são ambíguos ou não representam nenhum dos 20 aminoácidos. Terceiro, as sequências protéicas deveriam possui 100 ou mais aminoácidos.

4.3 – Predição dos Pares de Interação de Proteínas

Com o objetivo de predizer os pares de interação de proteínas para os três organismos alvos deste trabalho, o método conhecido como *Interolog Mapping* foi aplicado. Para a utilização desta abordagem, quatro bases de dados públicas que descrevem interações de proteínas e de domínios protéicos foram empregadas, são elas: Domine (http://domine.utdallas.edu/cgi-bin/Domine) (Raghavachari, Tasneem *et al.*, 2008), PSI-Base (http://psibase.kobic.re.kr/) (Gong, Yoon *et al.*, 2005), IntAct (http://www.ebi.ac.uk/intact/) (Aranda, Achuthan *et al.*, 2010), e String (http://string-db.org/) (Von Mering, Jensen *et al.*, 2005).

O primeiro passo para a execução desta etapa foi carregar em nosso servidores todas as interações descritas nestas bases além das sequências de proteína presentes nas mesmas. Posteriormente, as sequências dos proteomas preditos de *L. braziliensis*, *L. infantum* e *L. major* foram comparadas, através de alinhamento de sequência, contra as sequências de aminoácidos das proteínas provenientes das quatro bases de dados, o inverso também foi realizado. Estas comparações foram feitas utilizando o programa *blastp* do pacote Blastall (Sf, W *et al.*, 1990) para as bases PSI-Base, IntAct e String. Os programas do pacote Blastall possuem como estratégia de alinhamento de sequência o alinhamento local, e utilizam o algoritmo BLAST ("*Basic Local Alignment Search Tool*").

Já para a base de dados Domine, as comparações de sequências foram realizadas através do programa *hmmpfam*, que compara sequências protéicas contra Modelos Ocultos de Markov (HMM – "*Hidden Markov Models*"), e através do programa *hmmsearch*, que compara HMM contra sequências protéicas. Estes dois

programas são parte do pacote de programas conhecido como HMMER (Eddy, 1998), utilizado para lidar com HMM. Os HMMs são construídos a partir das sequências de proteínas de uma família protéica. Eles são constituídos de um perfil que incluem as probabilidades posição-específica de variação de aminoácido, bem como inserções e deleções. Isto pode indicar posições conservadas (importante para a família de proteínas), e não conservadas as quais são variáveis entre os membros da família. Por fim, foi necessário a utilização destes programas para a base de dados Domine porque esta utiliza os HMMs presentes na base de dados PFAM (Finn, Mistry *et al.*, 2010) para descrever suas proteínas.

Assim, a proteína *"X*" de uma base de dados é considerada homóloga a proteína "*A*" de um dos organismo de interesse se e somente se a proteína "*X*" for o melhor resultado de alinhamento para a proteína "*A*", e a proteína "*A*" for o melhor resultado de alinhamento para a proteína "*X*". Isto é chamado de Melhor Hit Bidirecional (BBH – "*Best Bidirectional Hit*"). Para cada BBH, diversas medidas foram extraídas. Quando um BBH era resultado do *blastp,* a identidade mínima, a similaridade mínima e a pontuação mínina de alinhamento (*alignScore*) entre as duas sequências foram extraídas. Além disso, a cobertura do alinhamento também foi utilizada. Contudo, quando o BBH era resultado dos programas do pacote HMMER, apenas a pontuação mínima de alinhamento foi extraída. Em suma, as seguinte fórmulas foram aplicadas:

$$identidade(AX) = (\min\{\max_{i,\ldots,k} identidade((A \leftarrow X)_i), \max_{j,\ldots,l} identidade((A \rightarrow X)_j)\})$$

$$similaridade(AX) = (\min\{\max_{i,\ldots,k} similaridade((A \leftarrow X)_i), \max_{j,\ldots,l} similaridade((A \rightarrow X)_j)\})$$

$$cobertura(AX) = (\min\{\max_{i,\ldots,k} cobertura((A \leftarrow X)_i), \max_{j,\ldots,l} cobertura((A \rightarrow X)_j)\})$$

$$alignScore(AX) = (\min\{\max_{i,\ldots,k} score((A \leftarrow X)_i), \max_{j,\ldots,l} score((A \rightarrow X)_j)\})$$

Aqui, "*A*" representa uma proteína de um dos organismos alvos, "*X*" representa uma proteína de uma das bases da dados, *k* é o número de resultados da comparação realizada utilizando "*X*" como pergunta e *l* é o número de resultados da comparação utilizando "*A*" como pergunta. Para cada comparação, os valores máximos de cada medidas são extraídos. Posteriormente, apenas os valores mínimos das duas comparações são utilizados para as futuras análises. Além disso, estas medidas foram calculadas para cada base de dados se e somente se o valor de *e-value* (chance do alinhamento com uma determinada pontuação e base de dados ocorrer ao acaso) do alinhamento para cada comparação fosse menor ou igual a 10^{-85} para o String e IntAct, a 10^{-45} para o Domine e a 10^{-10} para o PSI-Base.

Após esta etapa, as interações presentes nas bases de dados foram então mapeadas nos três proteomas. Para isso, primeiramente existia a informação de que "*X*" e "*Y*", que são proteínas de uma das bases de dados, interagem. Segundo, existia também a informação de que "*X*" era o BBH de "*A*", que é uma proteína de um dos três organismos de estudo, e "*Y*" era o BBH de "*B*", que é também uma proteína do mesmo proteoma de "*A*". Portanto, assume-se que "*A*" e "*B*" interagem (Figura 8).

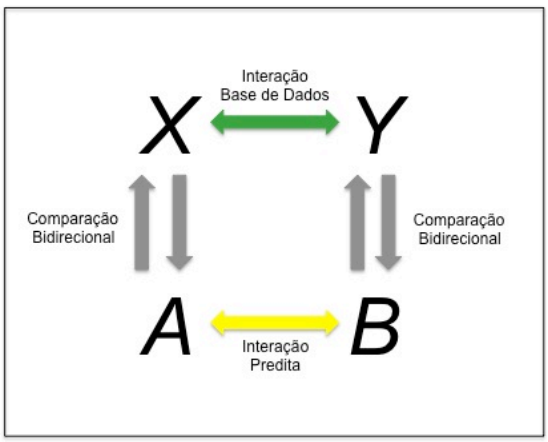


Figura 8 – Esquema ilustrando o racional utilizado na metodologia empregada para predição de interações entre proteínas (*Interolog Mapping*)

Em geral, cada uma das bases de dados utilizada neste estudo possuía uma pontuação de confiança para cada interação presente. Logo, estas pontuações foram utilizadas para compor a pontuação combinada final das interações aqui preditas. A única base de dados que não possuía esta pontuação para as suas interações foi PSI-Base. No final, para cada uma das bases de dados, a pontuação (*score*) das predições realizadas foi calculada como segue:

$$score_{STRING} = \left(\frac{\left(\frac{similaridade(AX) + similaridade(BY)}{2} \right) + \left(\frac{identidade(AX) + identidade(BY)}{2} \right) + \left(\frac{cobertura(AX) + cobertura(BY)}{2} \right)}{3} \right) \times scoreString(X,Y)$$

$$score_{IntAct} = \left(\frac{\left(\frac{similaridade(AX) + similaridade(BY)}{2} \right) + \left(\frac{identidade(AX) + identidade(BY)}{2} \right) + \left(\frac{cobertura(AX) + cobertura(BY)}{2} \right)}{3} \right) \times scoreIntAct(X,Y)$$

$$score_{Domine} = \sqrt{alignScore(AX) \times alignScore(BY)} \times scoreDomine(X,Y)$$

$$score_{PSIBase} = \left(\frac{\left(\frac{(similaridade(AX) + similaridade(BY))}{2} + \frac{(identidade(AX) + identidade(BY))}{2} + \frac{(cobertura(AX) + cobertura(BY))}{2} \right)}{3} \right) \times \sqrt{alignScore(AX) \times alignScore(BY}$$

4.4 – Cálculo de *Escore* de Confiança para Interações Protéicas Preditas

Com o objetivo de atribuir uma pontuação de confiança (*Escore* de Confiança) para as interações de proteínas preditas, um racional já descrito na literatura (Von Mering, Jensen *et al.*, 2005; Kim, Park *et al.*, 2008) foi adotado, e assim foi construído um *Escore* Combinado de Interação para metodologia aqui aplicada.

Este *Escore* Combinado de Interação, chamado aqui de *Escore* de Confiança, levou em consideração os *Escores* das predições calculados para as quatro bases de dados utilizadas, Domine, PSI-Base, IntAct e String, detalhados na seção anterior. Por fim, ele foi calculado conforme a fórmula abaixo:

$$score_{confiança(AB)} = 1 - \prod_{i \in E} (1 - S_i),$$

onde *score_confiança(AB)* é o *Escore* de Confiança de interação entre as proteínas "*A*" e "*B*" de um dos proteomas alvos de estudo, *E* representa os métodos que foram utilizados para predizer a interação, e S_i é o *Escore* do método *i* normalizado pelo maior valor de *Escore* encontrado para este método.

4.5 – Análise das Redes Preditas Frente a Modelos de Redes Descritos

A próxima etapa para avaliação das redes de proteínas aqui preditas foi a verificação de algumas características peculiares às redes biológicas. Segundo trabalhos da literatura (Ravasz, Somera *et al.*, 2002; Barabasi e Oltvai, 2004), a distribuição do número de interações e da média do índice de clusterização (agrupamento) dos nós de uma rede biológica seguem uma lei de potência:

Número de interações: $P(k) \sim k^{-\gamma}$

Índice de clusterização: $C(k) \sim k^{-1}$

Para a primeira distribuição, *P(k)* é a probabilidade de selecionar um nó com *k* número de interações, enquanto a segunda distribuição, *C(k)* é o valor médio do índice de clusterização na rede para nós com *k* número de interações. O valor γ

representa o expoente da distribuição. Assim, se as redes preditas possuem estas duas distribuições seguindo uma lei de potência, elas podem ser consideradas redes hierárquicas (Barabasi e Oltvai, 2004).

Portanto, utilizando estas duas distribuições como parâmetros, e utilizando o programa Cytoscape versão 2.8.3 (Shannon, Markiel *et al.*, 2003; Smoot, Ono *et al.*, 2011) juntamente com o *plug-in* NetworkAnalyzer versão 2.7 (Assenov, Ramirez *et al.*, 2008), as redes foram avaliadas.

Além disso, ainda utilizando o programa Cytoscape, porém com outro *plug-in* chamado Random Network versão 1.0 (http://sites.google.com/site/randomnetworkplugin/), as redes tiveram o índices de Coeficiente de Clusterização e a média do Caminho Mais Curto ("Shortest Path") comparados contra os mesmo índices, contudo gerados a partir de 1000 redes aleatórias criadas com base nas redes aqui preditas. Esta comparação foi realizada utilizando o teste-*z*.

4.6 – Anotação Funcional *Gene Ontology* (GO)

Para realizar atribuição de anotação funcional, o vocabulário de classificação definido pelo consórcio *Gene Ontology* (Ashburner, Ball *et al.*, 2000) (http://www.geneontology.org/) foi empregado.

Esta ontologia é composta de três domínios, são eles: Componente Celular ("*Cellular Component*"), que envolve descrições sobre partes de uma célula ou de seu ambiente extracelular; Função Molecular ("*Molecular Function*"), que é caracterizado por possuir termos que descrevem atividades elementares de um produto gênico ao nível molecular, tais como ligação ou catálise; e Processo Biológico ("*Biological*

Process"), que possui descrições de operações ou conjunto de eventos moleculares com um início e um fim definido, pertinentes ao funcionamento de unidades vivas integradas: células, tecidos, órgãos e organismos.

O esquema de anotação do GO para os organismos alvos deste trabalho teve sua origem da bases de dados TriTrypDB versão 4.1. Nela estavam disponíveis para cada uma das três ontologias do GO dois tipos de evidência de anotação. Uma era chamada de "anotação" e a outra de "predição". Portanto, com o objetivo de garantir um maior grau de confiança em relação à anotação funcional, quando possível os termos com a evidência de "anotação" foram utilizados para análises posteriores.

4.7 – Predição de Módulos Funcionais

Durante esta etapa do trabalho, o objetivo foi identificar os módulos (complexos ou grupos) funcionais que eram conservados entre as redes de interações de proteínas preditas para as três espécies de *Leishmania* alvos do estudo. Módulos funcionais podem ser entendidos como um grupo de proteínas funcionalmente ou fisicamente ligadas que trabalham juntas para alcançar ou realizar uma determinada função (Hartwell, Hopfield *et al.*, 1999). Além disso, de acordo com Ravasz e colaboradores (Ravasz, Somera *et al.*, 2002), redes de interação de proteína possuem arquitetura modular (hierárquica).

Assim, para executar a predição dos módulos funcionais conservados nas redes modeladas anteriormente, o programa networkBlast (Sharan, Suthram *et al.*, 2005) foi escolhido. Este programa possui um algoritmo que combina as redes de interação juntamente com informação relativa a similaridade de sequência com o objetivo de produzir uma rede que representa o alinhamento das redes de interação

que são fornecidas como entrada para o programa. Cada nó deste grafo define um grupo de proteínas similares enquanto ligações entre nós deste grafo definem prováveis complexos proteicos que são evolutivamente conservados nas redes que foram fornecidas inicialmente.

Mais detalhadamente, o programa networkBlast assume que duas condições devem ser preenchidas completamente para predição de módulos conservados entre redes de interação, são elas: 1) o conjunto de interações que compõe um módulo dentro de cada rede deve possuir uma estrutura semelhante; 2) sendo k o número de redes a serem analisadas, deve existir um correspondência entre os conjuntos de proteínas que compõe o módulo em diferentes redes, logo grupos de k proteínas, uma de cada rede, induzidas por esta correspondência, representem k proteínas com sequências similares.

Além disso, após a redes de interações serem alinhadas, um grupo de k proteínas distintas, uma de cada rede, compõe um nó do grafo alinhado se o grupo não puder ser dividido em duas partes com nenhuma similaridade de sequência entre elas. Para $k = 2,3$, esta condição traduz a necessidade de que cada proteína em um grupo que compõe um nó grafo alinhado tenha no mínimo uma outra proteína com sequência similar neste mesmo grupo. Assim, dois nós $(p_1,...,p_k)$ e $(q_1,...,q_k)$ no grafo alinhado estão conectados por uma ligação se e somente se uma das seguintes condições for respeitada em relação ao par de interação (p_i, q_i): 1) um par de proteínas diretamente interage e todos outros pares incluem proteínas com distância de no máximo dois em relação as redes de interação; 2) todos os pares de interação estão exatamente a uma distância de dois nas redes de interação; ou 3) no mínimo $max\{2, k - 1\}$ pares de proteínas diretamente interagem.

Portanto, um módulo no grafo alinhado corresponde a um módulo conservado. Assim, para cada rede *S*, o conjunto de proteínas incluídas nos nós do módulo define o módulo que é induzido em *S*.

Posteriormente à etapa de predição dos módulos, foi realizada a caracterização funcional dos mesmos. Para execução desta tarefa, um esquema de anotação funcional era necessário, e como descrito na seção anterior o vocabulário controlado do GO foi utilizado.

Com o esquema de anotação em mãos e considerando a ontologia somente em relação a Processo Biológico, foi realizada uma analise de enriquecimento de termos GO para cada módulo predito. Para isso, o módulo de programação em linguagem Perl chamado GO::TermFinder (Boyle, Weng *et al.*, 2004) foi utilizado. Este módulo de programação possui um conjunto de ferramentas implementadas para lidar com a informações provenientes da base do GO. Além disso, a significância estatística do enriquecimento de um determinado termo para um determinado complexo funcional (*P-value*) é calculado pelo GO::TermFinder através de uma distribuição hipergeométrica:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$

Aqui, *N* é igual ao número total de proteínas em uma das redes preditas que receberam no mínimo um termo GO, *M* representa o número de proteínas dentro do conjunto *N* que estão anotadas (diretamente ou indiretamente) a qualquer termo GO de interesse, *n* é o tamanho da lista de proteínas de interesse, neste caso, é o número

de proteínas no módulo funcional de interesse. Finalmente, *k* é o número de proteínas dentro do conjunto *n* que estão anotadas (diretamente ou indiretamente) com o termo GO de interesse. Além disso, como esta análise envolve teste de múltiplas hipóteses, um método de correção para o *P-value* teve que ser aplicado. O GO::TermFinder aplica o método de correção de Bonferroni.

4.8 – Análise Topológica

As métricas utilizadas com o objetivo de extrair informação biológica das redes de interações preditas foram calculadas utilizando o *plug-in* CytoHubba versão 1.6 (Lin, Chin *et al.*, 2008) dentro do programa Cytoscape versão 2.8.3. Neste estudo, o índices topológicos "*Degree"* e MCC ("*Maximal Centrality Clique*") foram utilizados.

O *Degree* é o índice topológico que está diretamente ligado à contagem do número de interações que uma proteína faz dentro da rede de interação.

Por outro lado, de acordo com os desenvolvedores do CytoHubba (http://hub.iis.sinica.edu.tw/cytoHubba/supplementary/index.htm), o índice topológico MCC demonstrou ter a maior sobreposição com proteínas essenciais da rede de interação de proteínas de *Saccharomyces cerevisae*. A sobreposição relatada foi de 80% para as 10 primeiras proteínas ranqueadas por este índice, e 70% para as 100 primeiras proteínas também ranqueadas pelo MCC. Portanto, devido o alto desempenho em ordenar as proteínas de um rede em relação à essencialidade para o organismo, o MCC foi utilizado para ranquear as proteínas das três redes de *Leishmania* aqui modeladas.

Além disso, a variabilidade das proteínas mais bem ranqueadas também foi avaliada com base em seus grupos de proteínas ortólogas que estão definidos na base de dados do TriTrypDB. Nesta base, as proteínas de Kinetoplastida são agrupadas em grupos de proteínas ortólogas através da informação existente na base de dados OrthoMCL (http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi) (Li, Stoeckert *et al.*, 2003). Assim, para cada grupo de ortólogos associado as proteínas mais bem ranqueadas, foi realizado um alinhamento múltiplo através do programa MAFFT (Katoh, Misawa *et al.*, 2002), e a identidade média foi avaliada através do programa *alistat* do pacote HMMER.

4.9 – Análise Evolutiva

Existem relatos na literatura de que as proteínas com alto valor do índice topológico *Degree* provavelmente são mais conservadas e mais ancestrais.

Portanto, com o objetivo de avaliar está correlação e assim também obter mais uma evidência do caráter biológico das redes aqui preditas, o grau de diversidade nucleotídica (ou índice de diversidade nucleotídica - π) (Nei e Li, 1979) dos genes que codificam as proteínas nas redes preditas foi comparado com seus respectivos *Degrees*.

O π pode ser entendido como um método para medir a quantidade de variação nucleotídica em uma população. Para isto, ele determina a proporção de nucleotídeos diferentes entre pares de sequências, e estas diferenças recebem pesos diferentes de acordo com as frequências das sequências. Assim, para um população o π poder ser calculado da seguinte maneira:

$$\pi = \sum_{ij} p_i p_j \pi_{ij}$$

Onde p_i e p_j representam a frequência da sequência *i* e *j* e π_{ij} é a proporção de nucleotídeos que diferem quando as sequências *i* e *j* são comparadas.

Para determinar o π, o programa Variscan (http://www.ub.edu/softevol/variscan/) (Vilella, Blanco-Garcia *et al.*, 2005) foi utilizado. Este programa possui um conjunto de ferramentas para análise de polimorfismo de DNA em grande quantidade de dados.

Os dados que foram utilizados como entrada para o Variscan foram os conjuntos de sequências de DNA de genes ortólogos alinhados pelo MAFFT. Estes conjuntos de ortólogos foram definidos na base de dados TriTrypDB e possuíam genes dos genomas dos seguintes organismos: *L. braziliensis*, *L. infantum*, *L. major*, *Leishmania mexicana*, *Trypanosoma brucei, Trypanosoma cruzi* e *Trypanosoma vivax*. Além disso, os grupos de ortólogos foram separados por faixas de *Degree* para a construção do gráfico que representa o resultado desta análise. As faixa de *Degree* utilizadas foram: de 2 a 10, de 11 a 20, de 21 a 30, de 31 a 40, de 41 a 50 e maior que 50.

4.10 – Análise das Proteínas Hipotéticas

Como já foi mencionado na introdução, de uma maneira geral, as proteínas hipotéticas são definidas como proteínas preditas computacionalmente a partir de sequências de nucleotídeos, geralmente os genomas, porém não existem evidências

experimentais de que elas existam. Além disso, não existem proteínas com função determinada que sejam similares às proteínas hipotéticas.

Além disso, os genomas de Tripanosomatídeos são conhecidos entre outras características por terem uma grande quantidade de proteínas hipotéticas, quantidade esta que chega aproximadamente a 60% do proteoma predito dos organismos deste grupo (El-Sayed, Myler, Bartholomeu *et al.*, 2005; Ivens, Peacock *et al.*, 2005; Peacock, Seeger *et al.*, 2007; Raymond, Boisvert *et al.*, 2012). E obviamente a caracterização destas proteínas é de suma importância uma vez que elas podem estar envolvidas em processos celulares essenciais, e processos relacionados à interação parasito-hospedeiro.

Portanto, devido a clara importância destas proteínas para o entendimento da biologia das *Leishmanias*, a possibilidade de utilizar as redes de interação de proteínas modeladas nesta estudo para inferir uma função às proteínas hipotéticas dos organismos alvo de estudo foi explorada.

Para execução desta etapa, o programa *FS-Weight* (Chua, Sung *et al.*, 2006) foi utilizado. Este programa utiliza uma abordagem que é baseada em associação funcional direta e indireta utilizando as redes de interação de proteínas como principal fonte de informação. Em uma rede de interação, vizinhos diretos ou indiretos (existem outras proteínas intermediárias) de uma proteína podem compartilhar algumas propriedades físicas e bioquímicas que permitem que eles se linguem a esta proteína. Assim, este método tem como vantagem o fato de não somente utilizar vizinhos diretos, o que limitaria a predição de função para as proteínas que tem pelo menos um único vizinho com anotação conhecida.

Mais detalhadamente, o *FS-Weight* realiza um cálculo de similaridade funcional entre duas proteínas, não sendo elas necessariamente vizinhas diretas na

rede de proteínas. Este cálculo é baseado no contexto topológico de ambas as proteínas e a confiança das interações que elas fazem. Isto é feito para reduzir os efeitos de interações falsas. Portanto, quanto mais proteínas em comum existirem interagindo com duas outras proteínas, maiores são as chances de que as duas proteínas compartilhem alguma função biológica.

Além disso, esta abordagem necessita de um esquema de anotação que já tenha sido utilizado para proteínas com funções conhecidas. Desta maneira, a ontologia do GO foi utilizada.

4.11 – Avaliação dos Métodos de Predição de Epítopos

Do mesmo modo como foi descrito na seção 4.1 relativa à avaliação do método de predição de interações protéicas, para avaliarmos os métodos de predição de epítopos também foi necessário obter os conjuntos de dados que representassem o controle positivo e o controle negativo.

Portanto, foram carregados dois conjuntos de epítopos testados experimentalmente, um conjunto de epítopos com afinidade para receptores de célula B e outro de epítopos com afinidade para receptores de células T CD8+ (MHC classe I). Estes conjuntos de epítopos foram derivados de proteínas de parasitos e foram extraídos da base de dados IEDB (http://www.iedb.org/) (Vita, Zarebski *et al.*, 2010). Os seguinte critérios foram empregados para a seleção dos epítopos: a) as proteínas as quais os epítopos pertenciam deveriam ser de protozoários parasitos; b) os epítopos de células T CD8+ selecionados deveriam ter afinidade pelo MHC I de camundongo (*Mus musculus*) ou do homem (*Homo sapiens*). Além disso, os epítopos selecionados

eram epítopos mínimos, experimentalmente validados como imunogênicos (controle positivo) ou não-imunogênicos (controle negativo).

No entanto, existiam sobreposições relativas à localização na sequência de aminoácido entre diversos epítopos de uma mesma proteína. Deste modo, para obter um conjunto de dados validado experimentalmente sem redundância para cada proteína, uma estratégia que foi denominada de "região consenso validada" foi empregada. Esta abordagem consiste em agrupar os epítopos em uma única região consenso que foi chamada de região consenso experimentalmente validada.

No final, o conjunto de dados experimentais para epítopos de célula B possuía 312 proteínas e 866 regiões consensos experimentalmente validadas incluindo controle positivo e controle negativo. Para os epítopos de MHC classe I, o conjunto de dados foi composto de 81 proteínas com 224 regiões consensos experimentalmente validadas incluindo também controle positivo e controle negativo.

Posteriormente à construção destes conjuntos de dados, os mesmos foram utilizados como entrada para o programa *formatdb* do pacote Blastall. Este programa formata as sequências de modo que possam ser alinhadas pelo algoritmo BLAST.

Outro passo importante desta etapa do trabalho foi a seleção dos algoritmos de predição de epítopos para serem avaliados. Para esta seleção, a possibilidade da implementação do algoritmo ser instalada localmente, e a confiança da predição reportada na literatura foram levados em consideração. Logo, os algoritmos avaliados foram BepiPred (Larsen, Lund *et al.*, 2006), BCPreds (Chen, Liu *et al.*, 2007; El-Manzalawy, Dobbs *et al.*, 2008), que inclui duas metodologias, AAP12 e BCPred12 para predição de epítopos de célula B. Para epítopos de MHC classe I, os seguinte preditores foram avaliados: NetCTL (Larsen, Lundegaard *et al.*, 2007) e NetMHC (Nielsen, Lundegaard *et al.*, 2004).

Além disso, os preditores de epítopos para célula T CD8+ foram testados para o alelo A2. Outros alelos HLA estão presentes na base de dados IEDB, porém para as proteínas de protozoários, eles estão sub-representados. Além disso, o alelo HLA-A2 está incluído no grupo que é expresso em 88% da população, o que enfatiza sua relevância.

Finalmente, depois da escolha dos preditores a serem avaliados, os mesmos foram então executados nas proteínas presentes nos controles. Após a execução, o resultado da predição foi então comparado com as regiões consenso experimentalmente validadas. Isto possibilitou o cálculo das taxas de TP, FP, TN e FN.

Mais detalhadamente, para a classificação dos epítopos preditos como TP ou FP, foi utilizado o resultado do programa *blastp* alinhando o conjunto de dados controle e a predição dos algoritmos. Os seguintes parâmetros foram utilizados para decidir se uma predição de epítopos iria ser considerada como TP (Figura 9): 1) o alinhamento local entre a predição de um epítopo ("*query*") e uma região consenso imunogênica validada experimentalmente ("*subject*") deveria ter no mínimo 50% de cobertura caso fosse um predição para célula B e 87% de cobertura caso fosse para célula T CD8+. Estes cortes de cobertura foram estabelecidos com base no tamanho mínimo dos epítopos experimentais presentes no IEDB que foi de 6 e de 8 aminoácidos para célula B e T CD8+, respectivamente. Os programas NetCTL e NetMHC fazem predição de epítopos com 9 aminoácidos, logo 87% de cobertura garante um tamanho mínimo de alinhamento de 8 aminoácidos. Por outro lado, os preditores AAP12 e BCPred12 predizem epítopos de 12 aminoácidos, portanto 50% de cobertura garante o tamanho mínimo de alinhamento de 6 aminoácidos. Em relação ao BepiPred, ele faz predição de epítopos de tamanhos variados, assim

somente as predições com no mínimo 6 aminoácidos foram consideradas na análise, e para predições variando entre 6 a 11 aminoácidos, o mínimo de cobertura necessário no alinhamento variou para garantir um alinhamento de no mínimo 6 aminoácidos; 2) O alinhamento local entre uma predição de epítopos e a região consenso validada experimentalmente (imunogênica) deveria ter 100% de identidade. Este parâmetro foi utilizado para assegurar que a predição estava realmente retornando a mesma região validada experimentalmente; 3) Finalmente, com o objetivo de garantir que a predição e a região validada experimentalmente pertenciam à mesma proteína, o nome da *query* e do *subject* deveriam ser os mesmos.

As predições classificadas como FP foram identificadas utilizando o mesmo racional descrito acima, porém os alinhamentos foram realizados contra as regiões consenso não-imunogênicas validadas experimentalmente.

Epítopos preditos que não alinharam respeitando os parâmetros acima ou que alinharam contra os dados imunogênicos e não-imunogênicos simultaneamente não foram considerados para análises posteriores.

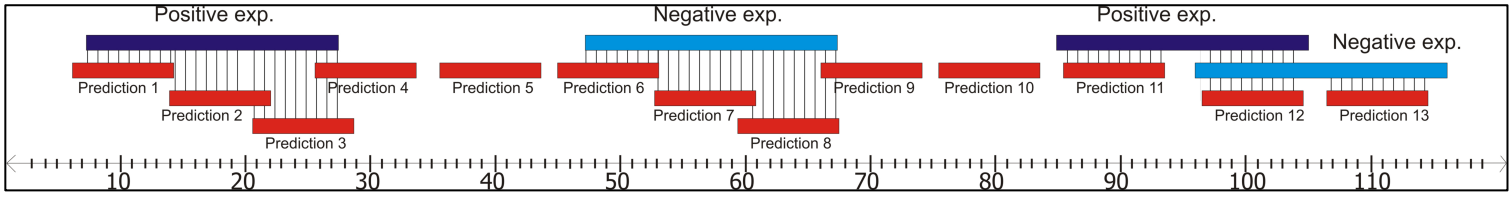


Figura 9 – Estratégia empregada para avaliar o desempenho dos preditores de epítopos. A barra com a escala representa a sequência teórica de uma proteína de 120 aminoácidos. Os retângulos em azul escuro representam um simples epítopo ou um consenso de epítopos sobrepostos que foram experimentalmente validados de acordo com a base IEDB (Positive Exp); os retângulos em azul claro representam uma única região não-imunogênica ou um consenso de regiões não-imunogênicas sobrepostas que foram validadas experimentalmente de acordo com a base IEDB (Negative Exp); os retângulos vermelhos representam epítopos preditos a partir dos algoritmos avaliados. Para predição de células B, os epítopos preditos foram considerados TP se eles alinhassem com no mínimo 50% de cobertura e 100% de identidade a uma região experimentalmente validade como positiva; Para predição de células T CD8+, os epítopos preditos foram considerados TP se eles alinhassem com no mínimo 87% de cobertura e 100% de identidade a uma região experimentalmente validade como positiva (Prediction 1, Prediction 2, Prediction 3 e Prediction 11). Para predicão de células B, os epítopos preditos foram considerados FP se

eles alinhassem com no mínimo 50% de cobertura e 100% de identidade a uma região validada experimentalmente como negativa; para predição de células T CD8+, os epítopos preditos foram considerados TP se eles alinhassem com no mínimo 87% de cobertura e 100% de identidade a uma região validada experimentalmente como negativa (Prediction 6, Prediction 7, Prediction 8 e Prediction 13). Predições não foram consideradas nas análises se elas não alinhassem respeitando os parâmetros citados acima (Prediction 4, Prediction 5, Prediction 9 e Prediction 10) ou se elas alinhassem simultaneamente em ambas regiões positivas e negativas experimentalmente validadas (Prediction 12).

Além das avaliações dos preditores acima citados, também foram avaliadas as combinações destas ferramentas. Desta maneira, o seguinte racional foi aplicado: 1) para uma dada proteína, as regiões experimentais foram indexadas e então, considerando a proteína (P) com três regiões experimentalmente validadas, estas seriam nomeadas como P1, P2 e P3; 2) se um dado algoritmo A prediz um epítopo que corresponde a P2 por exemplo e outro algoritmo B prediz um epítopo que também corresponde a P2 então as duas predições são consideradas como uma predição combinada; 3) se um algoritmo A prediz um epítopo que corresponde a P2 e um outro algoritmo B prediz um epítopo que corresponde a P1 ou P3 então as duas predições não são consideradas como predição combinada.

Finalmente, depois de todas as comparações para obtenção das predições consideradas TP e FP, e posterior derivação das predições FN e TN, foi possível empregar a mesma abordagem da seção 4.1 para avaliar os preditores. Assim, o cálculo da curva ROC e posteriormente do AUC foi realizado.

4.12 – Aplicação dos Métodos de Predição de Epítopos

Com o objetivo de explorar o potencial imunológico das proteínas mais bem ranqueadas conforme descrito na seção 4.8, as ferramentas de predição de epítopos

que tiveram os melhores desempenhos determinados na seção anterior foram aplicadas.

Para trabalhar com epítopos que potencialmente serão reconhecidos por células B, o programa BCPred12 foi utilizado. Já para os epítopos potencialmente reconhecidos por células T CD8+ e T CD4+, os programas NetCTL e NetMHCII foram empregados respectivamente.

Além disso, com o objetivo de avaliar se as proteínas mais bem ranqueadas, que são potenciais alvos para desenvolvimento de drogas e vacinas, possuíam alguma similaridade com proteínas de alguns hospedeiros mamíferos, os proteomas preditos de camundongo (*M. musculus*), homem (*H. sapiens*) e cachorro (*Canis lupus familiaris*) foram carregados a partir do repositório do NCBI ("*National Center for Biotechnology Information*" – www.ncbi.nlm.nih.gov) em 24 de Agosto de 2012 para os nossos servidores. Posteriormente as sequências de aminoácidos foram comparadas através do programa *blastp* do pacote Blastall.

4.13 – Integração de Dados

Para agregar e relacionar toda a informação produzida durante o desenvolvimento deste estudo, o Sistema Gerenciador de Banco de Dados Relacional (RDBMS – "*Relational Database Management System*") MySQL (http://www.mysql.com) foi utilizado. O uso de um sistema baseado em banco de dados relacional proporciona um modo eficiente de extrair correlações e resultados.

As operações de construção, gerenciamento e consulta do banco foram realizadas a partir da Interface Gráfica do Usuário (GUI – "*Graphical User Interface*") chamada MySQL Workbench (http://wb.mysql.com). Além disso, as

formatações de arquivos necessárias e a carga dos dados para o banco foram realizadas através de programas escritos em linguagem Perl utilizando os módulos de programação DBI e BioPerl. A Figura 10 exibi o modelo entidade-relacionamento (ER) criado para o banco de dados utilizado neste trabalho.

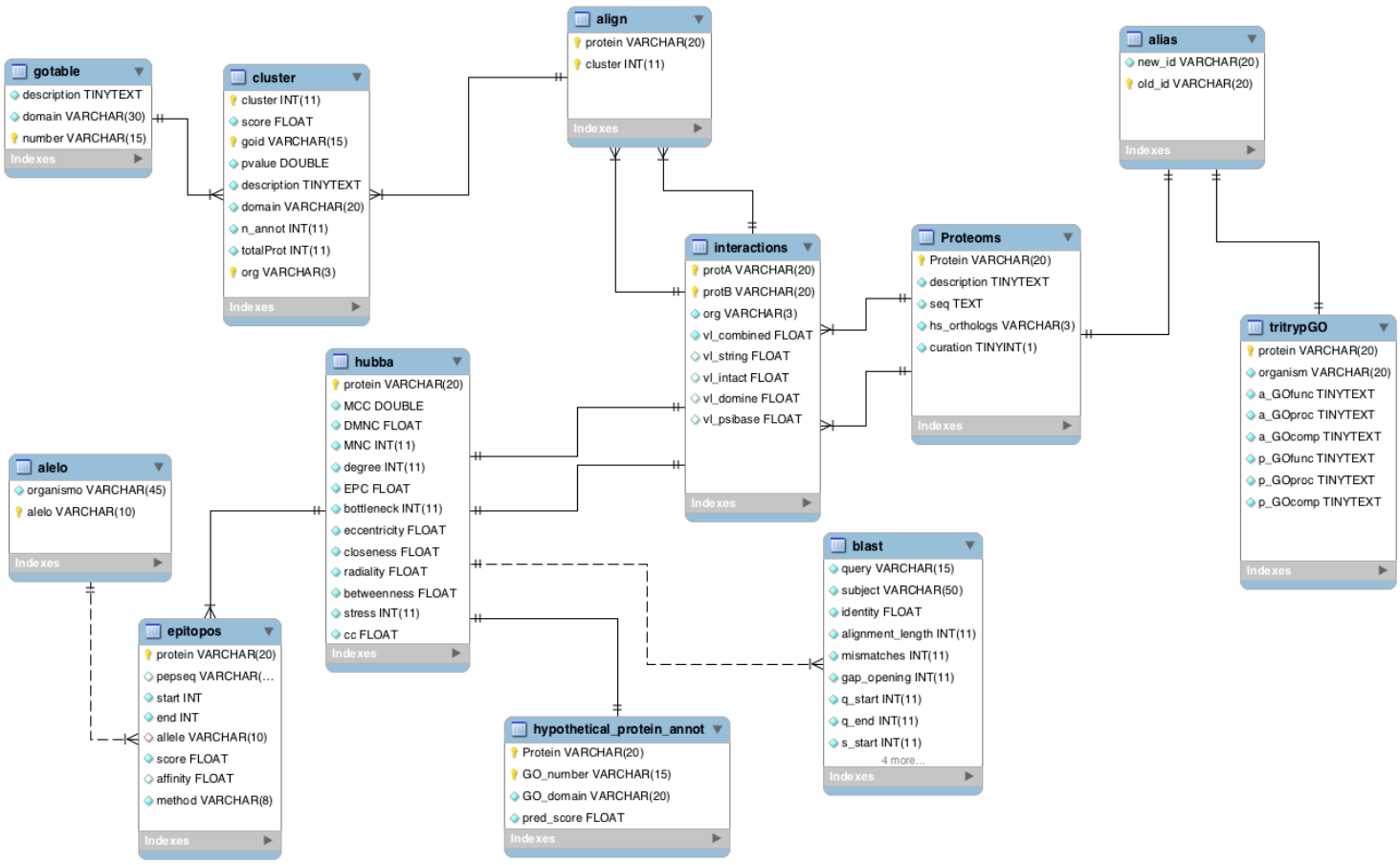


Figura 10 – Modelo entidade – relacionamento do banco de dados modelado paro a análise dos dados gerados neste projeto.

A principal tabela do banco de dados é chamada de *Proteoms*, e ela possui todas as proteínas já filtradas dos proteomas preditos de *L. braziliensis*, *L. infantum* e *L. major*. A tabela *Interactions* alberga os pares de interação preditos e os *Escores* das interações para todas as bases de dados utilizadas. A tabela *hubba* contem os valores dos índices topológicos calculados para as proteínas presentes nas redes aqui

modeladas. As tabelas *epítopos* e *alelo* estão relacionadas as análises de potencial imunológico das proteínas presentes na tabela *hubba*. Além disso, as tabelas *hypothetical_protein_annot* e *blast* albergam resultados da anotação das proteínas hipotéticas presentes nas redes de interação a partir de suas posições nas redes e os resultados das proteínas presentes nas redes quando comparadas às proteínas dos proteomas preditos de *M. musculus, H. sapiens* e *C. lupus familiaris*. As tabelas nomeadas de *cluster* e *align* estão vinculadas aos resultados das análises de predição de módulos funcionais conservados entre as três espécies de *Leishmania*. A tabela *tritrypGO* alberga as informações relativas às anotações das proteínas dos proteomas preditos das *Leishmanias* através da ontologia do GO presentes na base de dados TriTrypDB. A tabela *alias* possui a relação entre os identificadores atuais das proteínas das *Leishmanias* utilizado na base de dados TriTrypDB e os identificadores das proteínas relativos às versões dos genomas que foram utilizados para este estudo. Esta tabela foi carregada a partir de informações presentes no TriTrypDB. Por fim, a tabela *gotable* contem informações relativas ao termos anotadores da ontologia GO, está tabela foi gerada a partir dos arquivos disponíveis para carregamento no site do GO.

5 – Resultados

Os resultados aqui apresentados podem ser encontrados nos artigos publicados em revistas indexadas presentes no Anexo I deste trabalho.

5.1 – Avaliação de Desempenho dos Métodos de Predição

Os métodos utilizados na predição das interações de proteína e na predição de epítopos para receptores de célula B e moléculas de MHC classe I foram avaliados em relação aos seus desempenhos, e os resultados das avaliações estão a seguir.

5.1.1 – Método de Predição de Redes de Interação

Como descrito na seção de Materiais e Métodos, conjuntos de dados representando o controle positivo e o controle negativo foram construídos a partir da base de dados DIP, utilizando dados de interação de proteína de *E. coli* com o objetivo de avaliar a confiança e o desempenho da metodologia de predição de redes de interação de proteína, denominada "*Interolog Mapping*", aqui empregada.

Conjuntos de dados com uma alta qualidade foram obtidos, e estes eram compostos de 702 pares de interação proteicos, no caso do controle positivo. Já o controle negativo foi construído com uma razão de 5:1 em relação ao controle positivos, logo ele possuía 3.510 pares de interação de proteína. Estes dados foram então utilizados juntamente com a metodologia das curvas ROC para a avaliação do método de predição de redes de interação.

A acurácia da metodologia de redes de interação, medida pela área sob a curva ROC (AUC), pode ser visualizada na Tabela 2, e através do gráfico presente na Figura 11. O valor de AUC de 0,94 obtido para o *score_comb*, que é o método de pontuação desenvolvido neste trabalho, indica a robustez da abordagem aqui utilizada. No entanto, este resultado deve ser cuidadosamente considerado, uma vez que as bases de dados utilizadas nas predições possuem muitas interações protéicas de *E. coli*. Isto talvez possa conferir algum grau de viés à avaliação de desempenho da metodologia de "*Interolog Mapping*".

Tabela 2 – Avaliação de desempenho da abordagem utilizada para predição das redes de interação de proteína

Medida de Pontuação do *Interolog Mapping*	**Valor de AUC**
Método desenvolvido (*score_comb)*	0.94
Média Geométrica do *score* do alinhamento	0.74
Média Geométrica do *evalue* do alinhamento	0.57
Valor Máximo do *evalue*	0.55

Outros métodos previamente descritos na literatura (Yu, Luscombe *et al.*, 2004) de pontuação da interação a partir do "*Interolog Mapping*" foram também avaliados (Tabela 2 e Figura 11). É possível perceber a partir dos resultados que o método de pontuação desenvolvido neste trabalho teve um melhor desempenho em relação aos demais.

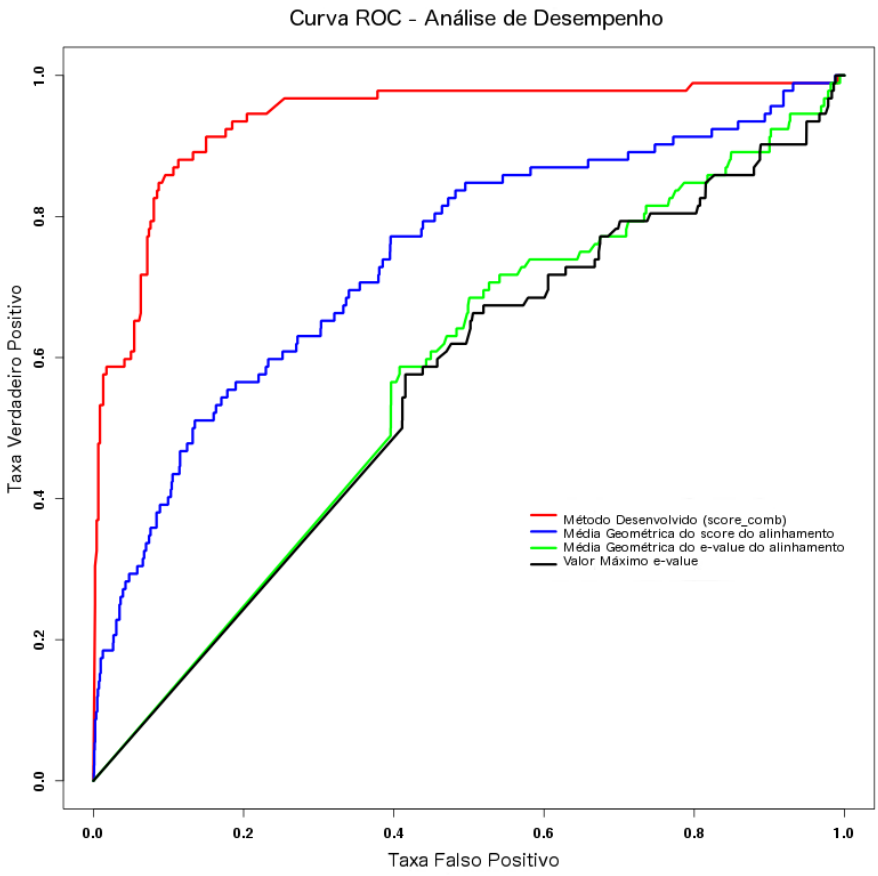


Figura 11 – Curvas ROC criadas a partir da análises de desempenho

5.1.2 – Métodos de Predição de Epítopos

Diversas abordagens para predição de peptídeos ou epítopos com afinidade de ligação para moléculas de MHC têm sido publicadas. Aqui, dois algoritmos publicados foram escolhidos para serem avaliados, NetCTL e NetMHC. A escolha deste algoritmos foi realizada levando em consideração o número de citações ISI dos trabalhos que descrevem os respectivos algoritmos e a disponibilidade para carregamento e implementação dos mesmos em nossos servidores.

Quando possível, com o objetivo de estabelecer parâmetros ideais para predição de epítopos em protozoários, os parâmetros dos algoritmos foram triados e avaliados para valores de AUC. Dentro desta estrutura, a pontuação de corte do

NetCTL variou de 0,50 a 0,90. O total de predições avaliadas para os dois algoritmos foi de 3.906 predições. Com o objetivo de avaliar o desempenho então destes algoritmos, epítopos preditos foram alinhados contra o conjunto de dados consenso validados experimentalmente para epítopos que se ligam em moléculas de MHC classe I. Além disso, uma análise de desempenho da combinação dos resultados utilizando a melhor pontuação de corte para cada algoritmo foi realizada (Tabela 3).

Tabela 3 – Avaliação de desempenho dos algoritmos de predição de epítopos

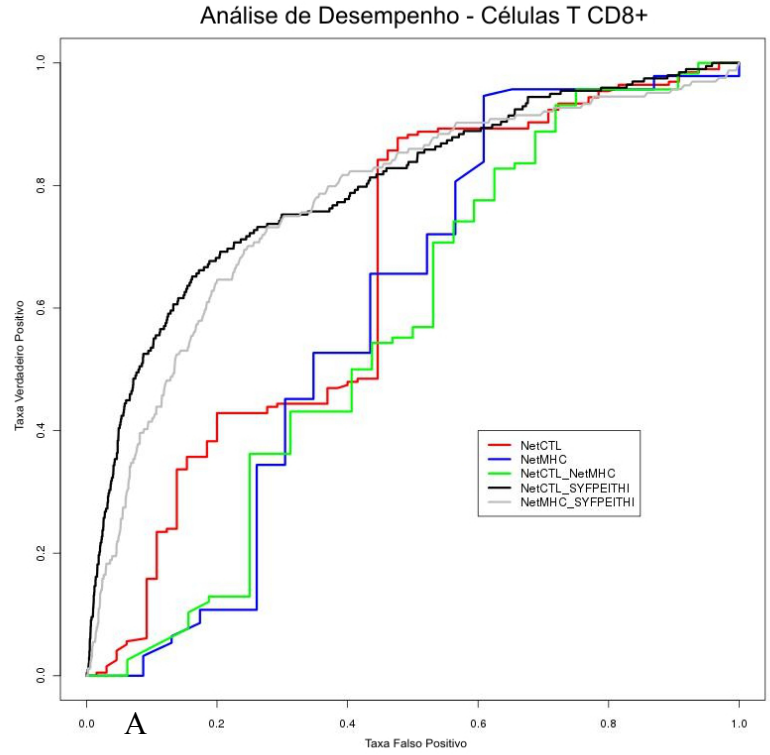
Algoritmos	Valores de AUC
MHC classe I	
NetCTL	0,66
NetMHC	0,60
NetCTL e NetMHC	0,64
Receptores Célula B	
AAP12	0,52
BCPred12	0,62
BepiPred	0,53
AAP12 e BCPred12	0,77
AAP12 e BepiPred	0,49
BCPred12 e BepiPred	0,58
AAP12, BCPred12 e BepiPred	0,57

O valor de AUC obtido para o NetCTL com uma pontuação de corte de 0,50 foi de 0,66, e para o NetMHC foi de 0,60. Este algoritmo não permite mudança na pontuação corte. Por outro lado, a análise de desempenho dos dois algoritmos combinados obteve um valor de AUC igual a 0,64 (Tabela 3 e Figura 12A).

Utilizando do mesmo racional descrito acima para algoritmos de predição de epítopos com afinidade de ligação para MHC classe I, três algoritmos atualmente disponíveis para predição de epítopos com afinidade de ligação em receptores de célula B (imunoglobulinas de superfície) foram selecionados. São eles: BepiPred, BCPred12 e AAP12. Quando possível, os parâmetros dos algoritmos também foram triados utilizando como referência os valores de AUC, para assim estabelecer os melhores valores para predição de epítopos em protozoários. Deste modo, o valor de

pontuação de corte para BepiPred variou de 0,15 a 0,90 e de 0,50 a 0,90 para AAP12 e BCPred12. O total de predições avaliadas para os três algoritmos foi de 187.187 predições. Portanto, com o objetivo de avaliar o desempenho dos algoritmos, os epítopos preditos foram alinhados contra conjunto de dados de regiões consenso experimentalmente validadas para epítopos de célula B. A análise de desempenho para os resultados combinados dos algoritmos também foi realizada utilizando novamente a melhor pontuação de corte para cada metodologia (Tabela 3 e Figura 12B).

Os valores de AUC obtidos foram iguais a 0,53 para BepiPred utilizando uma pontuação de corte de 0,40, a 0,52 para AAP12 com uma pontuação de corte de 0,80, e a 0,62 para BCPred12 utilizando um corte de 0,90 (Tabela 3). Em relação à analise dos resultados combinados, os seguintes resultados foram encontrados: 0,77 para AAP12 e BCPred12; 0,49 para AAP12 e BepiPred; 0,58 para BCPred12 e BepiPred, e 0,57 para AAP12, BCPred12 e Bepipred.



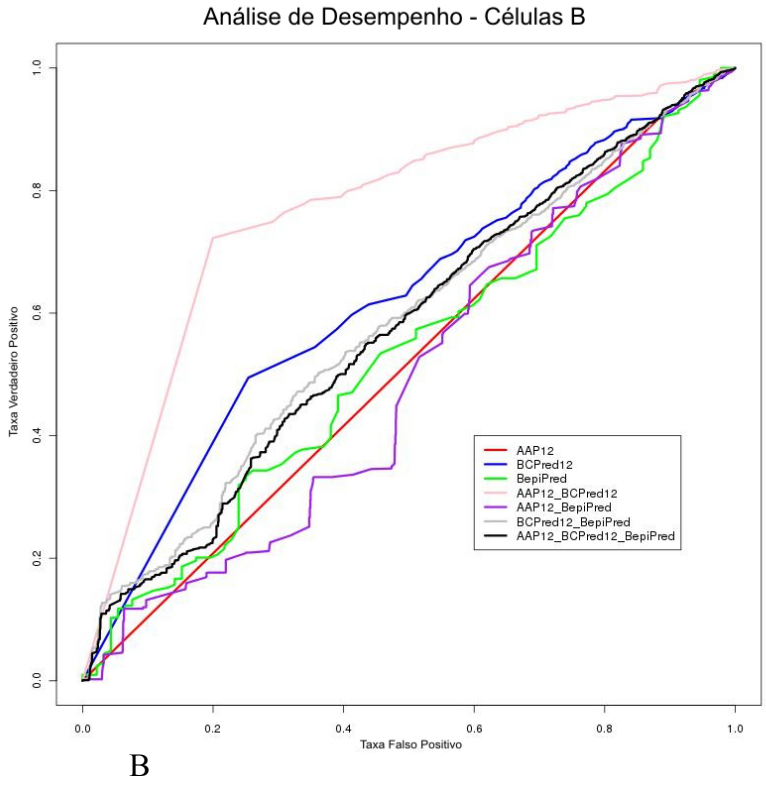


Figura 12 – Análise de desempenho dos algoritmos de predição de epítopos através da construção de curvas ROC.

5.2 – Filtragem dos Dados

De acordo com que foi dito na seção de metodologia, uma etapa de filtragem nos três proteomas preditos alvos do estudo foi realizada com o objetivo de selecionar sequências que tivessem sido anotadas corretamente.

Uma pequena porcentagem de proteínas foi excluída das análises uma vez que suas sequência apresentavam possíveis erros. Assim, os proteomas preditos de *L. braziliensis*, *L. infantum* e *L. major* perderam 4,33%, 4,78% e 2,95% de proteínas, respectivamente (Tabela 4).

Tabela 4 – Número de proteínas nos proteomas preditos dos organismos alvos antes e após a processo de filtragem

Organismo	Número Inicial Total de proteínas	Número Final Total de proteínas após filtragem	Número relativo de proteínas perdidas (%)
L. braziliensis	8310	7950	4,33
L. major	8408	8160	2,95
L. infantum	8216	7823	4,78

5.3 – Predição dos Pares de Interação de Proteínas

Subsequentemente ao processo de filtragem, os três proteomas preditos foram utilizados para a predição das redes de interação de proteína. Estas predições foram realizadas baseadas em diferentes base de dados tais como Domine, PSI-Base, IntAct e String. Utilizando estas bases, um esquema de pontuação chamado *score_comb* foi proposto e calculado para as interações preditas, e este varia de 0 a 1. Após este cálculo, foi possível demonstrar que as três redes preditas se encaixavam bem no modelo livre de escala (Tabela 5).

Tabela 5 – Características gerais das três redes de interação de proteínas preditas

Organismo	Número de Nós (Proteínas)	Número de Interações	Modelo Livre de Escala	
			Correlação	R^2*
L. braziliensis	1818	39420	0,941	0,816
L. major	1947	43531	0,925	0,815
L. infantum	1959	45235	0,940	0,829

*R^2 - Coeficiente de determinação: varia entre o intervalo de 0 a 1. Se seu valor é próximo de 1, isto quer dizer os dados podem ser explicados fortemente pelo modelo proposto. Se o valor está próximo de 0, então o modelo proposto fracamente explica os dados apresentados.

As redes preditas foram então comparadas contra 1.000 redes aleatórias. A comparação ocorreu entre os índices denominados Coeficiente de Agrupamento e Média do Caminho Mais Curto (Tabela 6). Os valores dos Coeficientes de Agrupamento das redes preditas foram muito maiores do que os das redes aleatórias adicionando uma camada extra de credibilidade para as redes preditas, além de ser um indicativo para o modelo de rede hierárquica.

Tabela 6 – Resultados das comparações do Coeficiente de Agrupamento e da Média do Caminho Mais Curto das três redes de interação contra os mesmos índices de 1.000 redes aleatórias

Leishmania braziliensis			
Índice	Rede Predita	Rede Aleatória	P-value
Coeficiente de Agrupamento	0.433	0.159±0,003	$p<0.05$
Média do Caminho Mais Curto	2.877	2.579±0,004	$p<0.05$
Leishmania major			
Índice	Rede Predita	Rede Aleatória	P-value
Coeficiente de Agrupamento	0.430	0.157±0.003	$p<0.05$
Média do Caminho Mais Curto	2.914	2.584±0.004	$p<0.05$
Leishmania infantum			
Índice	Rede Predita	Rede Aleatória	P-value
Coeficiente de Agrupamento	0.424	0.160±0.003	$p<0.05$
Média do Caminho Mais Curto	2.886	2.573±0.004	$p<0.05$

Como resultado, as três redes de interação preditas incorporaram 23%, 25% e 24% das proteínas dos proteomas preditos filtrados de *L. braziliensis*, *L. infantum* e *L. major* (Tabela 5). A Figura 13 exibe as três redes preditas neste estudo.

Além disso, os termos da ontologia GO foram utilizados para extrair um perfil de função das redes. Para esta análise, os termos ditos como preditos presentes na base de dados TritrypDB foram empregados ao invés dos termos ditos anotados. O motivo por de trás desta escolha está associado com o pequeno número de termos GO anotados para *L. braziliensis*, o que iria impedir a comparação da rede predita para esta espécie com as outras duas redes. Assim, as três ontologias presente no GO foram aplicadas (Processo Biológico – P , Componente Celular – C e Função Molecular – M), e resultados similares foram encontrados. Considerando uma frequência maior do que 2 para um dado termo GO, o total de interseção entre as redes preditas for 79%, 84% e 75% para P, C e M, respectivamente. Além disso, entre os 10 termos mais frequentes para cada ontologia, 8 deles para P, 7 deles para M e todos eles para C eram os mesmos para as três redes.

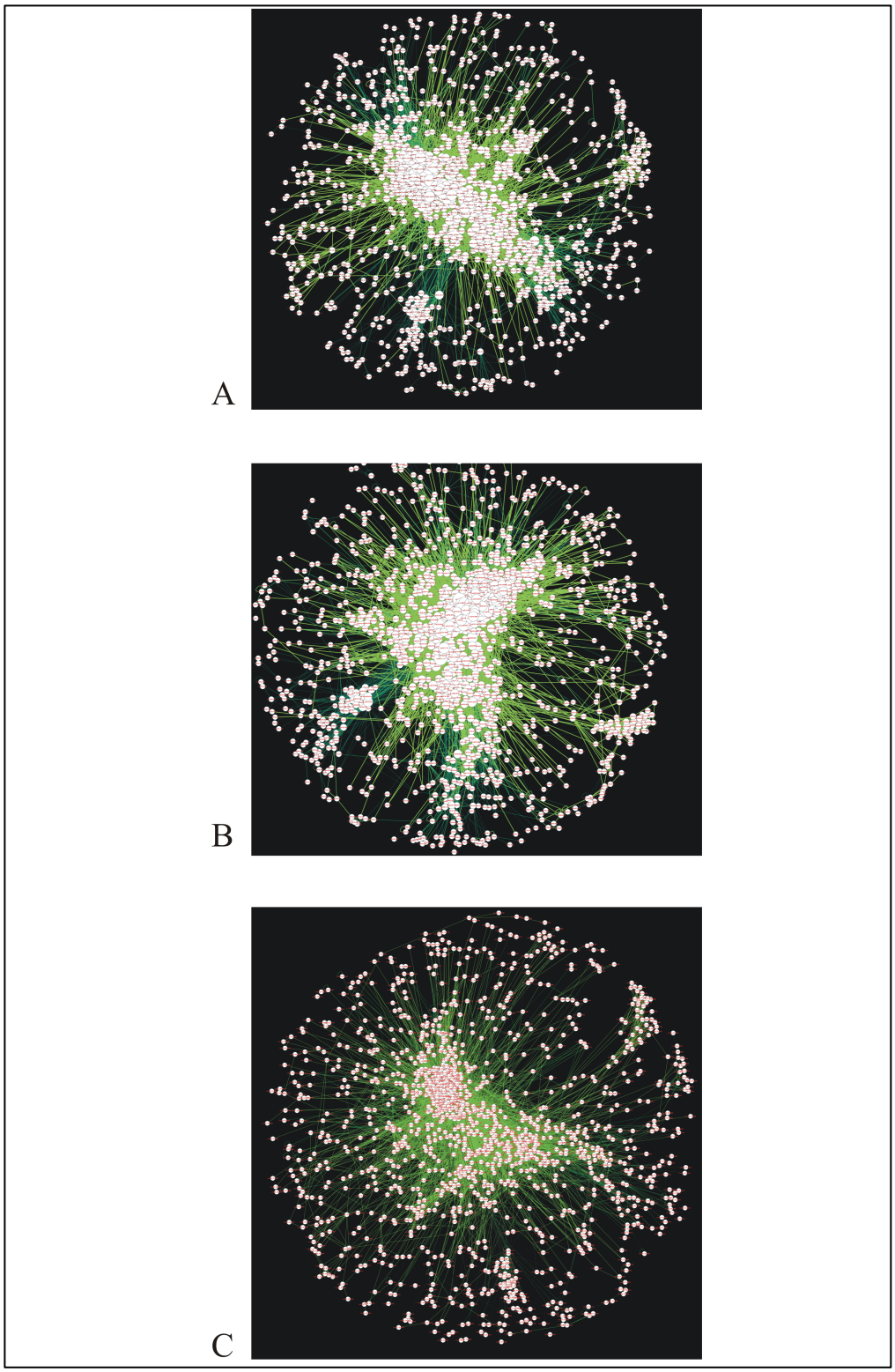


Figura 13 – Redes de interação de proteínas preditas neste estudo. (A) Rede predita utilizando o proteoma predito de *L. braziliensis*. (B) Rede predita utilizando o proteoma predito de *L. major*. (C) Rede predita utilizando o proteoma predito de *L. infantum*.

5.4 – Análise Evolutiva

Com o objetivo de obter informação relativa à correlação entre o número de interação que uma proteína faz (*Degree*) e o seu grau de conservação, o número de interações das proteínas presentes nas redes de interação modeladas foi comparado contra a diversidade nucleotídica dos genes que codificam as mesmas.

Com base nesta análise, foi possível perceber que quando as proteínas aumentam o número de interações que elas fazem, seus graus de diversidade, medidos aqui pelo π (índice de diversidade nucleotídica), diminuem (Figura 14).

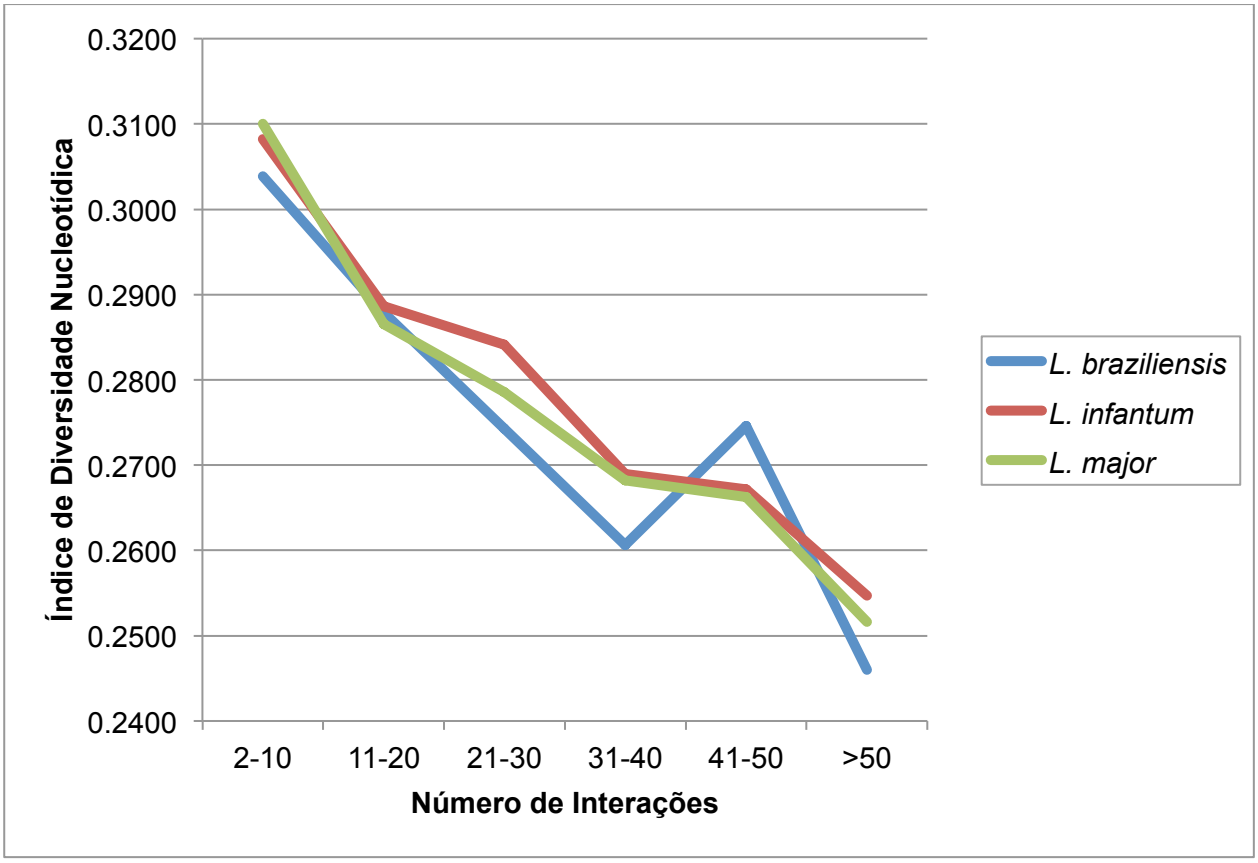


Figura 14 – Número de interações (*Degree*) versus análise de diversidade. Gráfico da mediana do índice de Diversidade Nucleotídica (π) (eixo Y) versus faixa de *Degree* (eixo X) das três redes modeladas.

Após a obtenção destes resultados para as três redes de interação de proteína preditas, é possível sugerir a existência de uma pressão evolutiva para a manutenção de uma baixa diversidade em proteínas com um alto número de interações.

5.5 – Caracterização dos Módulos

Durante esta etapa do trabalho, o algoritmo networkBLAST foi utilizado para identificar os módulos nas redes de interação de proteínas preditas. O número de módulos conservados compartilhados pelas três espécies do gênero *Leishmania* foi igual a 199. Apesar do milhões de ano de divergência propostos para as espécies analisadas, este resultado não é surpreendente considerando que uma alta sintenia já foi observada e reportada entre todas as espécies do gênero *Leishmania* sequenciadas.

Subsequentemente, como detalhado na seção de metodologia, uma anotação funcional para os módulos das redes foi realizada utilizando a hierarquia de termos Processo Biológico da ontologia GO e o conjunto de módulos de programação em Perl chamado GO::TermFinder. Esta abordagem permitiu a identificação de 153 módulos que tinham termos GO com uma frequência acima do esperado. Além disso, em casos onde um dado módulo da rede recebeu mais de um termo GO, o termo mais significante caracterizado pelo menor P-value na análise de enriquecimento foi escolhido para descrever o módulo. Vale a pena mencionar que diferentemente dos algoritmos padrões de agrupamento, a abordagem utilizada pelo networkBLAST pode produzir módulos com sobreposições. Este fato faz sentido do ponto de vista biológico, uma vez que uma proteína pode pertencer a mais de um módulo nas redes de interação.

Além disso, é também importante destacar que somente 57 termos GO únicos foram utilizados para descrever os 153 módulos preditos nas redes de interação, e os

termos mais frequentes foram atribuídos a módulos que estão provavelmente envolvidos em processos biológicos relacionados a estruturação de proteínas ("*protein folding*"), tradução, aminoacilação de tRNA para tradução de proteínas, geração de energia através da oxidação de compostos orgânicos e metabolismo de carboidratos.

Levando em consideração a significância biológica desta análise funcional, estes resultados foram sobrepostos com as análises topológicas.

5.6 – Análise Topológica e Predição de Epítopos

De acordo com a metodologia proposta, dois índices topológicos (*Degree* e MCC) foram utilizados para estudar as redes de interações preditas. Assim, as proteínas presentes nas redes foram organizadas de forma decrescente com base no índice MCC. Posteriormente, uma lista de proteínas, que potencialmente são centrais para vários cliques (subgrafos), e com alto grau de interação (alto *Degree*) foi obtida (Anexo II).

As análises posteriores que foram realizadas para esta lista de proteínas são: a) análise da variabilidade de aminoácidos presentes nos grupos de ortólogos; b) análise do grau de conservação contra proteínas de três hospedeiros potenciais (*M. musculus*, *C. lupus familiaris* e *H. sapiens*); c) predição computacional de epítopos.

Em relação a variabilidade destas proteínas, nossos resultados revelaram uma identidade média de 80% entre as 20 proteínas mais bem posicionadas em relação ao índice MCC e seus ortólogos (Anexo II). Portanto, foi possível notar que estas proteínas são relativamente conservadas entre os Kinetoplastídeos.

Além disso, somente duas proteínas, LbrM22_V2.0510 (ATPsae reguladora de proteassomo subunidade I) e LmjF36.1650 (proteassomo subunidade 5 beta), de *L.*

braziliensis e *L. major*, respectivamente, tiveram identidade maior que 60% quando comparadas contra os proteomas preditos dos potenciais hospedeiros. Ainda em relação à esta comparação, *L. infantum* apresentou duas proteínas com identidade maior do que 60%. São elas: LinJ36_V3.1730 (proteassomo subunidade 5 beta) e LinJ22_V3.0490 (ATPase reguladora de proteossomo subunidade 5).

Neste contexto, podemos sugerir que a baixa identidade apresentada pela grande maioria das proteínas analisadas pode ser interessante do ponto de vista do estudo para novos alvos de drogas e vacinas.

O racional em sugerir que estas proteínas poderiam ser utilizadas para propósitos médicos pode ser reforçado pela função predita dos módulos presentes nas redes de interação onde estas proteínas estão inseridas. Foi possível perceber que a maioria dos módulos estão envolvidos na renovação do repertório proteico ("*protein turnover*") dos organismos, o que já se sabe estar envolvido nas respostas à vacinação (Garlick, Mcnurlan *et al.*, 1980). Além disso, foram encontrados um total de 9 termos GO descrevendo estes módulos, e 7 deles são compartilhados pelas proteínas selecionadas pelo MCC para cada rede de interação predita.

Finalmente, em relação ao potencial imunológico para estas proteínas (Anexo II), todas elas tiveram mais de 5 epítopos preditos para receptores de célula B. Para a predição de epítopos para MHC classe I, 12 alelos diferentes foram testados e todas proteínas tiveram no mínimo 2 epítopos preditos com potencial afinidade de ligação a no mínimo 11 alelos. A última análise realizada foi referente a predição de epítopos para MHC classe II. O preditor utilizado para esta análise fornece junto com a predição de epítopos uma medida de afinidade de ligação entre o epítopo e o receptor. Esta medida é dividida em duas categorias: ligação fraca (WB – "*weak binding*") e ligação forte (SB – "*strong binding*"). Foram selecionadas apenas as predições que

foram classificadas como SB. Assim, todas as proteínas tiveram no mínimo dois epítopos com afinidade de ligação a pelos menos 1 alelo testado. O total de alelos testados foi 17.

5.7 – Anotação de Proteínas Hipotéticas

Com o objetivo de utilizar as redes de interação de proteína no processo de atribuição de algum nível de anotação funcional às proteínas hipotéticas, decidimos utilizar a abordagem chamada de FS-Weigth, a qual leva em consideração ambos vizinhos diretos e indiretos como detalhado na seção Materiais e Métodos.

Do número total de proteínas albergadas pelas redes preditas, aproximadamente 21% foram originalmente anotadas como proteínas hipotéticas. Deste conjunto de proteínas aproximadamente 40%, 48% e 55% receberam algum termo GO baseado na anotação utilizando as redes e o algoritmo FS-Weight (Tabela 7).

Tabela 7 – Predição de anotação para proteínas hipotéticas presentes nas redes

Organismo	Número de Nós (Proteínas)	Número de Proteínas Hipotéticas	Número de Proteínas Hipotéticas Anotadas (%)*
L. braziliensis	1818	381	153 (40%)
L. major	1947	416	200 (48%)
L. infantum	1959	415	229 (55%)

*Proteínas foram anotadas seguindo metodologia descrita no texto.

Além disso, é importante destacar que esta abordagem fornece uma pontuação para todas as predições de anotação que varia de 0 a 1, e apenas termos GO que receberam pontuação igual a 1 foram considerados.

Outro ponto importante desta análise foi que quando as informações a respeito dos módulos preditos nas redes foram cruzadas contra a anotação predita das proteínas hipotéticas, foi possível perceber que para as três redes de interação as proteínas hipotéticas são mais frequentes em módulos envolvendo metabolismo de RNA.

6 – Discussão

As doenças infecciosas estão ranqueadas entre as principais notícias na mídia tais como desastres naturais, situações de conflito e terrorismo. Doenças infecciosas emergentes com grande potencial de ameaça tais como síndrome respiratória aguda severa e uma pandemia de influenza, são usualmente itens com alto impacto na mídia, e consequentemente angariam maiores apoios financeiros. Doenças com alta prevalência tais como HIV/AIDS podem trazer melhores retornos financeiros de investimentos em pesquisa para novos tratamentos e desenvolvimento de vacinas do que infecções com menor prevalência. A leishmaniose é uma das doenças que raramente compartilham espaços na mídia e portanto se mantêm como sendo uma doença negligenciada (Piscopo e Mallia, 2006).

Além disso, as leishmanioses são doenças transmitidas por vetores insetos, e o impacto do aquecimento global na distribuição geográfica do insetos (flebótomos) infectados com parasitas sugerem que as leishmanioses podem se tornar amplamente disseminadas, aumentando sua gravidade como problema de saúde pública (Costa, Peters *et al.*, 2011). Mudanças políticas e socioeconômicas podem ter ainda um papel mais importante do que o aquecimento global na mudança da epidemiologia das leishmanioses. Os determinantes mais importantes para emergência das leishmanioses incluem aumento da pobreza, urbanização e migração humana. Em áreas empobrecidas, as casas com o piso e as paredes sem nenhuma cobertura juntamente com a falta de infraestrutura sanitária e coleta de lixo combinam para criar sítios em que os insetos transmissores das leishmanioses podem se proliferar (Piscopo e Mallia, 2006; Hotez, Bottazzi *et al.*, 2008).

De fato, nos últimos 20 anos no Brasil, o padrão epidemiológico por exemplo da leishmaniose visceral tem mudado de uma doença esporádica de áreas rurais para um doença de ocorrência em regiões peri-urbanas que afeta diversas camadas socioeconômicas, com uma tendência no aumento da mortalidade (Costa, Peters *et al.*, 2011).

Ainda dentro deste cenário, diversos problemas em relação as leishmanioses possuem mérito para discussão: resistência a drogas utilizadas no tratamento tem sido relatada em certas regiões do globo, fazendo necessário a mudança dos agentes de tratamento utilizados como primeira-linha; apesar do avanços no entendimento da resposta imune contra o parasito, e do sequenciamento do genoma de várias espécies do gênero *Leishamania*, uma vacina eficiente ainda não foi desenvolvida (Piscopo e Mallia, 2006).

Portanto, as leishmanioses são, até o momento, doenças sem maneiras eficientes de prevenção, e com mudanças nos perfis epidemiológicos. Logo este cenário atual demanda novos instrumentos para tratamento e controle destas doenças (Costa, Peters *et al.*, 2011).

Os mecanismos de atuação de um fármaco, da resistência por parte do parasito a uma determinada droga e os eventos envolvidos na resposta imune por parte do parasito são complexos. Estes têm sido estudados através do método científico tradicional de elaboração de hipótese, e validação experimental, particularmente através de abordagens reducionistas da biologia molecular. No entanto, ao mesmo tempo que estas abordagens são poderosas, elas oferecem uma visão limitada dos sistemas biológicos complexos (Pulendran, Li *et al.*, 2010). Nos genomas das três espécies do gênero *Leishmania* alvos deste estudo, foram preditos mais de 8.000 genes, assim uma busca por novos alvos para drogas e vacinas não é uma problema

simples. A Biologia de Sistemas, com uma proposta mais holística, pode nos oferecer uma solução para este problema.

Na tentativa de contribuir para elucidação deste problema, a abordagem de Biologia de Sistemas foi empregada neste trabalho. Particularmente, a modelagem de redes de interação de proteínas foi utilizada em conjunto a abordagens de vacinologia reversa. Proteínas são tradicionalmente identificadas em relação às suas ações individuais. No entanto, a visão pós-genômica expandiu o conceito relativo ao papel de uma proteína para um elemento integrante de uma rede de interações de proteínas em que esta proteína possui uma função celular dentro de módulos funcionais (Jeong, Mason *et al.*, 2001).

Assim, a primeira etapa realizada em nosso trabalho foi avaliar a metodologia que seria empregada na modelagem das redes. Para isso, foi necessário obter uma rede de interação de proteína modelada experimentalmente para ser utilizada como controle. A rede de interações utilizada para este fim pertence à espécie de bactéria *E. coli*. Assim, em posse deste dado, as análises de desempenho para a metodologia chamada *Interolog Mapping* foram realizadas.

Baseado no resultado obtido nesta etapa, valor de AUC igual a 0,94, é possível afirmar que a metodologia de predição das redes de interação de proteína juntamente com esquema de pontuação aqui desenvolvido é robusto, e as interações de proteína preditas possuem uma boa confiança. No entanto, como foi dito na seção de resultados, as bases de dados utilizadas no desenvolvimento metodológico possuem muitas interações descritas para *E. coli*, fato esse que pode ter trazido algum viés na nossa avaliação de desempenho cujo grau permanece difícil de ser avaliado. Contudo, utilizando as mesmas bases de dados e a mesma abordagem, outros esquemas de pontuação da interação foram avaliados. Estes esquemas também estão sujeitos aos

mesmos problemas do esquema de pontuação desenvolvido neste trabalho, assim seria esperado que o desempenho destes esquemas também fosse alto devido ao viés. Porém, é possível perceber pelos resultados (valores de AUC presentes na Tabela 2) que estes esquemas tiveram um desempenho muito inferior em relação ao esquema desenvolvido neste trabalho. Assim, mesmo com uma possível superestimativa da análise de desempenho devido à presença de interações de *E. coli* nas bases utilizadas, os outros métodos de pontuação tiveram seus valores de AUC muito abaixo, o que confere um maior grau de confiança no resultado obtido para nosso método.

Além disso, recentemente Kamburov A. e colaboradores (Kamburov, Grossmann *et al.*, 2012) descreveram um método, chamado CAPPIC ("*cluster-based assessment of protein-protein interaction confidence*") baseado na topologia das redes de interação para o cálculo de pontuação de confiança. Este método assume que as redes são modulares, e que proteínas que são específicas para um certo módulo funcional são esperadas ter mais interações com proteínas que são específicas para o mesmo módulo do que com outra proteínas. O desempenho alcançado para este método chegou a um valor de AUC de 0,94. Deste modo, este método foi empregado utilizando a rede predita de interações de proteína de *E. coli* avaliada em nossa análise de desempenho sem nenhum valor de pontuação. É possível perceber a partir dos resultados que a grande maioria das interações obtiveram uma pontuação acima de 0,8 (Figura 15), o que corrobora a predição de interações realizada utilizando o esquema de pontuação desenvolvido neste trabalho.

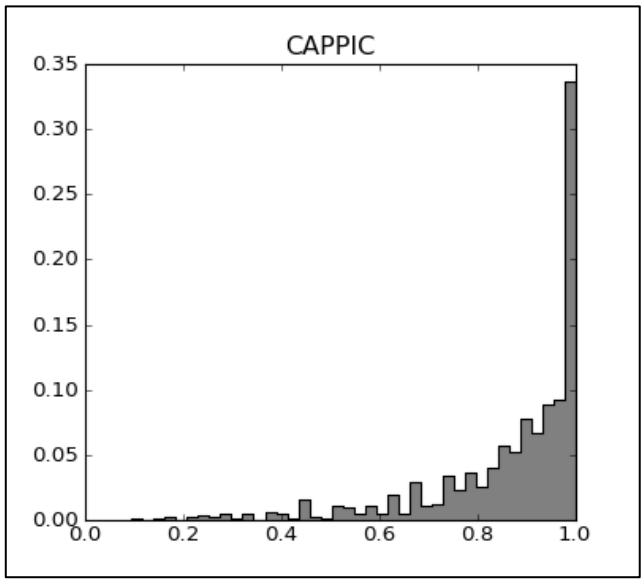


Figura 15 – Resultado do cálculo de pontuação de confiança para a rede de interação de proteína de *E. coli* predita por nossa metodologia utilizando o método CAPPIC. Eixo X – valor da pontuação de confiança calculado. Eixo Y – frequência das interações de proteína para um determinado valor de pontuação.

Com o intuito de verificar também a distribuição de pontuação de confiança para as redes das espécies do gênero *Leishmania* aqui modeladas, o método CAPPIC foi utilizado juntamente com as três redes de interação preditas. Estas também não continham nenhum cálculo de pontuação no momento que foram carregadas pelo servidor do método CAPPIC. Assim é possível notar que a maioria das interações recebem valores altos de pontuação, e que os perfis das distribuições são similares entre si (Figura 16). Novamente, através de uma outra perspectiva de análise, demonstramos a robustez da abordagem empregada neste estudo para a predição de interações proteicas e a alta conservação existente entre as três espécies de leishmania alvos do trabalho.

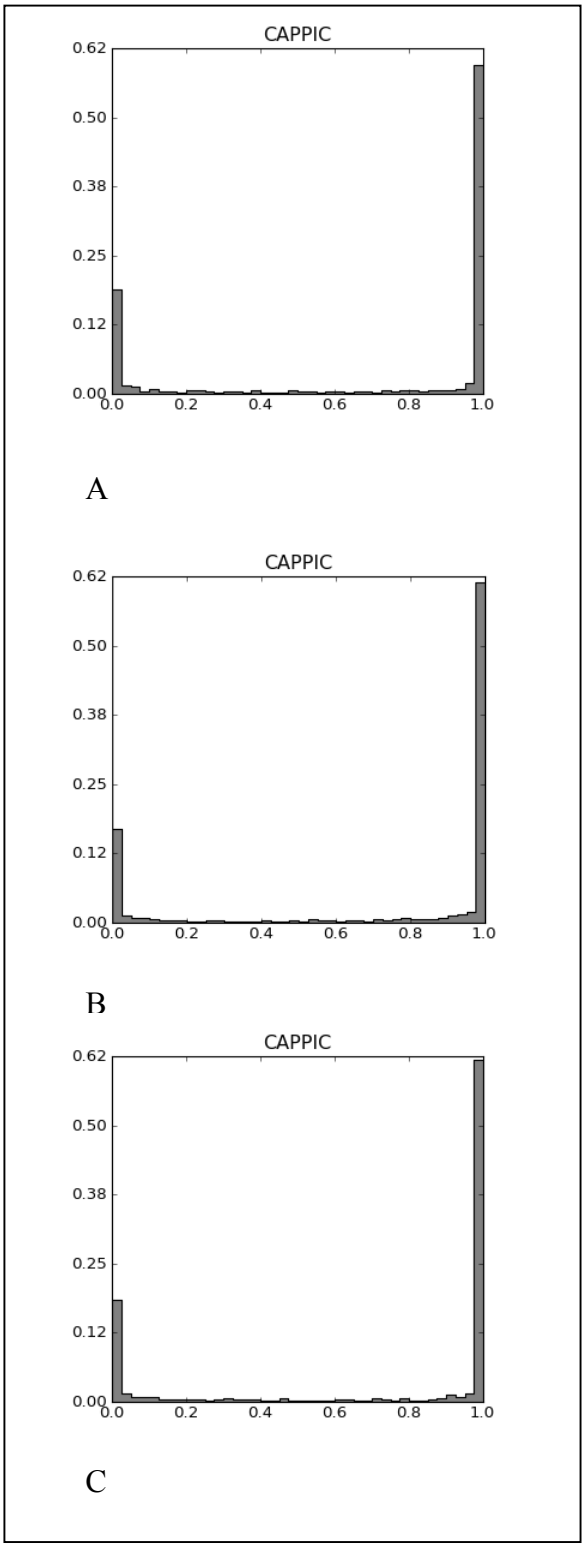


Figura 16 - Resultado do cálculo de pontuação de confiança para as redes de interação de proteína dos organismos alvos preditas através da nossa metodologia utilizando o método CAPPIC. Eixo X – valor da pontuação de confiança calculado. Eixo Y – frequência das interações de proteína para um determinado valor de pontuação. A – Distribuição de pontuação para a rede de *L. braziliensis*. B –

Distribuição de pontuação para a rede de *L. infantum*. C – Distribuição de pontuação para a rede de *L. major*.

Em relação à qualidade dos dados utilizados na predição das redes de interação dos organismos alvos do estudo, a perda de proteínas com o passo de filtragem foi pequena (detalhes na seção resultados), e isto reflete a qualidade da anotação gênica do genoma de *L. braziliensis*, *L. infantum* e *L. major*. Este fato possui uma grande importância uma vez que o principal dado de entrada para a metodologia aqui aplicada foram as sequências proteicas preditas destes genomas, e o resultado final está ligado diretamente à qualidade destes dados.

Ainda dentro do contexto da qualidade da predição computacional, na seção de resultados, a avaliação das redes de interação preditas para os organismos alvos baseada em alguns modelos de redes conhecidos tais como o modelo livre de escala foi descrita para garantir a confiança das mesmas. É possível sugerir que as redes aqui preditas são consistentes, uma vez que elas apresentam características que são comuns às redes biológicas descritas atualmente (Barabasi e Albert, 1999; Barabasi e Oltvai, 2004), ou seja, elas apresentam um grande número de proteínas com poucas interações e poucas proteínas com muitas interações. Além disso, quando as redes modeladas foram comparadas contra redes aleatórias (Tabela 6), foi possível perceber que os valores de Coeficiente de Agrupamento das nossas redes são muito maiores do que os das redes aleatórias, fato este que novamente sugere a solidez da predição das redes e a falta de interações espúrias. Portanto, ambos resultados podem ser utilizados para ilustrar a confiança da abordagem *Interolog Mapping*, e para reforçar o resultado encontrado em nossa avaliação de desempenho. Outro fato importante em relação a estes resultados é o fato de que somente estas evidências poderiam já garantir uma avaliação consistente das interações preditas, conforme foi feito no trabalho de Flórez

e colaboradores (Florez, Park *et al.*, 2010). No entanto, ainda assim realizamos a avaliação de desempenho da metodologia utilizando um conjunto de dados experimentais de alta qualidade como controle.

Em relação aos números presentes nas redes de interação, os resultados aqui encontrados são comparáveis aos encontrados para *L. major* por Flórez e colaboradores (Florez, Park *et al.*, 2010), que encontraram aproximadamente 16% do proteoma predito de *L. major* na rede de interações de proteína predita. De acordo com os autores, a razão para o pequeno número de proteínas mapeadas dentro da rede é um reflexo de baixos níveis de similaridade entre as espécies de leishmania e o conteúdo das bases utilizadas na predição. Por outro lado, as diferenças encontradas entre o número de interações preditas observadas em nosso estudo e no de Flórez e colaboradores pode ser explicada pelas diferentes fontes de informação e abordagens utilizadas.

Posteriormente à avaliação das redes, a primeira análise realizada nas três redes de interação foi uma anotação funcional utilizando a ontologia do *Gene Ontonlogy*. Deste modo, foi possível perceber que os termos mais frequentes para as três redes de interação dentro da ontologia de Função Molecular são relacionados com a função de ligação. Este fato faz sentido uma vez que as proteínas presentes nas redes preditas interagem umas com as outras. Por outro lado, dentro da ontologia Componente Celular, termos associados com complexos proteicos tais como proteassomo e ribossomo foram os mais observados. Novamente, isto é de alguma maneira esperado pois um conjunto de proteínas interagindo provavelmente irá formar complexos, além disso as redes biológicas possuem, como já mencionado, arquitetura modular, logo diversos complexos coexistem nas redes. Contudo, para a ontologia de Processos Biológicos, nós obtivemos uma alta variedade de termos, o

que pode ser explicado pelo fato de que a mesma proteína está sujeita a participar de diversos processos biológicos diferentes em uma célula.

Uma análise de cunho evolutivo também foi realizada com o objetivo de verificar se existia alguma tendência relacionada ao número de interações e a diversidade de sequência em uma proteína. Os resultados obtidos indicam que o número de interações e a diversidade são inversamente proporcionais; isto quer dizer que o aumento da diversidade é representado pela queda no número de interações que uma proteína faz dentro da rede de interações. Em redes de interações de proteína-proteína, proteínas apresentando diversas interações (alto valor de *Degree*) são geralmente chamadas de nós *hubs* e amplos estudos genômicos (Jeong, Mason *et al.*, 2001; Lee, Lehner *et al.*, 2008) têm mostrado que deleções de uma proteína do tipo *hub* é mais provável de ser letal do que deleções de proteínas não *hubs*, isto foi denominado regra da centralidade-letalidade. Este fato faz sentido do ponto de vista biológico uma vez que estas proteínas (*hubs*) provavelmente estão envolvidas em diferentes processos biológicos dentro da célula com relativo sucesso e, neste contexto, se uma mutação aleatória ocorre, ela provavelmente produzirá um resultado negativo.

Portanto, os pontos discutidos até este momento ilustram que as redes de interações preditas aqui para as espécies *L. braziliensis*, *L. infantum* e *L. major* são biologicamente consistentes. Caso contrário, nós teríamos uma tendência de proteínas com uma ampla diversidade, isto é, pouco conservadas, sendo *hubs*.

Um outro ponto importante, ao qual foi dado ênfase dentro deste estudo, foi a análise de modularidade realizada para as redes preditas. Modularidade é uma medida da estrutura das redes, e em muitos trabalhos prévios tem sido relatado que redes biológicas são modulares (Hartwell, Hopfield *et al.*, 1999; Ravasz, Somera *et al.*,

2002; Barabasi, Ravasz *et al.*, 2003; Barabasi e Oltvai, 2004; Pereira-Leal, Enright *et al.*, 2004; Brohee e Van Helden, 2006). Esta característica é importante para a robustez da rede uma vez que uma arquitetura modular garante que uma falha no sistema seja isolada (Barabasi e Oltvai, 2004). Alguns pesquisadores consideram que os módulos funcionais sejam um nível crítico na organização biológica (Hartwell, Hopfield *et al.*, 1999). Assim, se nós estamos interessados em desestabilizar as redes de interação com propósitos de desenvolvimento de novos alvos para droga e vacina, nós precisamos conhecer os módulos presentes dentro das redes preditas.

Dentro deste contexto, com o objetivo de medir a modularidade, uma análise de agrupamento foi realizada para a identificação de módulos funcionais conservados. Esta análise foi feita utilizando a ferramenta denominada de networkBLAST. Como já foi mencionado anteriormente, os módulos preditos por este programa não são mutualmente exclusivos, isto é, uma proteína de um módulo pode fazer parte de outro módulo. Biologicamente, este tipo de predição faz sentido pois módulos funcionais não precisam ser rígidos, estruturas fixas. Uma dada proteína pode pertencer a diferentes módulos inúmeras vezes. A função de um módulo pode ser quantitativamente regulada, ou trocada entre diferentes funções qualitativas, por sinais químicos de outros módulos. Funções de mais alto nível podem ser construídas através da conexão de módulos. Por exemplo, o supermódulo cuja função é a distribuição acurada dos cromossomos para células filhas na mitose contém módulos que unem o fuso mitótico, módulos que monitoram o alinhamento cromossomal ao fuso mitótico, e um oscilador de ciclo celular que regula as transições entre interfase e mitose (Hartwell, Hopfield *et al.*, 1999).

Assim, utilizando a ferramenta networkBLAST, foram encontrados 153 módulos funcionais conservados que tiveram uma função atribuída através da análise

de enriquecimento de termos da ontologia GO. Além disso, estes módulos puderam ser agrupados num total de 57 funções diferentes (57 termos GOs diferentes). Destas 57 funções, foi possível perceber que as mais frequentes estavam relacionadas ao enovelamento correto de proteínas, tradução, aminoacilação de tRNA para tradução, geração de energia através da oxidação de compostos orgânicos e metabolismo de carboidratos.

Baseado nestes resultados, é possível perceber que existem muitos complexos proteicos (módulos) que são essenciais para os organismos em estudo. Assim, é de fundamental importância explorar com maior profundidade estes complexos juntamente com as informações topológicas das proteínas das redes com potencial para serem eleitas novos alvos para desenvolvimento de drogas e vacinas.

Além disso, outras fontes de informações foram integradas à análise topológica tais como potencial imunológico, grau de conservação da sequência de aminoácido entre os ortólogos, e grau de conservação comparado a proteínas dos potenciais hospedeiros, homem, cachorro e camundongo. Esta integração das informações fornece um melhor entendimento que pode ser valioso para a seleção de novos potenciais alvos biológicos.

Deste modo, utilizando deste racional, nós sugerimos uma lista de proteínas (Anexo II) que pode ser atrativa para propósitos médicos. Estas proteínas possuem baixa identidade contra proteínas dos hospedeiros, elas são potencialmente reconhecidas per receptores de células B e células T, e são altamente conservadas quando comparadas com seus ortólogos. Além disso, elas parecem ser centrais para muitos processos biológicos uma vez que possuem altos valores de MCC e *Degree*. Algumas proteínas desta lista fazem parte de módulos funcionais importantes citados anteriormente. Para os módulos envolvidos no enovelamento proteico correto,

existem 2 proteínas de cada um dos organismos alvos. Para os módulos com função de tradução atribuída, existem 3 proteínas na lista de *L.major* e *L. braziliensis* sendo que uma das proteínas da última espécie (LbrM26_V2.0980) está anotada como proteína hipotética. As proteínas hipotéticas presentes nas redes passaram por um processo de anotação discutido mais à frente. Uma das anotações preditas para esta proteína é a atividade de síntese de espermidina, que já teve sua atividade reportada na inibição da síntese de óxido nítrico e no auxílio da transcrição do RNA (Wan e Wilkins, 1993; Hu, Mahmoud *et al.*, 1994). Assim, esta proteína poderia também estar relacionada ao escape do sistema imune do hospedeiro. Para a espécie *L. infantum*, nenhuma proteína da lista está inserida em módulos preditos para o processo biológico de tradução. Para os módulos envolvidos em aminoacilação de tRNA para tradução, foram encontradas 2 e 3 proteínas nas lista para as espécies *L. major* e *L. infantum*, respectivamente. Para *L. braziliensis*, também foram mapeadas três proteínas na lista, contudo uma estava anotada como proteína hipotética (LbrM09_V2.0920). Entre as anotações atribuídas a esta proteína em nosso estudo estão: componente do ribossomo, e atividade de síntese também em espermidina. Assim, devido ao seu contexto funcional e topológico, torna-se um alvo interessante a ser estudado.

Deste modo, se as proteínas presentes nesta lista forem neutralizadas todo o sistema de interações de proteínas pode sofrer danos graves. Razão pela qual a topologia e a funcionalidade das mesmas devem ser estudadas.

Antes de aprofundar com mais detalhe na discussão destes dados, vale a pena abrir a discussão para a avaliação dos algoritmos de predição do potencial imunogênico das proteínas. Esta foi uma etapa importante deste estudo, uma vez que ela permitiu uma escolha mais consciente das ferramentas a serem utilizadas nas

análises posteriores. Além disso, a predição de epítopos por métodos computacionais representa uma das abordagens mais promissoras para desenvolvimento de vacinas. Contudo existem diversos problemas neste processo relativo aos genomas de tripanosomatídeos. Neste contexto, a falta de um conjunto de dados experimentais validados de epítopos de protozoários com tamanho suficiente para validação da predição *in silico* de epítopos representa uma séria limitação.

Atualmente diversos métodos de predição de epítopos já foram desenvolvidos, contudo nenhum deles possui dados de parasitos protozoários suficientemente representados como dado de treinamento para o algoritmo de predição (Peters, Bulik *et al.*, 2003; Larsen, Lund *et al.*, 2006; El-Manzalawy, Dobbs *et al.*, 2008; Yasser, Dobbs *et al.*, 2008; Nielsen e Lund, 2009; Nielsen, Lundegaard *et al.*, 2009). Consequentemente, os resultados de desempenho destes preditores descritos na literatura devem ser interpretados com cuidado. O senso comum é que o desempenho dos métodos de predição de epítopos criticamente depende do conjunto de dados utilizado como treinamento e também do viés composicional das proteínas deste conjunto. Em relação à predição de epítopos em genomas de parasitos, estes problemas são evidentes considerando que estes organismos têm um conteúdo genômico com proteínas com perfis físico-químicos peculiares, e que são sub-representadas nos dados de treinamento.

Portanto, a avaliação de desempenho realizada aqui focou no desempenho frente ao genomas de parasitos. A comparação entre os algoritmos foi realizada com base também em valores de AUC, isto é, a probabilidade de que uma predição positiva selecionada aleatoriamente tenha uma pontuação mais alta do que uma predição negativa selecionada aleatoriamente (Wu e Flach, 2005). Como já foi mencionado ne seção de métodos, nosso dado experimental utilizado como controle

teve sua origem na bases de dados do IEDB, que atualmente representa a principal fonte de epítopos lineares e conformacionais. Além disso, ele utiliza uma métrica que leva em conta o número de referência, número de ensaios positivos, e o número total de ensaios para cada epítopo, o que é crucial para obter um subconjunto de epítopos validados experimentalmente com um alto grau de confiança para as análises de desempenho.

Em relação aos algoritmos para células T CD8+, nossos resultados de AUC indicam uma pequena diferença nos desempenhos relacionados aos algoritmos NetCTL e NetMHC. Os valores foram iguais a 0,66 e 0,60, respectivamente. Se considerarmos que já foram reportados métodos de predições de epítopos do mesmo tipo com uma acurácia que em muitas vezes permitiu valores de AUC entre 0,95 a 0,99 (Larsen, Lundegaard *et al.*, 2007), ambos algoritmos não alcançaram o desempenho esperado. De fato, esta não é a primeira vez que um desempenho abaixo do esperado para algoritmos de predições de epítopos é relatada na literatura. Em um estudo recente, 167 peptídeos de 9 aminoácidos do vírus *Influenza A* foram preditos como ligantes potenciais pelo NetMHC, e apenas 89 deles (53% do total) foram confirmados como verdadeiros ligantes (Wang, Lamberth *et al.*, 2007). Assim, novamente acreditamos que a sub-representação de proteínas de protozoários nos dados de treinamento dos algoritmos, e o viés composicional das sequências de protozoários tiveram um forte impacto nos métodos de predição e consequentemente no desempenho dos mesmos. Deste modo, para destacar a diferença de desempenho dos algoritmos testados em frente a diferentes conjuntos de dados, e excluir a influência da abordagem realizada, nós avaliamos os desempenhos dos algoritmos dentro da mesma estrutura contudo utilizando proteínas humanas disponíveis para carregamento no site do NetCTL (http://www.cbs.dtu.dk/services/NetCTL/). Os

resultados para ambos algoritmos NetCTL e NetMHC foram considerados melhores do que os resultados obtidos para o conjunto de dados de protozoários. O valor de AUC para NetCTL foi de 0,80 e para o NetMHC foi de 0,77 (Figura 12A). A avaliação de desempenho realizada aqui não incluiu algoritmos de predição de epítopos para células T CD4+, uma vez que não existiam dados experimentais suficientes. Na prática, a predição de peptídeos que se liguem a moléculas de MHC está longe de ser perfeita, contudo este fato não anula todos os avanços alcançados nesses últimos anos nesta área de pesquisa (Ostell e Kans, 2006).

Em relação aos epítopos com afinidade para receptores de célula B, os resultados de AUC indicaram um melhor desempenho para o algoritmo BCPred12 quando comparado aos algoritmos AAP12 e BepiPred (Tabela 3). Novamente os desempenhos observados foram inferiores daqueles atualmente observados para este tipo de predição (Larsen, Lund *et al.*, 2006). Esta diferença pode ser explicada pelas mesmas razões que foram discutidas para a predição de epítopos de célula T CD8+. Além disso, também não é primeira vez que um desempenho ruim é reportado na literatura para a predição de potenciais epítopos com afinidade de ligação em anticorpos (Sebatjane, Pretorius *et al.*, 2010).

Lafuente e Reche 2009 [referencia] acreditam que o lançamento do *Critical Assessment of Techniques for Epitope Prediction* beneficiará esta área de pesquisa. Dentro deste evento, métodos computacionais serão utilizados para predição cega *de novo* de peptídeos, que são imunogênicos, de proteínas que, para propósitos de avaliação, têm sido verificadas experimentalmente (Ostell e Kans, 2006). Juntamente a este fato e à abordagem e os resultados obtidos neste trabalho, acreditamos que este evento será útil para trazer avanços no campo de predição de epítopos.

Apesar dos contrapontos citados até então na discussão em relação às predições de epítopos, a análise de desempenho dos algoritmos combinados sugere que a combinação de predições pode ser uma abordagem promissora. Para algoritmos de célula B, quando a análise de desempenho foi realizada combinando resultados de diferentes algoritmos, o melhor desempenho foi encontrado para AAP12 e BCPred12 que alcançaram juntos um AUC igual a 0,77 o que está dentro da faixa já reportada na literatura (Larsen, Lund *et al.*, 2006).

Retomando a discussão das análises realizadas para as proteínas presentes no Anexo II, observamos que a grande maioria delas não possuíam alto nível de identidade contra as proteínas dos proteomas preditos dos hospedeiros. Isto é uma característica desejável para proteínas que serão selecionadas para desenvolvimento de drogas e/ou vacinas. Das quatro proteínas, divididas entre as três espécies, citadas na seção de resultados que apresentaram uma identidade acima de 60% com proteínas dos hospedeiros, todas estão inseridas em módulos de catálise de proteínas. Sendo assim, a neutralização das mesmas pode acarretar danos nos processos de catálise dos hospedeiros. Selecionando as outras proteínas, é possível evitar efeitos colaterais indesejáveis.

Outra característica importante é o alto grau de conservação destas proteínas quando comparadas contra seus ortólogos. Como foi dito nos resultados, a média de identidade das proteínas da lista e seus ortólogos foi de 80%. Esta característica é interessante do ponto de vista imunológico e farmacêutico uma vez que ela possibilita a garantia de um amplo espectro de ação para vacinas e drogas a serem desenvolvidas. Deste modo, vacinas e drogas poderiam estar aptas a terem efeito contra mais de uma linhagem de leishmania de uma determinada espécie, e contra diversas espécies propriamente ditas. O fato das proteínas mais bem ranqueadas serem conservadas

notoriamente era esperado, uma vez que a análise com um enfoque evolutivo realizada aqui demonstrou esta tendência.

Por fim, as proteínas mais bem ranqueadas foram utilizadas para uma análise do potencial imunológico das mesmas. As análises foram feitas utilizando os preditores de epítopos com os melhores desempenhos avaliados neste trabalho. Segundo os resultados obtidos, todas as proteínas presentes na lista gerada tiveram um grande número de epítopos com afinidade para receptores de célula B preditos. Para as predições de epítopos com afinidade para receptores MHC classe I e MHC classe II (célula T CD8+ e célula T CD4+, respectivamente), todas as proteínas tiveram epítopos preditos para diversos alelos de MHC. Esta diversidade de reconhecimento é fundamental para a produção de uma resposta imune efetiva e ampla, e para que a vacina desenvolvida seja capaz de proteger uma maior parcela da população.

Dentre as proteínas com destaque topológico no trabalho de Flórez e colaboradores (Florez, Park *et al.*, 2010), três delas foram encontradas na rede modelada para *L. major*. Duas delas (LmjF36.1360 e LmjF25.2370) são adenilato quinases, estas possuem *Degree* acima de 20 e estão inseridas nos mesmos módulos funcionais: processo metabólico de nucleosídeo difosfato e processo biosintético de compostos contendo purina. Inibição destas proteínas causou baixo crescimento em promastigotas de *Leishmania donovani* (Villa, Pérez‐Pertejo *et al.*, 2003) e existem ortólogos destas proteínas nas outras duas espécies alvos deste trabalho, o que é vantajoso do ponto de vista para desenvolvimento de drogas e vacinas. A terceira proteína mapeada, LmjF36.2380, descrita como esterol metiltransferase, se mostrou pouco importante do ponto de vista topológico em nossa rede, com um *Degree* igual a 4.

Outra importante análise realizada com a redes modeladas em mãos foi a utilização das mesmas em uma abordagem para atribuir funções potenciais para proteínas presentes nas redes até então preditas como hipotéticas. Dentro do contexto dos genomas dos tripanosomatídeos, o estudo de proteínas hipotéticas possui uma importância enorme uma vez que alguns organismos deste táxon, o qual alberga entre outros organismos aqueles que são alvos de estudo deste trabalho, possuem aproximadamente 60% de seus proteomas preditos compostos de proteínas não caracterizadas. Este cenário é mantido atualmente mesmo considerando a era das "*omicas*" devido à maioria dos estudos frequentemente focarem em elementos dos organismos que já estão bem entendidos e estabelecidos em cenário molecular. Portanto, a oportunidade de expansão do conhecimento além do que é conhecido e esperado é raramente alcançada (Pawłowski, 2008).

Além disso, a maioria dos pesquisadores não estão interessados em investigar o dado molecular que muitas vezes é de difícil interpretação na luz do conhecimento biológico atual, por exemplo dados sobre as proteínas hipotéticas (Pawłowski, 2008). No entanto, as abordagens de Biologia de Sistemas podem auxiliar na melhoria destes números. Assim, existe um grupo de métodos no contexto da Biologia de Sistemas que objetiva explorar a informação derivada das redes para auxiliar na predição de função de proteínas. Desta maneira, vários métodos permitem uma predição de função generalista utilizando características livres da abordagem de similaridade de sequência ("*homology-free*") (Pawłowski, 2008).

Um exemplo de aplicação de um estudo de redes de interação para elucidar a função de uma proteína não caracterizada pode ser encontrado no trabalho de Cui e colaboradores (Cui, Zhang *et al.*, 2009), no qual eles construíram uma rede de interação de proteína-proteína para *Mycobacterium tuberculosis* utilizando a

abordagem de *Interolog Mapping*. Neste estudo, uma proteína hipotética com um alto valor de *Degree* foi encontrada e evidências para sua função foram derivadas do fato desta interagir com o mesmo grupo de subunidades ATPase de transportadores ABC que uma proteína conhecida interage. Portanto, este racional de atribuição de função baseado em vizinhos de uma proteína em uma rede de interação pode ser extremamente útil.

Em nossos resultados, aproximadamente 50% das proteínas hipotéticas presentes nas redes receberam alguma anotação funcional (Tabela 7). Além disso, os módulos mais frequentes, onde estas proteínas estavam inseridas, estão relacionados ao metabolismo de RNA. Isto pode ser interessante uma vez que existe atualmente uma grande quantidade de estudos envolvendo diferentes tipos de RNA e seus papéis em distintos fenômenos biológicos.

7 – Conclusão

Este estudo foi o primeiro a modelar três redes de interação de proteínas para três espécies diferentes de organismo do gênero *Leishmania* (*L. braziliensis*, *L. infantum* e *L. major*), e a comparar estas redes preditas entre elas.

Além disso, o método de predição de interações utilizado neste trabalho foi previamente avaliado em relação ao seu desempenho. Deste modo, um novo esquema de pontuação de confiança para as interações preditas foi proposto, e provou-se ser confiável.

Portanto, após as predições e avaliações das redes modeladas, foi observado que era possível extrair informações relacionadas à biologia dos organismos estudados. Utilizando as informações topológicas, proteínas com potencial para serem alvos de desenvolvimento de drogas e vacinas foram selecionadas.

Contudo, a seleção de novos alvos para drogas e vacinas apresenta-se como um problema multifatorial e complexo, assim um número maior de camadas de informação auxiliou na seleção das proteínas. Informações sobre a conservação das proteínas entre os seus ortólogos e entre as proteínas dos hospedeiros deste parasitos, e informações a respeito do potencial imunológico das mesmas foram adicionadas.

Em relação ao potencial imunológico, os algoritmos utilizados para as predições de epítopos foram previamente avaliados também em relação aos seus desempenhos. Os algoritmos mais bem colocados foram então utilizados nas predições. Outro ponto que merece atenção em relação a esta avaliação foi a apresentação do pequeno desempenho em geral dos algoritmos frente aos dados de protozoários. Este fato se deve provavelmente à falta de treinamento dos algoritmos

junto aos dados desses parasitos e às próprias peculiaridades apresentadas pelos proteomas desses organismos.

De volta às predições das redes de interação, foi possível, utilizando as informações de vizinhança para as proteínas presentes nos modelos, inferir algumas pistas relacionadas à função de algumas proteínas, que a princípio não possuíam nenhuma informação relacionada às suas funções moleculares na célula.

Finalmente, baseado nas evidência relatadas aqui, acreditamos que as redes de interações de proteínas modeladas neste estudo são biologicamente consistentes, e que podem ser utilizadas como ferramentas para diferentes tipos de estudos nestes organismos.

8 – Referências Bibliográficas

ARANDA, B. et al. The IntAct molecular interaction database in 2010. **Nucleic Acids Research,** v. 38, p. D525-D531, Jan 2010. ISSN 0305-1048. Disponível em: < <Go to ISI>://000276399100085 >.

ASHBURNER, M. et al. Gene Ontology: tool for the unification of biology. **Nature Genetics,** v. 25, n. 1, p. 25-29, May 2000. ISSN 1061-4036. Disponível em: < <Go to ISI>://000086884000011 >.

ASLETT, M. et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. **Nucleic Acids Research,** v. 38, p. D457-D462, Jan 2010. ISSN 0305-1048. Disponível em: < <Go to ISI>://000276399100073 >.

ASSENOV, Y. et al. Computing topological parameters of biological networks. **Bioinformatics,** v. 24, n. 2, p. 282-284, Jan 2008. ISSN 1367-4803. Disponível em: < <Go to ISI>://000252498500020 >.

BADER, G. D.; HOGUE, C. W. An automated method for finding molecular complexes in large protein interaction networks. **Bmc Bioinformatics,** v. 4, Jan 2003. ISSN 1471-2105. Disponível em: < <Go to ISI>://000181510200001 >.

BALDI, P.; ATIYA, A. F. How delays affect neural dynamics and learning. **Neural Networks, IEEE Transactions on,** v. 5, n. 4, p. 612-621, 1994. ISSN 1045-9227.

BAMBINI, S.; RAPPUOLI, R. The use of genomics in microbial vaccine development. **Drug discovery today,** v. 14, n. 5-6, p. 252, 2009. ISSN 1359-6446.

BARABASI, A. L.; ALBERT, R. Emergence of scaling in random networks. **Science,** v. 286, n. 5439, p. 509-512, 1999. ISSN 0036-8075. Disponível em: < <Go to ISI>://WOS:000083121200054 >.

BARABASI, A. L.; OLTVAI, Z. N. Network biology: Understanding the cell's functional organization. **Nature Reviews Genetics,** v. 5, n. 2, p. 101-U15, 2004. ISSN 1471-0056. Disponível em: < <Go to ISI>://WOS:000188602400012 >.

BARABASI, A. L.; RAVASZ, E.; OLTVAI, Z. Hierarchical organization of modularity in complex networks. **Statistical Mechanics of Complex Networks,** v. 625, p. 46-65, 2003 2003. ISSN 0075-8450. Disponível em: < <Go to ISI>://WOS:000185214000003 >.

BATES, P. A. Transmission of< i> Leishmania</i> metacyclic promastigotes by phlebotomine sand flies. **International journal for parasitology,** v. 37, n. 10, p. 1097-1106, 2007. ISSN 0020-7519.

BLATT, M.; WISEMAN, S.; DOMANY, E. Superparamagnetic clustering of data. **Physical Review Letters,** v. 76, n. 18, p. 3251-3254, Apr 1996. ISSN 0031-9007. Disponível em: < <Go to ISI>://A1996UG82900002 >.

BORJA-CABRERA, G. P. et al. Effective immunotherapy against canine visceral leishmaniasis with the FML-vaccine. **Vaccine,** v. 22, n. 17, p. 2234-2243, 2004. ISSN 0264-410X.

BOYLE, E. I. et al. GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. **Bioinformatics,** v. 20, n. 18, p. 3710-3715, Dec 2004. ISSN 1367-4803. Disponível em: < <Go to ISI>://000225786600064 >.

BRITTO, C. et al. Conserved linkage groups associated with large-scale chromosomal rearrangements between Old World and New World< i> Leishmania</i> genomes. **Gene,** v. 222, n. 1, p. 107-117, 1998. ISSN 0378-1119.

BROHEE, S.; VAN HELDEN, J. Evaluation of clustering algorithms for protein-protein interaction networks. **Bmc Bioinformatics,** v. 7, 2006. ISSN 1471-2105. Disponível em: < <Go to ISI>://WOS:000242187800001 >.

CHEN, J. et al. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. **Amino Acids,** v. 33, n. 3, p. 423-428, 2007. ISSN 0939-4451.

CHEN, P. P. S. The entity-relationship model—toward a unified view of data. **ACM Transactions on Database Systems (TODS),** v. 1, n. 1, p. 9-36, 1976. ISSN 0362-5915.

CHUA, H. N.; SUNG, W. K.; WONG, L. Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. **Bioinformatics,** v. 22, n. 13, p. 1623-1630, Jul 2006. ISSN 1367-4803. Disponível em: < <Go to ISI>://000238905700012 >.

CHUANG, H. Y. et al. Network-based classification of breast cancer metastasis. **Molecular Systems Biology,** v. 3, Oct 2007. ISSN 1744-4292. Disponível em: < <Go to ISI>://000250700500001 >.

CLAYTON, C. E. Life without transcriptional control? From fly to man and back again. **The EMBO journal,** v. 21, n. 8, p. 1881, 2002.

CODD, E. F. A relational model of data for large shared data banks. **Communications of the ACM,** v. 13, n. 6, p. 377-387, 1970. ISSN 0001-0782.

COSTA, C. H. N. et al. Vaccines for the leishmaniases: proposals for a research agenda. **PLoS Neglected Tropical Diseases,** v. 5, n. 3, p. e943, 2011. ISSN 1935-2735.

CUI, T. et al. Uncovering new signaling proteins and potential drug targets through the interactome analysis of Mycobacterium tuberculosis. **Bmc**

Genomics, v. 10, Mar 19 2009. ISSN 1471-2164. Disponível em: < <Go to ISI>://WOS:000265791900003 >.

EDDY, S. R. Profile hidden Markov models. **Bioinformatics,** v. 14, n. 9, p. 755-763, 1998. ISSN 1367-4803. Disponível em: < <Go to ISI>://000077489900002 >.

EL-MANZALAWY, Y.; DOBBS, D.; HONAVAR, V. Predicting linear B-cell epitopes using string kernels. **Journal of Molecular Recognition,** v. 21, n. 4, Jul-Aug 2008. ISSN 0952-3499. Disponível em: < <Go to ISI>://WOS:000258006400006 >.

EL-SAYED, N. M. et al. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. **Science,** v. 309, n. 5733, p. 409-415, 2005. ISSN 0036-8075.

_____. Comparative genomics of trypanosomatid parasitic protozoa. **Science,** v. 309, n. 5733, p. 404-409, 2005. ISSN 0036-8075.

ELTABAKH, M. Y.; OUZZANI, M.; AREF, W. G. BDBMS--a database management system for biological data. **arXiv preprint cs/0612127**, 2006.

ENRIGHT, A. J. et al. Protein interaction maps for complete genomes based on gene fusion events. **Nature,** v. 402, n. 6757, p. 86-90, 1999. ISSN 0028-0836. Disponível em: < <Go to ISI>://WOS:000083638600048 >.

ENRIGHT, A. J.; VAN DONGEN, S.; OUZOUNIS, C. A. An efficient algorithm for large-scale detection of protein families. **Nucleic Acids Research,** v. 30, n. 7, p. 1575-1584, 2002. ISSN 0305-1048.

FERENC, J.; LIU, W. C.; MIKE, A. Trophic field overlap: A new approach to quantify keystone species. **Ecological Modelling,** v. 220, n. 21, p. 2899-2907, 2009. ISSN 0304-3800. Disponível em: < <Go to ISI>://WOS:000272332900009 >.

FINN, R. D. et al. The Pfam protein families database. **Nucleic Acids Research,** v. 38, p. D211-D222, Jan 2010. ISSN 0305-1048. Disponível em: < <Go to ISI>://000276399100032 >.

FLOREZ, A. F. et al. Protein network prediction and topological analysis in Leishmania major as a tool for drug target selection. **Bmc Bioinformatics,** v. 11, 2010. ISSN 1471-2105. Disponível em: < <Go to ISI>://WOS:000283062500001 >.

FRASER, H. B. et al. Evolutionary rate in the protein interaction network. **Science,** v. 296, n. 5568, p. 750-752, Apr 26 2002. ISSN 0036-8075. Disponível em: < <Go to ISI>://WOS:000175281700060 >.

FREEMAN, L. C. SET OF MEASURES OF CENTRALITY BASED ON BETWEENNESS. **Sociometry,** v. 40, n. 1, p. 35-41, 1977. Disponível em: < <Go to ISI>://WOS:A1977CZ20900004 >.

GARLICK, P. J. et al. STIMULATION OF PROTEIN-SYNTHESIS AND BREAKDOWN BY VACCINATION. **British Medical Journal,** v. 281, n. 6235, 1980 1980. ISSN 0959-8138. Disponível em: < <Go to ISI>://WOS:A1980KA41100007 >.

GOMEZ, S. M.; NOBLE, W. S.; RZHETSKY, A. Learning to predict protein-protein interactions from protein sequences. **Bioinformatics,** v. 19, n. 15, Oct 12 2003. ISSN 1367-4803. Disponível em: < <Go to ISI>://WOS:000186179000003 >.

GONG, S. et al. PSIbase: A database of Protein Structural Interactome map (PSIMAP). **Bioinformatics,** v. 21, n. 10, p. 2541-2543, 2005. ISSN 1367-4803. Disponível em: < <Go to ISI>://WOS:000229285600058 >.

HAAG, J.; O'HUIGIN, C.; OVERATH, P. The molecular phylogeny of trypanosomes: evidence for an early divergence of the Salivaria. **Molecular and biochemical parasitology,** v. 91, n. 1, p. 37-49, 1998. ISSN 0166-6851.

HANNAERT, V.; OPPERDOES, F. R.; MICHELS, P. A. M. Comparison and evolutionary analysis of the glycosomal glyceraldehyde-3-phosphate dehydrogenase from different Kinetoplastida. **Journal of Molecular Evolution,** v. 47, n. 6, p. 728-738, 1998. ISSN 0022-2844.

HARRINGTON, E. D.; JENSEN, L. J.; BORK, P. Predicting biological networks from genomic data. **Febs Letters,** v. 582, n. 8, p. 1251-1258, 2008. ISSN 0014-5793.

HARTWELL, L. H. et al. From molecular to modular cell biology. **Nature,** v. 402, n. 6761, p. C47-C52, Dec 1999. ISSN 0028-0836. Disponível em: < <Go to ISI>://000084014100007 >.

HE, F. et al. The prediction of protein-protein interaction networks in rice blast fungus. **Bmc Genomics,** v. 9, 2008. ISSN 1471-2164. Disponível em: < <Go to ISI>://WOS:000262193300001 >.

HELLER, K. Shane Crotty: Exploring immune memory. **The Journal of Experimental Medicine,** v. 206, n. 5, p. 974-975, 2009. ISSN 0022-1007.

HERWALDT, B. L. Leishmaniasis. **Lancet,** v. 354, n. 9185, p. 1191-1199, Oct 1999. ISSN 0140-6736. Disponível em: < <Go to ISI>://000082954100044 >.

HISHIGAKI, H. et al. Assessment of prediction accuracy of protein function from protein–protein interaction data. **Yeast,** v. 18, n. 6, p. 523-531, 2001. ISSN 1097-0061.

HOTEZ, P. J. et al. The Neglected Tropical Diseases of Latin America and the Caribbean: A Review of Disease Burden and Distribution and a Roadmap for Control and Elimination. **Plos Neglected Tropical Diseases,** v. 2, n. 9, Sep 2008. ISSN 1935-2735. Disponível em: < <Go to ISI>://WOS:000261807500004 >.

HU, J.; MAHMOUD, M. I.; EL-FAKAHANY, E. E. Polyamines inhibit nitric oxide synthase in rat cerebellum. **Neuroscience letters,** v. 175, n. 1, p. 41-45, 1994. ISSN 0304-3940.

HUGHEY, R.; KROGH, A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. **Computer applications in the biosciences: CABIOS,** v. 12, n. 2, p. 95-107, 1996. ISSN 1367-4803.

HUYNEN, M. A.; BORK, P. Measuring genome evolution. **Proceedings of the National Academy of Sciences of the United States of America,** v. 95, n. 11, p. 5849-5856, May 1998. ISSN 0027-8424. Disponível em: < <Go to ISI>://000073852600004 >.

IVENS, A. C. et al. The genome of the kinetoplastid parasite, Leishmania major. **Science,** v. 309, n. 5733, p. 436-442, 2005. ISSN 0036-8075.

JANSEN, R.; GERSTEIN, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. **Current Opinion in Microbiology,** v. 7, n. 5, Oct 2004. ISSN 1369-5274. Disponível em: < <Go to ISI>://WOS:000224575700015 >.

JANSEN, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. **Science,** v. 302, n. 5644, Oct 17 2003. ISSN 0036-8075. Disponível em: < <Go to ISI>://WOS:000185963200044 >.

JEONG, H. et al. Lethality and centrality in protein networks. **Nature,** v. 411, n. 6833, p. 41-42, 2001. ISSN 0028-0836. Disponível em: < <Go to ISI>://WOS:000168432800033 >.

JORDÁN, F.; LIU, W.-C.; VAN VEEN, F. J. F. Quantifying the importance of species and their interactions in a host-parasitoid community. **Community Ecology,** v. 4, n. 1, p. 9, 2003.

KAMBUROV, A. et al. Cluster-based assessment of protein-protein interaction confidence. **BMC bioinformatics,** v. 13, n. 1, p. 262, 2012. ISSN 1471-2105.

KATOH, K. et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic Acids Research,** v. 30, n. 14, p. 8, 2002.

KIM, J.-G. et al. Predicting the Interactome of Xanthomonas oryzae pathovar oryzae for target selection and DB service. **Bmc Bioinformatics,** v. 9, Jan 24 2008. ISSN 1471-2105. Disponível em: < <Go to ISI>://WOS:000253686300001 >.

KING, A. D.; PRZULJ, N.; JURISICA, I. Protein complex prediction via cost-based clustering. **Bioinformatics,** v. 20, n. 17, p. 3013-3020, Nov 2004. ISSN 1367-4803. Disponível em: < <Go to ISI>://000225361400015 >.

KOLEV, N. G. et al. The transcriptome of the human pathogen Trypanosoma brucei at single-nucleotide resolution. **PLoS pathogens,** v. 6, n. 9, p. e1001090, 2010. ISSN 1553-7374.

KORBER, B.; LABUTE, M.; YUSIM, K. Immunoinformatics comes of age. **PLoS computational biology,** v. 2, n. 6, p. e71, 2006. ISSN 1553-7358.

KRYLOV, D. M. et al. Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. **Genome Research,** v. 13, n. 10, p. 2229-2235, 2003. ISSN 1088-9051.

LACOUNT, D. J. et al. A protein interaction network of the malaria parasite Plasmodium falciparum. **Nature,** v. 438, n. 7064, p. 103-107, 2005. ISSN 0028-0836. Disponível em: < <Go to ISI>://WOS:000232979000050 >.

LAINSON, R.; WARD, R. D.; SHAW, J. J. Leishmania in phlebotomid sandflies: VI. Importance of hindgut development in distinguishing between parasites of the Leishmania mexicana and L. braziliensis complexes. **Proceedings of the Royal Society of London. Series B. Biological Sciences,** v. 199, n. 1135, p. 309-320, 1977. ISSN 0962-8452.

LARSEN, J. E. P.; LUND, O.; NIELSEN, M. Improved method for predicting linear B-cell epitopes. **Immunome research,** v. 2, n. 1, p. 2, 2006. ISSN 1745-7580.

LARSEN, M. V. et al. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. **Bmc Bioinformatics,** v. 8, Oct 31 2007. ISSN 1471-2105. Disponível em: < <Go to ISI>://WOS:000252550400001 >.

LEE, I. et al. A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. **Nature Genetics,** v. 40, n. 2, p. 181-188, Feb 2008. ISSN 1061-4036. Disponível em: < <Go to ISI>://000252732900016 >.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. **Genome Research,** v. 13, n. 9, p. 2178-2189, 2003. ISSN 1088-9051.

LIANG, X. et al. trans and cis splicing in trypanosomatids: Mechanism, factors, and regulation. **Eukaryotic cell,** v. 2, n. 5, p. 830-840, 2003. ISSN 1535-9778.

LIN, C. Y. et al. Hubba: hub objects analyzer - a framework of interactome hubs identification for network biology. **Nucleic Acids Research,** v. 36, p. W438-W443, 2008. ISSN 0305-1048. Disponível em: < <Go to ISI>://WOS:000258142300081 >.

LIRA, R. et al. Evidence that the high incidence of treatment failures in Indian kala-azar is due to the emergence of antimony-resistant strains of Leishmania donovani. **Journal of Infectious Diseases,** v. 180, n. 2, p. 564-567, Aug 1999. ISSN 0022-1899. Disponível em: < <Go to ISI>://WOS:000081767000049 >.

LUKEŠ, J. et al. Analysis of ribosomal RNA genes suggests that trypanosomes are monophyletic. **Journal of molecular evolution,** v. 44, n. 5, p. 521-527, 1997. ISSN 0022-2844.

LUNDEGAARD, C. et al. Modeling the adaptive immune system: predictions and simulations. **Bioinformatics,** v. 23, n. 24, p. 3265-3275, 2007. ISSN 1367-4803.

LYNN, M. A.; MCMASTER, W. R. < i> Leishmania</i>: conserved evolution–diverse diseases. **Trends in parasitology,** v. 24, n. 3, p. 103-105, 2008. ISSN 1471-4922.

MAMITSUKA, H. Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models. **Proteins Structure Function and Genetics,** v. 33, n. 4, p. 460-474, 1998. ISSN 0887-3585.

MANI, K. M. et al. A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. **Molecular Systems Biology,** v. 4, Feb 2008. ISSN 1744-4292. Disponível em: < <Go to ISI>://000253761300007 >.

MARCOTTE, E. M. et al. Detecting protein function and protein-protein interactions from genome sequences. **Science,** v. 285, n. 5428, p. 751-753, 1999. ISSN 0036-8075. Disponível em: < <Go to ISI>://WOS:000081765100059 >.

MARTÍNEZ-CALVILLO, S. et al. Transcription initiation and termination on Leishmania major chromosome 3. **Eukaryotic cell,** v. 3, n. 2, p. 506-517, 2004. ISSN 1535-9778.

MASLOV, S.; SNEPPEN, K. Specificity and stability in topology of protein networks. **Science,** v. 296, n. 5569, p. 910-913, May 2002. ISSN 0036-8075. Disponível em: < <Go to ISI>://000175442500044 >.

MATTHEWS, L. R. et al. Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". **Genome research,** v. 11, n. 12, p. 2120-2126, 2001. ISSN 1088-9051.

MULEY, V. Y.; RANJAN, A. Effect of Reference Genome Selection on the Performance of Computational Methods for Genome-Wide Protein-Protein Interaction Prediction. **PloS one,** v. 7, n. 7, p. e42057, 2012. ISSN 1932-6203.

MURRAY, H. W. et al. Advances in leishmaniasis. **Lancet,** v. 366, p. 17, 2005.

NADIM, A. et al. Effectiveness of leishmanization in the control of cutaneous leishmaniasis. **Bulletin de la Société de Pathologie Exotique et de ses Filiales,** v. 76, n. 4, p. 377, 1983. ISSN 0037-9085.

NEI, M.; LI, W. H. MATHEMATICAL-MODEL FOR STUDYING GENETIC-VARIATION IN TERMS OF RESTRICTION ENDONUCLEASES. **Proceedings of the National**

Academy of Sciences of the United States of America, v. 76, n. 10, p. 5269-5273, 1979 1979. ISSN 0027-8424. Disponível em: < <Go to ISI>://WOS:A1979HR46000110 >.

NIELSEN, M.; LUND, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. **BMC bioinformatics,** v. 10, n. 1, p. 296, 2009. ISSN 1471-2105.

NIELSEN, M. et al. Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. **Bioinformatics,** v. 20, n. 9, p. 1388-1397, 2004. ISSN 1367-4803.

_____. Reliable prediction of T ‐ cell epitopes using neural networks with novel sequence representations. **Protein Science,** v. 12, n. 5, p. 1007-1017, 2009. ISSN 1469-896X.

NILSSON, D. et al. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. **PLoS pathogens,** v. 6, n. 8, p. e1001037, 2010. ISSN 1553-7374.

NOAZIN, S. et al. First generation leishmaniasis vaccines: A review of field efficacy trials. **Vaccine,** v. 26, n. 52, p. 6759-6767, 2008. ISSN 0264-410X. Disponível em: < <Go to ISI>://WOS:000261898400008 >.

OSTELL, J. M.; KANS, J. A. The NCBI data model. **Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, Volume 39**, p. 121-144, 2006. ISSN 0470110600.

PASTOR-SATORRAS, R.; SMITH, E.; SOLE, R. V. Evolving protein interaction networks through gene duplication. **Journal of Theoretical Biology,** v. 222, n. 2, p. 199-210, May 2003. ISSN 0022-5193. Disponível em: < <Go to ISI>://000182894200006 >.

PAWŁOWSKI, K. Uncharacterized/hypothetical proteins in biomedical 'omics' experiments: is novelty being swept under the carpet? **Briefings in functional genomics & proteomics,** v. 7, n. 4, p. 283-290, 2008. ISSN 2041-2649.

PEACOCK, C. S. et al. Comparative genomic analysis of three Leishmania species that cause diverse human disease. **Nature Genetics,** v. 39, n. 7, p. 839-847, 2007. ISSN 1061-4036.

PELLEGRINI, M. et al. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. **Proceedings of the National Academy of Sciences of the United States of America,** v. 96, n. 8, p. 4285-4288, 1999. ISSN 0027-8424. Disponível em: < <Go to ISI>://WOS:000079766500017 >.

PEREIRA-LEAL, J. B.; ENRIGHT, A. J.; OUZOUNIS, C. A. Detection of functional modules from protein interaction networks. **Proteins-Structure Function and**

Genetics, v. 54, n. 1, p. 49-57, Jan 2004. ISSN 0887-3585. Disponível em: < <Go to ISI>://000187806400006 >.

PETERS, B. et al. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. **The Journal of Immunology,** v. 171, n. 4, p. 1741-1749, 2003. ISSN 0022-1767.

PISCOPO, T. V.; MALLIA, A. C. Leishmaniasis. **Postgraduate Medical Journal,** v. 82, n. 972, p. 649-657, 2006. ISSN 0032-5473. Disponível em: < <Go to ISI>://WOS:000241628600006 >.

PIZZA, M. et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. **Science,** v. 287, n. 5459, p. 1816-1820, 2000. ISSN 0036-8075.

PUJANA, M. A. et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. **Nature Genetics,** v. 39, n. 11, p. 1338-1349, Nov 2007. ISSN 1061-4036. Disponível em: < <Go to ISI>://000250575900018 >.

PUJOL, A. et al. Unveiling the role of network and systems biology in drug discovery. **Trends in Pharmacological Sciences,** v. 31, n. 3, p. 115-123, 2010. ISSN 0165-6147. Disponível em: < <Go to ISI>://WOS:000276136500004 >.

PULENDRAN, B.; LI, S.; NAKAYA, H. I. Systems vaccinology. **Immunity,** v. 33, n. 4, p. 516-529, 2010. ISSN 1074-7613.

QI, Y. J.; KLEIN-SEETHARAMAN, J.; BAR-JOSEPH, Z. Random forest similarity for protein-protein interaction prediction from multiple sources. **Pacific Symposium on Biocomputing 2005**, 2005 2005. Disponível em: < <Go to ISI>://WOS:000230169100044 >.

RAGHAVACHARI, B. et al. DOMINE: a database of protein domain interactions. **Nucleic Acids Research,** v. 36, p. D656-D661, 2008. ISSN 0305-1048. Disponível em: < <Go to ISI>://WOS:000252545400118 >.

RAVASZ, E. et al. Hierarchical organization of modularity in metabolic networks. **Science,** v. 297, n. 5586, p. 1551-1555, 2002. ISSN 0036-8075. Disponível em: < <Go to ISI>://WOS:000177697300057 >.

RAYMOND, F. et al. Genome sequencing of the lizard parasite Leishmania tarentolae reveals loss of genes associated to the intracellular stage of human pathogenic species. **Nucleic Acids Research,** v. 40, n. 3, p. 1131-1147, Feb 2012. ISSN 0305-1048. Disponível em: < <Go to ISI>://WOS:000300422400025 >.

SATO, T. et al. The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. **Bioinformatics,** v. 21, n. 17, p. 3482-3489, 2005. ISSN 1367-4803.

SAUER, U.; HEINEMANN, M.; ZAMBONI, N. Genetics - Getting closer to the whole picture. **Science,** v. 316, n. 5824, p. 550-551, 2007. ISSN 0036-8075.

SEBATJANE, S. I. et al. *In vitro* and *in vivo* evaluation of five low molecular weight proteins of *Ehrlichia ruminantium* as potential vaccine components. **Veterinary immunology and immunopathology,** v. 137, n. 3, p. 217-225, 2010. ISSN 0165-2427.

SF, A. et al. Basic Local Alignment Tool. **Journal of Molecular Biology,** v. 5, n. 215, p. 7, 1990.

SHANNON, P. et al. Cytoscape: A software environment for integrated models of biomolecular interaction networks. **Genome Research,** v. 13, n. 11, p. 2498-2504, Nov 2003. ISSN 1088-9051. Disponível em: < <Go to ISI>://000186357000016 >.

SHARAN, R. et al. Conserved patterns of protein interaction in multiple species. **Proceedings of the National Academy of Sciences of the United States of America,** v. 102, n. 6, p. 1974-1979, Feb 2005. ISSN 0027-8424. Disponível em: < <Go to ISI>://000227072900033 >.

SHEKHAR, S.; CHAWLA, S. Spatial databases: a tour. **Upper Saddle River, New Jersey,** v. 7458, 2003.

SIEGEL, T. N. et al. Genome-wide analysis of mRNA abundance in two life-cycle stages of Trypanosoma brucei and identification of splicing and polyadenylation sites. **Nucleic acids research,** v. 38, n. 15, p. 4946-4957, 2010. ISSN 0305-1048.

SIMPSON, A. G. B. et al. Early evolution within kinetoplastids (Euglenozoa), and the late emergence of trypanosomatids. **Protist,** v. 155, n. 4, p. 407-422, 2004. ISSN 1434-4610.

SIMPSON, A. G. B.; STEVENS, J. R.; LUKES, J. The evolution and diversity of kinetoplastid flagellates. **Trends in Parasitology,** v. 22, n. 4, p. 168-174, 2006. ISSN 1471-4922. Disponível em: < <Go to ISI>://WOS:000236873000008 >.

SING, T. et al. ROCR: visualizing classifier performance in R. **Bioinformatics,** v. 21, n. 20, p. 3940-3941, 2005. ISSN 1367-4803.

SKRABANEK, L. et al. Computational prediction of protein-protein interactions. **Molecular Biotechnology,** v. 38, n. 1, p. 1-17, 2008. ISSN 1073-6085. Disponível em: < <Go to ISI>://WOS:000251796300001 >.

SMOOT, M. E. et al. Cytoscape 2.8: new features for data integration and network visualization. **Bioinformatics,** v. 27, n. 3, p. 431-432, Feb 2011. ISSN 1367-4803. Disponível em: < <Go to ISI>://000286991300023 >.

STEVENS, J. R. et al. The molecular evolution of Trypanosomatidae. **Advances in Parasitology,** v. 48, p. 1-53, 2001. ISSN 0065-308X.

SUNDAR, S. Drug resistance in Indian visceral leishmaniasis. **Tropical Medicine & International Health,** v. 6, n. 11, p. 849-854, 2002. ISSN 1365-3156.

TEIXEIRA, S. M. et al. Trypanosomatid comparative genomics: contributions to the study of parasite biology and different parasitic diseases. **Genetics and Molecular Biology,** v. 35, n. 1, p. 1-17, 2012. ISSN 1415-4757.

TITZ, B.; SCHLESNER, M.; UETZ, P. What do we learn from high-throughput protein interaction data? **Expert review of proteomics,** v. 1, n. 1, p. 111-121, 2004. ISSN 1478-9450.

VAPNIK, V.; VASHIST, A. A new learning paradigm: learning using privileged information. **Neural networks: the official journal of the International Neural Network Society,** v. 22, n. 5-6, p. 544, 2009. ISSN 0893-6080.

VILELLA, A. J. et al. VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. **Bioinformatics,** v. 21, n. 11, p. 2791-2793, Jun 1 2005. ISSN 1367-4803. Disponível em: < <Go to ISI>://WOS:000229441500033 >.

VILLA, H. et al. Molecular and functional characterization of adenylate kinase 2 gene from Leishmania donovani. **European Journal of Biochemistry,** v. 270, n. 21, p. 4339-4347, 2003. ISSN 1432-1033.

VITA, R. et al. The immune epitope database 2.0. **Nucleic acids research,** v. 38, n. suppl 1, p. D854-D862, 2010. ISSN 0305-1048.

VON MERING, C. et al. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. **Nucleic Acids Research,** v. 33, p. D433-D437, Jan 1 2005. ISSN 0305-1048. Disponível em: < <Go to ISI>://WOS:000226524300089 >.

WAN, C. Y.; WILKINS, T. A. Spermidine facilitates PCR amplification of target DNA. **Genome Research,** v. 3, n. 3, p. 208-210, 1993. ISSN 1088-9051.

WANG, M. et al. CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening. **Vaccine,** v. 25, n. 15, p. 2823-2831, 2007. ISSN 0264-410X.

WINCKER, P. et al. The Leishmania genome comprises 36 chromosomes conserved across widely divergent human pathogenic species. **Nucleic acids research,** v. 24, n. 9, p. 1688-1694, 1996. ISSN 0305-1048.

WU, S.; FLACH, P. A scored AUC metric for classifier evaluation and selection. Proceedings of the ICML 2005 Workshop on ROC Analysis in Machine Learning (International Machine Learning Society, Bonn, Germany, 2005), 2005.

WUCHTY, S. Evolution and topology in the yeast protein interaction network. **Genome Research,** v. 14, n. 7, p. 1310-1314, Jul 2004. ISSN 1088-9051. Disponível em: < <Go to ISI>://WOS:000222434200011 >.

_____. Topology and weights in a protein domain interaction network - a novel way to predict protein interactions. **Bmc Genomics,** v. 7, May 2006. ISSN 1471-2164. Disponível em: < <Go to ISI>://000239343900001 >.

XENARIOS, I. et al. DIP: the Database of Interacting Proteins. **Nucleic Acids Research,** v. 28, n. 1, Jan 1 2000. ISSN 0305-1048. Disponível em: < <Go to ISI>://WOS:000084896300083 >.

YASSER, E. L. M.; DOBBS, D.; HONAVAR, V. Predicting flexible length linear B-cell epitopes. Computational Systems Bioinformatics, 2008, NIH Public Access. p.121.

YU, H. Y. et al. Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. **Genome Research,** v. 14, n. 6, Jun 2004. ISSN 1088-9051. Disponível em: < <Go to ISI>://WOS:000221852400013 >.

ZHANG, L. V. et al. Predicting co-complexed protein pairs using genomic and proteomic data integration. **Bmc Bioinformatics,** v. 5, Apr 16 2004. ISSN 1471-2105. Disponível em: < <Go to ISI>://WOS:000221697900001 >.

ZHOU, Q. et al. A gene regulatory network in mouse embryonic stem cells. **Proceedings of the National Academy of Sciences,** v. 104, n. 42, p. 16438-16443, 2007. ISSN 0027-8424.

9 – Anexos

9.1 – Anexo I

Anexo I – Artigos Aceitos Para Publicação

ARTIGO 1

Date: Oct 31 2012 1:31AM
To: "Jeronimo C. Ruiz" jeronimo@cpqrr.fiocruz.br,jero.ruiz@gmail.com
From: "PLOS ONE" plosone@plos.org
Subject: PLOS ONE Decision: Accept [PONED1216405R2]

PONED1216405R2
Computational prediction of protein-protein interactions in Leishmania predicted proteomes
PLOS ONE

Dear Dr Ruiz,

I am pleased to inform you that your manuscript has been deemed suitable for publication in PLOS ONE. Your manuscript will now be passed on to our Production staff, who will check your files for correct formatting and completeness. After this review, they may return your manuscript to you so that you can make necessary alterations and upload a final version. Before uploading, you should check the PDF of your manuscript very closely. THERE IS NO AUTHOR PROOFING. You should therefore consider the corrected files you upload now as equivalent to a production proof. The text you supply at this point will be faithfully represented in your published manuscript exactly as you supply it. This is your last opportunity to correct any errors that are present in your manuscript files. If you or your institution will be preparing press materials for this manuscript, you must inform our press team in advance. Please contact them at ONEpress@plos.org. If you have any questions, concerns, or problems, please contact us at plosone@plos.org, and thank you for submitting your work to our journal.

With kind regards,
John Parkinson
Academic Editor
PLOS ONE

[**NOTE:** If reviewer comments were submitted as an attachment file, they will be accessible only via the submission site. Please log into your account, locate the manuscript record, and check for the action link "View Attachments". If this link does not appear, there are no attachment files to be viewed.]

Research article: "Computational prediction of protein-protein interactions in *Leishmania* predicted proteomes"

Antonio M. Rezende[1,2*], Edson L. Folador[1,3], Daniela de M. Resende[1,4], Jeronimo C. Ruiz[1]*

[1]Laboratório de Parasitologia Celular e Molecular, Centro de Pesquisa René Rachou – FIOCRUZ, Belo Horizonte, Minas Gerais, Brazil

[2]Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Minas Gerais, Brazil

[3]Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, Rio de Janeiro, Brazil.

[4]Laboratório de Pesquisas Clínicas, Universidade Federal de Ouro Preto, Ouro Preto, Minas Gerais , Brazil.

.

*Corresponding author

Email adresses:

AMR: am.rezende@cpqrr.fiocruz.br

JCR: jeronimo@cpqrr.fiocruz.br, jero.ruiz@gmail.com

ELF: edson.folador@cpqrr.fiocruz.br

DMR: dmresende@ef.ufop.br; dani.melo.resende@gmail.com

Abstract

The Trypanosomatids parasites *Leishmania braziliensis*, *Leishmania major* and *Leishmania infantum* are important human pathogens. Despite of years of study and genome availability, effective vaccine has not been developed yet, and the chemotherapy is highly toxic. Therefore, it is clear just interdisciplinary integrated studies will have success in trying to search new targets for developing of vaccines and drugs.

An essential part of this rationale is related to protein-protein interaction network (PPI) study which can provide a better understanding of complex protein interactions in biological system.

Thus, we modeled PPIs for Trypanosomatids through computational methods using sequence comparison against public database of protein or domain interaction for interaction prediction (Interolog Mapping) and developed a dedicated combined system score to address the predictions robustness.

The confidence evaluation of network prediction approach was addressed using gold standard positive and negative datasets and the AUC value obtained was 0.94.

As result, 39,420, 43,531 and 45,235 interactions were predicted for *L. braziliensis*, *L. major* and *L. infantum* respectively. For each predicted network the top 20 proteins were ranked by MCC topological index. In addition, information related with immunological potential, degree of protein sequence conservation among orthologs and degree of identity compared to proteins of potential parasite hosts was integrated. This information integration provides a better understanding and usefulness of the predicted networks that can be valuable to select new potential biological targets for drug and vaccine development.

Network modularity which is a key when one is interested in destabilizing the PPIs for drug or vaccine purposes along with multiple alignments of the predicted PPIs were performed revealing patterns associated with protein turnover.

In addition, around 50% of hypothetical protein present in the networks received some degree of functional annotation which represents an important contribution since approximately 60% of *Leishmania* predicted proteomes has no predicted function.

Introduction

According to the World Health Organization (www.who.int), there are roughly 12 million people infected with parasites from the *Leishmania* genus, which can cause visceral, cutaneous, or mucosal leishmaniasis [1], with an annual incidence from one to two million. Leishmaniasis is considered a neglected tropical disease responsible for a high estimated burden in Latin America [2].

For case control and the treatment of leishmaniasis, the major drugs used are either expensive, toxic, or both, and frequently require long periods of supervised therapy [2]. In addition, the pentavalent antimony based drugs that are the major chemical compounds used for leishmaniasis treatment have many side effects, such as pain, erythema, edema, abdominal pain, nausea, thrombocytopenia or leucopenia, and cardiotoxicity [1]. Furthermore, many reports of parasite resistance have been published [3-6]. It is worth mentioning that there are other medicines against leishmaniasis, but some of them are not economically feasible for many endemic countries [1].

To aggravate this situation, there are no effective vaccines for leishmaniasis. Despite abundant clinical and experimental evidence suggesting that leishmaniasis can be prevented by vaccination, the only proven vaccine agent in human beings is live *Leishmania major*, and it is discontinued because of unacceptable lesions in some recipients [1].

Therefore, based on the facts cited above, the necessity to develop new drugs and vaccine approaches is apparent. In order to reach this goal, new targets should be evaluated and the choice and evaluation method should consider the many different aspects of the complex biology of the agents of leishmaniasis. This challenging task can be achieved by integrating different data sets (e.g.; genome, transcriptome, proteome) in a systemic approach. Currently, this biology-based interdisciplinary approach focusing on the study of complex interactions in the biological system is called Systems Biology [7].

One of the main branches of Systems Biology refers to network studies. Here, there are different types of networks: protein-protein interaction network, metabolic

network, regulatory network, etc. These networks can provide valuable information about different characteristics of an organism. More specifically, on a protein-protein interaction network (PPI) it represents a set of proteins of an organism, and how they interact with each other [8]. Moreover, the PPIs are undirected networks, in general are scale-free [9] and modular [10].

Currently, there are many different experimental methods to predict a PPI, among them we have the yeast two-hybrid method and affinity purification coupled with mass spectrometry [8]. Nevertheless, they may not be feasible for all proteins for all organisms, and they are susceptible to systematic errors. Thus, a number of computational approaches have been developed to predict protein-protein interactions based on protein or nucleotide sequence in large-scale [11]. Some of the computational approaches most known are the Phylogenetic Profile [12,13], Genome Neighborhood [8], Gene Fusion [13,14], Sequence Co-evolution [15], and comparison against the interaction public database or Interolog Mapping [16-19].

In this work, the Interolog Mapping method was used. Specifically on this approach, it assumes that if two proteins have a great sequence similarity against two proteins from a public database, and these latter ones interact, then the former ones interact too.

Therefore, the main point of this work is to predict a PPI network for each one of the target organisms, *Leishmania braziliensis*, *Leishmania major* and *Leishmania infantum*. Ultimately, we intend to use these networks to identify proteins and protein interactions that can be used as new targets for drugs and vaccines development.

Methodology

1 – Evaluation of PPI Prediction Approach

In order to evaluate the confidence of our network prediction methodology and consequently predict PPI networks for *Leishmania* sp, a performance evaluation was conducted.

The gold standard positive dataset was extracted from DIP (Database of Interacting Proteins) [7,20]. The DIP database contains experimentally determined interactions between proteins, integrates information from many sources and is manually curated by experts. Given the information consistency of this database and taking into account the amount of information concerning PPI networks, *E.coli* was selected for the performance evaluation. Regarding the specific selection of positive pairs, we considered some points addressed on a recent work of Muley and Ranjan [21]. In this context, 702 interactions were selected as positive gold standard dataset.

The negative standard dataset used for the performance evaluation was built based on the works of [22-26]. In summary, considering all possible interactions in the model organism and subtracting the experimentally validated ones, a random selection was performed and only pairs containing proteins located in different subcellular localizations were maintained. A ratio of 1:5 between positive and negative interaction pairs was used resulting into 3,510 negative interactions.

Using these gold standards datasets and the model organism we could identify the true positive (TP) or true negative (TN) protein pairs predicted by our network prediction methodology. The properly performance evaluation was made using ROC (Receiver Operating Characteristic) curves using the ROCR package for R (http://www.r-project.org/) [27]. A ROC curve is a plot of the False Positive Rate (FPR) against the True Positive Rate (TPR or sensitivity) for a given approach prediction. A random prediction will give value of 0.5 for the area under the ROC curve (or AUC) and a perfect prediction method would have an AUC value equal to 1 [21].

2 – Data Filtering

Before starting with the network prediction, a filtering was performed on the predicted parasite proteomes to remove possible annotation errors. The proteome versions utilized here were version 2, final version, and version 3 for *L. braziliensis, L. major* and *L. infantum*, respectively. The following three criteria were utilized in this filtering. First, protein sequences should start with the methionine amino acid. Second, protein sequences should not have illegal characters such as X, B, Z, U, and "*" that are ambiguous or do not represent any of the 20 amino acids. Third, they should be bigger than 100 amino acids.

3 – Predictions of Protein-Protein Interaction Pairs

To predict the protein-protein interaction pairs for the three organisms (*L. braziliensis*, *L. major* and *L. infantum*), the Interolog Mapping method was used. To apply the approach, we used four public databases namely: Domine [28], PSI-Base [29], IntAct [30], and String [31]. Here, it is worth mentioning the String database are not limited to direct, physical interactions between two proteins. Indirect interactions between proteins also exist such as genetic and metabolic interaction. Nevertheless, according to the last work describing the String database [32], most association currently can not be specified with much precision in terms of their mode of interaction. Thus the fundamental unit stored in String is the 'functional association'. In addition, the String has flat files which have some degree of description about the interactions. If we consider in these files the term "binding" as physical interaction, we obtain nearly 94% of all interactions present in String. Besides, the other terms present in these files do not guarantee that the interactions are not physical interaction. Therefore, the impact of indirect interactions in our networks is minor. The first step here was to download all the interactions and all the protein sequences present in those databases. After that, the sequences from the predicted proteomes of the three protozoa were compared against the protein sequences from the databases and vice versa. To perform this comparison, we used the *blastp* from the BLAST software package [33] for searching sequences from PSI-Base, IntAct, and String. For the Domine database, the sequence comparisons were made by *hmmpfam* (sequence against HMM) and *hmmsearch* (HMM against sequence) from the HMMER software

package [34], since the Domine uses the HMMs (Hidden Markov Models) present in the PFAM database [35] to describe its proteins.

Therefore, a protein *"X*" from a database is only considered as a homolog to protein *"A"* from one of the three organisms if protein *"X"* is the best hit for protein *"A"*, and protein *"A"* is the best hit for protein *"X"*. This is called the Best Bidirectional Hit (BBH). For each BBH, several measures were extracted. When a BBH came from *blastp* result, we extracted from it the minimum identity, minimum similarity and minimum alignment score between two sequences. In addition, the alignment coverage was extracted. When a BBH came from HMMER software, we extracted just minimum alignment score. In summary, the following formulas were applied:

$$identity(AX) = (\min\{\max_{i,\ldots,k} identity((A \leftarrow X)_i), \max_{j,\ldots,l} identity((A \rightarrow X)_j)\})$$

$$similarity(AX) = (\min\{\max_{i,\ldots,k} similarity((A \leftarrow X)_i), \max_{j,\ldots,l} similarity((A \rightarrow X)_j)\})$$

$$coverage(AX) = (\min\{\max_{i,\ldots,k} coverage((A \leftarrow X)_i), \max_{j,\ldots,l} coverage((A \rightarrow X)_j)\})$$

$$alignScore(AX) = (\min\{\max_{i,\ldots,k} score((A \leftarrow X)_i), \max_{j,\ldots,l} score((A \rightarrow X)_j)\})$$

Here, *"A"* is a protein from a target organism, *"X"* is a protein from a database, *k* is the number of results from the comparison made using *"X"* as a query and *"l"* is the number of results from the comparison made using *"A"* as a query. From each comparison, the maximum values of each measure were taken. Afterward, just the minimum values from the two comparisons were used further. In addition, these measures were calculated for each database if the *e-value* for each comparison was smaller than 10^{-85} for String and IntAct results, 10^{-45} for Domine and 10^{-10} for PSI-Base.

We then mapped the interactions present in the databases on the three proteomes. To do that, we firstly knew that "*X"* and "*Y"*, which are proteins from a database, interact. Second, we knew that "*X"* was the BBH for "*A",* which is a protein from a target organism, and "*Y"* was the BBH for "*B",* which is also a protein from the same target organism of *"A"* protein. Hence, we assumed that *A* and *B* interact.

In general each database has a confidence score for its interactions, thus these scores were used to compose the final combined score. In our case, we were not able to find this kind of score for PSI-Base repository. In the end, for each database, the score for each prediction was calculated according to the followings:

$$score_{STRING} = \left(\frac{\left(\frac{similarity(AX) + similarity(BY)}{2} \right) + \left(\frac{identity(AX) + identity(BY)}{2} \right) + \left(\frac{coverage(AX) + coverage(BY)}{2} \right)}{3} \right) \times scoreString(X,Y)$$

$$score_{IntAct} = \left(\frac{\left(\frac{similarity(AX) + similarity(BY)}{2} \right) + \left(\frac{identity(AX) + identity(BY)}{2} \right) + \left(\frac{coverage(AX) + coverage(BY)}{2} \right)}{3} \right) \times scoreIntAct(X,Y)$$

$$score_{Domine} = \sqrt{alignScore(AX) \times alignScore(BY)} \times scoreDomine(X,Y)$$

$$score_{PSIBase} = \left(\frac{\left(\frac{(similarity(AX) + similarity(BY))}{2} + \frac{(identity(AX) + identity(BY))}{2} + \frac{(coverage(AX) + coverage(BY))}{2} \right)}{3} \right) \times \sqrt{alignScore(AX) \times alignScore(BY}$$

4 – Calculating Confidence Score for Protein-Protein Interaction

In order to attribute a confidence score for the predicted interactions, we adopted the same rational described by [16,36] and built a dedicated interaction combined score for our methodology. This combined score takes into account the prediction scores obtained from Domine, PSI-Base, IntAct and String Databases from each protein interaction and is calculated according to the formula shown below:

$$score_{comb(AB)} = 1 - \prod_{i \in E} (1 - S_i),$$

where *score_comb(AB)* is the combined score for the interaction between proteins "*A"* and "*B"*, *E* is all the methods that were used to predict the interactions, and S_i is the score normalized by the biggest value calculated for the method *i*.

Many observed networks fall into the class of scale-free networks, meaning that they have power-law (or scale-free) degree distributions and this does not occur with random networks. Thus, after the calculation of *score_comb*, the three predicted PPIs were tested against the scale-free model for PPIs suggested by Barabasi and Oltvai [37] and the hierarchical model suggested by Ravasz *et al* [10]. The evaluation was made using Network Analyzer Version 2.7 [38] plug-in at Cytoscape Version 2.8.3 [39,40]. Besides, our networks had their Clustering Coefficient and Mean Shortest Path compared against 1,000 random networks produced by Random Network Version 1.0 (http://sites.google.com/site/randomnetworkplugin/) plug-in at Cytoscape. For that, the empirical *P-values* were calculated.

5 – Gene Ontology Annotation

For the functional annotation attribution we adopted the classification vocabulary defined by the Gene Ontology Consortium [41] (GO - http://www.geneontology.org/). The ontology covers three domains: cellular component, the parts of a cell or its extracellular environment; molecular function, the elemental activities of a gene product at the molecular level, such as binding or catalysis; and biological process, operations or sets of molecular events with a defined beginning and end, pertinent to the functioning of integrated living units: cells, tissues, organs, and organisms.

The GO annotation schema adopted in this work came from the public Kinetoplastid database TriTrypDB version 4.1 (http://tritrypdb.org/tritrypdb/) [42]. This database provides for each of three GO ontologies two kinds of evidence of annotation, one is called annotated and the other predicted. In order to guarantee a

higher confidence on the functional annotation, when possible the annotated terms were used for further analysis.

6 – Predicting Functional and Conserved Modules

At this part of the work, our goal was to identify functional modules that are conserved in the predicted networks. Functional modules can be understood as a group of proteins functionally or physically linked that work together to reach a distinct function [43]. Moreover, according to Ravasz *et al* [10], PPIs in general have a modular or hierarchical architecture.

Then, to perform this prediction, we chose the networkBLAST program [44] that performs two basic tasks: a) the comparison of multiple PPI networks; and b) the prediction of functional modules. The algorithm also combines interactions along with sequence information in order to produce a network alignment graph. Each node in this graph defines a group of similar proteins whereas links between nodes defines putative complexes that are evolutionarily conserved across the three predicted networks. Interactions reliability scores are assigned using a probabilistic model and the similarity information necessary for that is obtained from a comparison of all versus all sequences present in the predicted PPI networks.

Afterwards, in order to characterize the clusters or modules found, a functional annotation schema is required. As described in an earlier section, the GO functional annotation was used.

Following the functional annotation described above and considering the Biological Process ontology, a GO term enrichment analysis was performed using the GO::TermFinder [45] for each cluster. In this approach, the statistical significance is determined using the hypergeometric distribution to calculate the *P*-value:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}$$

here, N is equal to total number of proteins in the background distribution, which is the number of proteins in a PPI network that received at least one GO term, M represents the number of proteins within that background distribution that are annotated (either directly or indirectly) to any GO term of interest. n is the size of the list of proteins of interest (in our case it is the number of proteins in the module of interest). Finally, k is the number of proteins within that list or module which are annotated to the GO term of interest. Besides, as we were dealing with multiple hypotheses test, a correction for each P-value should be applied. Here, GO::TermFinder applied the Bonferroni correction.

7 – Topological Analysis

The metrics used in order to extract biological information from the predicted PPIs were calculated using the CytoHubba Version 1.6 plug-in [46] at Cytoscape. In this work we used Degree and Maximal Centrality Clique (MCC). According to CytoHubba developer site (http://hub.iis.sinica.edu.tw/cytoHubba/supplementary/index.htm), the MCC topological index showed the highest overlap with known essential proteins of PPI network of S*accharomyces cerevisiae.* The reported overlap was 80% for the top 10 proteins and 70% for the top 100 proteins of the network. Considering this outstanding performance, we use the MCC index to rank the top 20 proteins from the three predicted PPI network.

Moreover, the variability of the top ranked proteins was also assessed based on the ortholog group information present in the TriTrypDB. In this database, the proteins of the Kinetoplastids are clustered in groups based on OrthoMCL database information (http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi). Thus, for each ortholog group associated with the top ranked proteins a multiple sequence alignment

was performed using MAFFT [47] and the mean identity evaluated with the *alistat* program from HMMER package.

Complementarily, the immunologic potential of the selected top ranked proteins was addressed using BCPred12 [48], which is a predictor for potential epitopes recognized by B cells, NetCTL [49] and NetMHCII [50] which are predictors for potential epitopes with affinity binding to MHC class I and II alleles respectively. Finally, the predicted proteomes of *Mus musculus* (mouse), *Canis lupus familiaris* (dog) and *Homo sapiens* (human) were downloaded from NCBI repositories (http://www.ncbi.nlm.nih.gov/) on August 24, 2012 and used to address the sequence similarity between these genomes and the top 20 proteins ranked by MCC.

8 – Evolutionary Analysis

It has been described that the proteins with high Degree (the degree of a node in a network is the number of connections it has to other nodes) probably are proteins more conserved and ancient [51-53]. Then, in order to assess this assertion, we compared Degree and the nucleotide diversity index (π) [54] of the proteins present in the predicted PPI networks. This measure was obtained first by defining a Degree range in the predicted networks. The ranges were 2 to 10, 11 to 20, 21 to 30, 31 to 40, 41 to 50, and greater than 50. The selected proteins jointly with their orthologs extracted from the TriTrypDB were aligned using MAFFT, and then the π was calculated for each ortholog group using the Variscan program [55].

9 – Hypothetical Proteins Analysis

In the strict sense, hypothetical proteins are defined as proteins computational predicted from nucleic acid sequences that have not been shown to exist by any

experimental evidence. Furthermore, these proteins are characterized by low identity to the known annotated proteins in public domain databases.

The term "conserved hypothetical proteins" is also broadly employed and describes a fraction of genes in sequenced genomes that are found in organisms from several phylogenetic lineages but that have not been functionally characterized and described at the protein chemical level.

Trypanosomatids genomes are known to have a large amount of hypothetical proteins (~60%) [56,57], and these might be involved in essential cellular processes. Therefore, due to the importance and amount of these proteins in the genomes that we are working with and the possibility to use the PPI network to infer a function for them, we decided to apply an approach called FS-Weight [58] to try to obtain a clue on the possible functions for the hypothetical proteins.

The FS-Weight method, which stands for Functional Similarity Weight, is based on direct and indirect functional association using PPI networks as the main input. Either direct or indirect neighbors of a protein may share some physical or biochemistry features that allow them to bind to this protein. Therefore, this method has as an advantage that it does not use only direct interaction partners, which would limit prediction to proteins that have at least one interaction partner with known annotation, actually, FS-Weight also uses indirect interaction partners which increases the chance of predicting a protein function [58]. Furthermore, it calculates a functional similarity between two proteins, not necessarily from direct partners, based on the topological context of both proteins and the reliability of the interactions they do. This calculation is applied in order to reduce the effects of including erroneous interactions. Hence, the more common proteins exist interacting with two any proteins the chances that these two proteins share some biological function are higher. In addition, FS-Weight gives greater weight to common neighbors than non-common ones [58]. It is also worth mentioning that the FS-Weight performance was not re-evaluated for Leishmania species. The work that described the approach utilized data of *S. cerevisae* to validate the method. Therefore, some caution must be taken in using the function predictions made for hypothetical proteins present in our networks.

Moreover, to apply this annotation approach it is necessary to use an annotation schema that has already been used for the proteins with known functions. For this purpose we used the three GO ontologies already described.

Results

1 – Evaluation of PPI Prediction Approach

As detailed at Methods section, in order to evaluate the confidence of our network prediction methodology, gold standards positive and negative datasets were built from DIP database using the protein interaction data from *E.coli*, used here as model organism. This high quality control dataset that integrates 702 positive protein pairs and 3,510 negative protein pairs was used in the performance evaluation made by Receiver Operating Characteristics (ROC) graphs.

The accuracy of the proposed methodology measured by the area under the ROC curve can be addressed on Table 1 and through the plot presented on Figure 1. The AUC value of 0.94 obtained for *score_comb* indicates the robustness of the approach adopted. However, this result should be considered carefully since the databases used for the interolog-mapping contain many *E. coli* interactions. This might lead the evaluation of the confidence of our networks to some degree of bias.

2 – Filtering of data and PPIs prediction

As mentioned earlier, a filtering step was performed on the three proteomes in study in order to select sequences that were correctly annotated. A small percentage of proteins were excluded from our analyses since they presented possible errors. Then, the predicted proteome of *L. braziliensis*, *L. major* and *L. infantum* lost 4.33%, 2.95%, and 4.78% of proteins, respectively (Table 2).

Subsequent to the filtering process, the three proteomes were used for PPIs prediction. These predictions were made based on different databases, such as Domine, PSI-Base, IntAct and String. Using these evidences, we proposed and calculated a combined score for the interactions predicted in the PPIs which ranged from 0 to 1. Afterwards, it was possible to demonstrate the wellness of fit of scale-free models for the three predicted PPIs (Table 3).

The PPIs predicted were then compared against 1,000 random networks. The Clustering Coefficient and the Mean Shortest Path were compared (Table 3). The values of the Clustering Coefficient of the PPIs are much greater than the random networks adding an extra layer of credibility for the predicted networks.

As a result, the predicted PPIs incorporated 23%, 24%, and 25% (Table 4) of the proteins from the filtered proteomes of *L. braziliensis* (Table S1)*, L. major* (Table S2) and *L. infantum* (Table S3) respectively. Figure 2 shows one of the three networks.

Furthermore, we used GO terms to try to draw a function profile of the networks. For this analysis, we used the predicted terms present in TritrypDB database instead of annotated terms. The rationale underlying this choice was associated with the small number of GO terms annotated for *L. braziliensis* that would prevent its comparison against the other two leishmanias. The three ontologies (Biological Process, Cellular Component, and Molecular Function) were applied and similar results were found. Considering a frequency larger than 2 for a given GO term, it is worth pointing out that the total intersection among the predicted networks was 79%, 84%, and 75% for Biological Process, Cellular Component, and Molecular Function, respectively. In fact, from the top 10 most frequent GO terms for each ontology, 8 of them for Biological Process, 7 of them for Molecular Function and all of them for Cellular Component are the same for the three networks.

3 – Evolution Analysis

In order to obtain information relative to the correlation between the number of interactions that a protein does and its conservation degree, we compared the

number of interactions of the proteins of our networks against the nucleotide diversity of the genes that encode them (Figure 3). Based on this analysis, as the proteins increase the number of interactions that they participate in, their diversity degree, measured here by π (nucleotide diversity index), decrease. From the results obtained for the three predicted networks we can suggest the existence of an evolutionary pressure for the maintenance of a lower diversity in proteins with a high number of interactions.

4 – Characterizing modules

At this point, the algorithm networkBLAST was used to identify the modules in the PPIs. The number of conserved modules shared by the three species of *Leishmania* was 199. Despite over millions years of proposed divergence for the analyzed species, this result is not surprising considering that a high synteny was already observed and reported between all sequenced *Leishmania* species [59].

Subsequently, as detailed in Methods section, a function annotation assignment to the network modules was performed using the Biological Process hierarchy of the Gene Ontology and the Perl programming modules GO::TermFinder. This approach allowed the identification of 153 modules which had GO terms with a frequency higher than the expected. In that cases where a given network module received more than one GO term, the most significant one characterized by the smallest P-value in the enrichment analysis was chosen. A complete description including the results obtained for all 153 networks modules annotated can be found in the Table S4. It worths to mention that differently from standard clustering algorithms the networkBLAST approach can produce overlapping modules which makes sense from the biological point of view since one protein can belong to more than one network module.

It is also important to highlight that only 57 unique GO terms were used to describe the 153 network modules predicted and the most frequent terms were assigned to modules which are likely involved in biological processes related to

protein folding, translation, tRNA aminoacylation for protein translation, energy derivation by oxidation of organic compounds and carbohydrate metabolism.

Taking into consideration the biological significance of this functional analysis, these results were overlapped with topological analysis.

5 – Topological analysis

According to our proposed methodology, two topological indexes (Degree and MCC) were utilized to study the interaction networks predicted here (Table S5). Then, we sorted the proteins present in the PPIs using the MCC index, and we obtained a list of proteins that are central for different cliques (subgraphs) and with high interaction degree (Table S6).

The following analyses were conducted for that list of proteins: a) amino acid variability present in orthologs groups; b) degree of conservation against proteins of three potential hosts (*M. musculus*, *C. lupus familiaris* and *H. sapiens*); and c) epitope computational prediction.

Regarding the variability of these proteins, our results revealed an average identity of 80% between the top 20 proteins ranked by MCC index and their orthologs. Therefore, it was possible to notice that these proteins were relatively conserved among the Kinetoplastids.

On the other hand, only two proteins, LbrM22_V2.0510 (proteasome regulatory ATPase subunit 1) and LmjF36.1650 (proteasome beta 5 subunit), from *L. braziliensis* and *L .major* respectively, had identity higher than 60% when compared against the host proteomes. In addition,

L. infantum presented 2 proteins with identity higher than 60%, they are LinJ36_V3.1730 (proteasome beta 5 subunit) and LinJ22_V3.0490 (proteasome regulatory ATPase subunit 5).

In this context, we can suggest that the low identity presented by the great majority of the top ranked proteins can be interesting for drug and vaccine studies.

The rationale in suggesting that these proteins could be used for medical purposes can be reinforced by the predicted function of the network modules that they are inserted in. We noted that the most of modules were involved in protein turnover which is known to be involved in responses to vaccination [60]. In addition, we found a total of 9 GO terms describing these modules and 7 of them are shared by the top 20 ranked proteins of each predicted PPI network.

Finally, in respect to immunological potential for these proteins, all of them had more than 5 epitopes predicted for B cells receptors. For the epitope prediction for MHC class I, 12 different alleles were tested and all tested proteins had at minimum of 2 predicted epitopes with potential binding affinity to at least 11 alleles. The last analysis was for epitope predictions of MHC class II. The predictor used for this analysis provides along with the epitope prediction a measure of binding affinity between the epitope and the receptor, and this measure is divided in two categories: weak binding (WB) and strong binding (SB). We selected just predictions which were categorized as SB. Thus, all proteins have at least 2 epitopes with binding affinity to at minimum of 1 allele tested. The total of tested alleles was 17.

6 – Annotation prediction for hypothetical proteins

In order to address the usefulness of the predicted network to assign some level of functional annotation to hypothetical proteins, we decided to use an approach called FS-Weight that takes into account both direct and indirect neighbors as detailed at Method section. From the total number of proteins covered by the networks, approximately 21% were originally annotated as hypothetical. From this set of proteins, nearly 40%, 48%, and 55% of them received some GO term based on the FS-Weight approach (Table 4). In addition, it is important to point out that this approach provides a score for all annotation prediction that ranges from 0 to 1, and that just GO terms which received a score equal to 1 were considered. Furthermore, when we crossed the information on modules against the hypothetical proteins that received a putative function it was possible to note that for the three networks the hypothetical proteins are more frequently present in modules involved in RNA

metabolism. All proteins that received a functional annotation are available in Table S7.

Discussion

Based on the results obtained regarding the accuracy of the proposed approach for network prediction (AUC value equal to 0.94), we can state that the prediction methodology is relatively reliable (Figure 1), and the predicted protein interactions own a good confidence. However, as it was said on Results section, the databases used for the methodology have many interactions of *E. coli*. This might make the performance evaluation a practice of circular reasoning, and thus lead it to some degree of bias.

In addition, we compared our interaction score schema against others previously published [61]. It is clear from the obtained results (AUC values presented in Table 1) that the score schema we used outperformed the others. Therefore, we applied our interaction score schema for leishmania PPI networks predictions.

Furthermore, the lost associated with the filtering step (detailed in results) was small and this result reflects the quality of genome annotation for the three different leishmanias, which is valuable since our main input data was the protein sequences and the final results depended on the quality of them.

Still on the computational prediction quality issue, in our results we described the assessment of the PPIs based on some known network models such as scale-free model [9] to guarantee their confidence. It is possible to suggest that the PPI networks predicted are consistent as they present features which are common for biological networks currently described. In addition, when the PPIs were compared to random networks (Table 3), it was possible to notice that the values of the Clustering Coefficient of the PPIs are much greater than the random networks, a find that once again suggests the PPIs prediction strength and the absence of spurious interactions. Both results can be used to illustrate the confidence of interolog mapping approach

and to reinforce the result found for its evaluation performance, even when there might be a possibility of bias on the evaluation.

In terms of the number of proteins present in the PPIs, our findings are comparable to those found for *L. major* by Flórez 2010, which found nearly 16% of the *L. major* predicted proteome in a predicted PPI. According to the authors, the reason for the small number of proteins mapped in PPIs is a reflection of low levels of similarity between leishmania species and the used database content. On the other hand, the differences between the predicted number of interactions observed in our work and Flórez 2010 can be explained by the different sources of information and approaches used.

Following the network assessment, the first analysis performed in the three PPIs was a Gene Ontology functional annotation. Moreover, it is also noteworthy that the most frequent terms for the three networks regarding Molecular Function ontology are related with binding function, which makes sense since the proteins present in the PPIs are predicted to interact with each other. On the other hand, about the Cellular Component category, we observed terms associated with protein complexes such as proteasome and ribosome. Again, this was somehow expected since a set of interacting proteins possibly are going to form complexes. However, for Biological Process, we obtained a higher diversity of terms that can be hypothesized to be explained by the fact that the same protein can participate in many different processes in a cell.

We also performed an evolution analysis in order to verify whether there was any trend related to the number of interactions and protein sequence diversity. Our results indicate that the number of interactions and diversity are inversely proportional, meaning that as the diversity increases, the number of interactions decreases. In protein-protein interaction networks, proteins presenting several interactions (high degree) are generally called *hub* nodes and genome-wide studies [62,63] have shown that the deletion of a hub protein is more likely to be lethal than the deletion of a non-hub protein (centrality-lethality rule). In addition, this finding makes sense because these proteins probably are involved in different biological process within a cell with relative success and, in this context, if a random mutation happens, it will likely produce a negative outcome.

Therefore, the points raised herein show that the predicted PPI networks are biologically consistent. Otherwise, we had a trend of proteins with a wide diversity of conservation as hubs.

Another point addressed in our analysis was the network modularity of the predicted PPIs. Modularity is one measure of the structure of networks and many previous works have reported that biological networks are modular [10,37,43,64]. This feature is important for their robustness since a modular architecture guarantees that a system failure is isolated [37]. Thus, if we are interested in destabilizing the PPIs for drug or vaccine purposes, we need to know the modules present in the networks.

In this context, aiming to measure modularity, a clustering analysis was performed in order to identify conserved modules. As it was stated in results section, we found 153 conserved modules which had a function assigned by the enrichment analysis, and these modules can be grouped in 57 different functions.

Based on these findings, it is possible to note that there are many protein complexes (modules) that are essential for the studied organisms. Thus, it is worth to explore in more details these complexes along with the topological information of the network proteins with the potential to elect new potential proteins targets for vaccine and drug development.

In addition, other sources of information were integrated to topological analysis, such as immunological potential, degree of protein sequence conservation among orthologs and degree of identity compared to proteins of potential parasite hosts (human, dog and mouse).This information integration provides a better understanding that can be valuable to select new potential biological targets.

Using this rationale, we suggested a list of proteins (Table S6) that can be attractive for medical purposes. These proteins have a low identity against proteins from hosts, they are potentially recognized by B cells and T cell receptors and are highly conserved compared to their orthologs. In addition, they seem to be central for many biological processes as they have high values of MCC and degree indexes, thus if they are neutralized all the system of protein interaction might suffer severe damage.

Moreover, those proteins do not have high level of identity against the proteins from host proteomes, a desirable characteristic for proteins that will be selected for vaccine development and/or drug therapy. Consequently, side effects can be avoided. Other important feature is the high level of conservation of them when compared against their orthologs; this can guarantee a wide spectrum of action. In the end, they have several potential epitopes which are fundamental for the most important kinds of immunological responses.

Finally, we are interested in using the PPI network information in an annotation framework to assign a putative function to the currently predicted hypothetical proteins. Within the Trypanosamatids context, the study of hypothetical proteins has huge importance, since some organisms, which comprise a part of this group such as the ones that are targets of this work, have around 60% of their predicted proteomes composed of uncharacterized proteins [56,57]. This scenario is kept current even within the 'omics' age because the majority of studies often focus on already well understood and established molecular scenarios. Therefore, the opportunity to expand knowledge further than the known and expected is rarely attempted [65].

Furthermore, the majority of researchers are not interested in investigating the molecular data that are hard to interpret in the light of current biological knowledge, i.e. data on hypothetical proteins [65]. However, the Systems Biology approaches can help to improve these numbers. Thus, there is a group of methods in the Systems Biology context that aims at exploiting information derived from networks to elucidate functional prediction. Hence, various classification methods allow for general function predictions utilizing 'homology-free' protein sequence features [65].

An example of the application of a network study to elucidate a function of an uncharacterized protein can be found in the work of Cui *et al* where they built a protein-protein interaction network for *Mycobacterium tuberculosis* using an homology protein mapping approach [66]. In this study, a hypothetical protein with a high degree of interaction was found and evidence for its function came from the fact that it interacts with the same group of ABC transporter ATPase subunits as does a known protein [66]. Thus, this rationale of assigning a function based on the neighbors of a protein can be extremely useful.

In our results, around 50% of the hypothetical proteins present in the networks received some functional annotation. Moreover, the most frequent modules, where those proteins are present, are related to RNA metabolism. This could be interesting as there is currently a huge amount of studies involving different types of RNA and their roles in distinct biological phenomena.

Finally, our group is involved in analysis concerning "Intrinsically Unstructured Proteins" (IUPs) and our previous results (still unpublished) link many features of this group of proteins with the group of hypothetical proteins in Trypanosomatids (data not published). This should be investigated in the future since there are many articles showing how important the IUPs are for the protein-protein interaction networks [67-70].

Conclusion

This work was the first to predict three protein-protein interaction networks for three different species of *Leishmania* and to compare them to each other. A new interaction score schema was proposed and proved to be reliable. Using this strategy, we observed that it is possible to extract important information related to the biology of the studied organism. In addition, using the topological information, we can select proteins that are potential targets for drugs and vaccine development. However, since vaccine and drug prediction represent a complex and multifactorial problem, more data, such as structural data, expression data, etc could be added in order to choose the proteins for future studies in a more efficient way.

In addition, addressing the network information, it was possible to infer some clues regarding some hypothetical proteins that did not have any information related to their molecular functions in the cell.

In summary, based on the evidences reported here, we believe that the networks modeled are biologically consistent and can be useful as tool for different kinds of studies on these organisms.

References

1. Murray HW, Berman JD, Davies CR, Saravia NG (2005) Advances in leishmaniasis. Lancet 366: 17.

2. Hotez PJ, Bottazzi ME, Franco-Paredes C, Ault SK, Periago MR (2008) The Neglected Tropical Diseases of Latin America and the Caribbean: A Review of Disease Burden and Distribution and a Roadmap for Control and Elimination. Plos Neglected Tropical Diseases 2: e300. doi:10.1371/journal.pntd.0000300.

3. Hadighi R, Boucher P, Khamesipour A, Meamar AR, Roy G, et al. (2007) Glucantime-resistant Leishmania tropica isolated from Iranian patients with cutaneous leishmaniasis are sensitive to alternative antileishmania drugs. Parasitology Research 101: 1319-1322.

4. Rojas R, Valderrama L, Valderrama M, Varona MX, Ouellette M, et al. (2006) Resistance to antimony and treatment failure in human Leishmania (Viannia) infection. Journal of Infectious Diseases 193: 1375-1383.

5. Lira R, Sundar S, Makharia A, Kenney R, Gam A, et al. (1999) Evidence that the high incidence of treatment failures in Indian kala-azar is due to the emergence of antimony-resistant strains of Leishmania donovani. Journal of Infectious Diseases 180: 564-567.

6. Rijal S, Yardley V, Chappuis F, Decuypere S, Khanal B, et al. (2007) Antimonial treatment of visceral leishmaniasis: are current in vitro susceptibility assays adequate for prognosis of in vivo therapy outcome? Microbes and Infection 9: 529-535.

7. Sauer U, Heinemann M, Zamboni N (2007) Genetics - Getting closer to the whole picture. Science 316: 550-551.

8. Harrington ED, Jensen LJ, Bork P (2008) Predicting biological networks from genomic data. Febs Letters 582: 1251-1258.

9. Barabasi AL, Albert R (1999) Emergence of scaling in random networks. Science 286: 509-512.

10. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. Science 297: 1551-1555.

11. Skrabanek L, Saini HK, Bader GD, Enright AJ (2007) Computational Prediction of Protein-Protein Interactions. Molecular Biotechnology 38: 17.

12. Huynen MA, Bork P (1998) Measuring genome evolution. Proceedings of the National Academy of Sciences of the United States of America 95: 5849-5856.

13. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, et al. (1999) Detecting protein function and protein-protein interactions from genome sequences. Science 285: 751-753.

14. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86-90.

15. Sato T, Yamanishi Y, Kanehisa M, Toh H (2005) The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21: 3482-3489.

16. Kim JG, Park D, Kim BC, Cho SW, Kim YT, et al. (2008) Predicting the Interactome of Xanthomonas oryzae pathovar oryzae for target selection and DB service. Bmc Bioinformatics 9: 41.

17. Florez AF, Park D, Bhak J, Kim BC, Kuchinsky A, et al. (2010) Protein network prediction and topological analysis in Leishmania major as a tool for drug target selection. Bmc Bioinformatics 11: 484.

18. Matthews LR, Vaglio P, Reboul J, Ge H, Davis BP, et al. (2001) Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". Genome research 11: 2120-2126.

19. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, et al. (2004) Annotation transfer between genomes: protein–protein interologs and protein–DNA regulogs. Genome research 14: 1107-1118.

20. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, et al. (2000) DIP: the Database of Interacting Proteins. Nucleic Acids Research 28: 289-291.

21. Muley VY, Ranjan A (2012) Effect of Reference Genome Selection on the Performance of Computational Methods for Genome-Wide Protein-Protein Interaction Prediction. PloS one 7: e42057.

22. Gomez SM, Noble WS, Rzhetsky A (2003) Learning to predict protein-protein interactions from protein sequences. Bioinformatics 19: 1875-1881.

23. Jansen R, Yu HY, Greenbaum D, Kluger Y, Krogan NJ, et al. (2003) A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302.

24. Zhang LV, Wong SL, King OD, Roth FP (2004) Predicting co-complexed protein pairs using genomic and proteomic data integration. Bmc Bioinformatics 5: 38.

25. Jansen R, Gerstein M (2004) Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Current Opinion in Microbiology 7: 535:545.

26. Qi YJ, Klein-Seetharaman J, Bar-Joseph Z (2005) Random forest similarity for protein-protein interaction prediction from multiple sources. Pacific Symposium on Biocomputing 2005: 531-542.

27. Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940-3941.

28. Raghavachari B, Tasneem A, Przytycka TM, Jothi R (2008) DOMINE: a database of protein domain interactions. Nucleic Acids Research 36: D656-D661.

29. Gong S, Yoon G, Jang I, Bolser D, Dafas P, et al. (2005) PSIbase: A database of Protein Structural Interactome map (PSIMAP). Bioinformatics 21: 2541-2543.

30. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, et al. (2010) The IntAct molecular interaction database in 2010. Nucleic Acids Research 38: D525-D531.

31. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8-a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research 37: D412-D416.

32. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic acids research 39: D561-D568.

33. SF A, W G, W M, EW M, DJ L (1990) Basic Local Alignment Tool. Journal of Molecular Biology 5: 7.

34. Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14: 755-763.

35. Finn RD, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The Pfam protein families database. Nucleic Acids Research 38: D211-D222.

36. von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, et al. (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. Nucleic Acids Research 33: D433-D437.

37. Barabasi AL, Oltvai ZN (2004) Network biology: Understanding the cell's functional organization. Nature Reviews Genetics 5: 101-U115.

38. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M (2008) Computing topological parameters of biological networks. Bioinformatics 24: 282-284.

39. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, et al. (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research 13: 2498-2504.

40. Smoot ME, Ono K, Ruscheinski J, Wang PL, Ideker T (2011) Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics 27: 431-432.

41. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nature Genetics 25: 25-29.

42. Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, et al. (2010) TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Research 38: D457-D462.

43. Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. Nature 402: C47-C52.

44. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, et al. (2005) Conserved patterns of protein interaction in multiple species. Proceedings of the National Academy of Sciences of the United States of America 102: 1974-1979.

45. Boyle EI, Weng SA, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder - open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710-3715.

46. Lin CY, Chin CH, Wu HH, Chen SH, Ho CW, et al. (2008) Hubba: hub objects analyzer - a framework of interactome hubs identification for network biology. Nucleic Acids Research 36: W438-W443.

47. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research 30: 3059-3066.

48. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. Journal of Molecular Recognition 21: 243-255.

49. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, et al. (2007) Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. Bmc Bioinformatics 8: 424.

50. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. BMC bioinformatics 10: 296.

51. Wuchty S (2004) Evolution and topology in the yeast protein interaction network. Genome Research 14: 1310-1314.

52. Wuchty S (2006) Topology and weights in a protein domain interaction network - a novel way to predict protein interactions. Bmc Genomics 7: 122-133.

53. Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C, Feldman MW (2002) Evolutionary rate in the protein interaction network. Science 296: 750-752.

54. Nei M, Li WH (1979) MATHEMATICAL-MODEL FOR STUDYING GENETIC-VARIATION IN TERMS OF RESTRICTION ENDONUCLEASES. Proceedings

of the National Academy of Sciences of the United States of America 76: 5269-5273.

55. Vilella AJ, Blanco-Garcia A, Hutter S, Rozas J (2005) VariScan: Analysis of evolutionary patterns from large-scale DNA sequence polymorphism data. Bioinformatics 21: 2791-2793.

56. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, et al. (2005) The genome of the kinetoplastid parasite, Leishmania major. Science 309: 436-442.

57. Peacock CS, Seeger K, Harris D, Murphy L, Ruiz JC, et al. (2007) Comparative genomic analysis of three Leishmania species that cause diverse human disease. Nature Genetics 39: 839-847.

58. Chua HN, Sung WK, Wong L (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. Bioinformatics 22: 1623-1630.

59. El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, et al. (2005) Comparative genomics of trypanosomatid parasitic protozoa. Science 309: 404-409.

60. Garlick PJ, McNurlan MA, Fern EB, Tomkins AM, Waterlow JC (1980) STIMULATION OF PROTEIN-SYNTHESIS AND BREAKDOWN BY VACCINATION. British Medical Journal 281.

61. Yu HY, Luscombe NM, Lu HX, Zhu XW, Xia Y, et al. (2004) Annotation transfer between genomes: Protein-protein interologs and protein-DNA regulogs. Genome Research 14.

62. Jeong H, Mason SP, Barabasi AL, Oltvai ZN (2001) Lethality and centrality in protein networks. Nature 411: 41-42.

63. Lee I, Lehner B, Crombie C, Wong W, Fraser AG, et al. (2008) A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. Nature Genetics 40: 181-188.

64. Barabasi AL, Ravasz E, Oltvai Z (2003) Hierarchical organization of modularity in complex networks. Statistical Mechanics of Complex Networks 625: 46-65.

65. Pawłowski K (2008) Uncharacterized/hypothetical proteins in biomedical 'omics' experiments: is novelty being swept under the carpet? Briefings in functional genomics & proteomics 7: 283-290.

66. Cui T, Zhang L, Wang X, He Z-G (2009) Uncovering new signaling proteins and potential drug targets through the interactome analysis of Mycobacterium tuberculosis. Bmc Genomics 10.

67. Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, et al. (2007) Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. Bmc Genomics 9: .

68. Haynes C, Iakoucheva LM (2006) Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. Nucleic Acids Research 34: 305-312.

69. Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN (2005) Flexible nets - The roles of intrinsic disorder in protein interaction networks. Febs Journal 272: 5129-5148.

70. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nature Reviews Molecular Cell Biology 6: 197-208.

Acknowledgment

We thank for the following agencies for their past and current support for our research: CPqRR – FIOCRUZ (Centro de Pesquisas René Rachou – FIOCRUZ); CAPES (Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior); FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) grant APQ-02382-10 and APQ-01835-10. CNPq (National Counsel of Technological and Scientific Development) grant 476539/2010-2, GENOPROT (Rede Integrada de Estudos Genômicos e Proteômicos) grant 560943/2010-5.

Figure Legends

Figure 1. Performance evaluation of approached used to predict PPI networks using the ROC curve. Here the predictions were compared against a gold standard data of interactions extracted from DIP database for *E. coli* (see text for details).

Figure 2. Protein-Protein Interaction for *L. infantum* visualized using Cytoscape 2.8.3 and the Edge-weighted spring embedded layout.

Figure 3. Degree versus diversity analysis. Graph of median of Nucleotide Diversity (π) measure (Y axis) versus Degree range (X axis) of three PPIs.

Tables

Table 1. Performance evaluation of approach used to predict PPI networks

Measure of confidence	**AUC value**
Developed method	0.94
Geometric mean of score	0.74
Geometric mean of *evalue*	0.57
Maximum *evalue*	0.55

Table 2. Number of proteins in the predicted proteome of the target organisms before and after the filtering

Organism	Total of proteins	Total of proteins after filtering	Relative number of lost proteins (%)
L. braziliensis	8310	7950	4.33
L. major	8408	8160	2.95
L. infantum	8216	7823	4.78

Table 3. Fitting results for scale-free model, and Clustering Coefficient and Mean Shortest Path for PPIs compared against the same measure extracted from 1000 Random PPIs

Leishmania braziliensis			
Scale free model	Correlation	R^2	
	0.941	0.816	
Random model			
Measure	Modeled PPI	Random PPIs	P-value
Clustering Coefficient	0.433	0.159±0,003	$p<0.05$
Mean Shortest Path	2.877	2.579±0,004	$p<0.05$
Leishmania major			
Scale free model	Correlation	R^2	
	0.925	0.815	
Random model			
Measure	Modeled PPI	Random PPIs	P-value
Clustering Coefficient	0.430	0.157±0.003	$p<0.05$
Mean Shortest Path	2.914	2.584±0.004	$p<0.05$
Leishmania infantum			
Scale free model	Correlation	R^2	
	0.940	0.829	
Random model			
Measure	Modeled PPI	Random PPIs	P-value
Clustering Coefficient	0.424	0.160±0.003	$p<0.05$
Mean Shortest Path	2.886	2.573±0.004	$p<0.05$

Table 4. General features of the three predicted PPI Networks

Organism	Number of Nodes (Proteins)	Number of Interactions	Number of hypothetical protein	Number of hypothetical protein annotated (%)*

L. braziliensis	1818	39420	381	153 (40%)
L. major	1947	43531	416	200 (48%)
L. infantum	1959	45235	415	229 (55%)

*Proteins were annotated following the methodology described in the text.

Supporting Information Legends

Table S1 - *Leishmania braziliensis* PPI network – Description of all predicted protein interactions and their confidence scores for *L. braziliensis* predicted proteome.

Table S2 - *Leishmania major* PPI network – Description of all predicted protein interactions and their confidence scores for *L. major* predicted proteome.

Table S3 - *Leishmania infantum* PPI network – Description of all predicted protein interactions and their confidence scores for *L. infantum* predicted proteome.

Table S4 - Annotation of Functional Modules (Clusters) predicted for the three PPI networks – Description of predicted modules including their scores, *p-values* and GO term id.

Table S5 – Topological analysis – Values of MCC and *Degree* indexes calculated for the proteins present in the three PPI networks modeled

Table S6 – Analysis of top 20 ranked proteins of each PPI network – Detailed description of top 20 ranked proteins by MCC index including product description, MCC and *Degree* values, host protein analysis, orthologs analysis and immunological analysis.

Table S7 – Annotation of hypothetical proteins – Predicted annotation based on FS-Weight and GO ontology assigned to hypothetical proteins present in the networks.

ARTIGO 2

De: BioMed Central Editorial [editorial@biomedcentral.com]

Enviado: domingo, 11 de novembro de 2012 20:38

Para: Jeronimo Ruiz - Fiocruz

Assunto: Your manuscript is acceptable for publication in principle.

Authors: Daniela M Resende, Antonio M Rezende, Nesley JD Oliveira, Izabella CA Batista, Rodrigo Corrêa-Oliveira, Alexandre B Reis and Jeronimo C Ruiz

Title : An assessment on epitope prediction methods for protozoa genomes

Journal: BMC Bioinformatics

MS : 4657522726961915

Dear Dr. Ruiz,

Peer review of your manuscript (above) is now complete, and we are delighted, in principle, to accept the manuscript for publication in BMC Bioinformatics.

However before acceptance, our editorial production team needs to check the format of your manuscript, to ensure that it conforms to the standards of the journal. They will get in touch with you shortly to request any necessary changes or to confirm that none are needed.

If you have any problems or questions regarding your manuscript, please do get in touch.

Best wishes,

Neil

Nathaniel Nazareno

BioMed Central Editorial Office

on behalf of Dr. Catherine Rice

Tel: +44 (0) 20 3192 2013

e-mail: editorial@biomedcentral.com

Web: http://www.biomedcentral.com/

Centro de Pesquisas René Rachou/CPqRR -- A Fiocruz em Minas Gerais.

Rene Rachou Research Center/CPqRR -- The Oswaldo Cruz Foundation in the State of Minas Gerais, Brazil.

www.cpqrr.fiocruz.br<http://www.cpqrr.fiocruz.br>

An assessment on epitope prediction methods for protozoa genomes

Daniela M Resende[1,2#], Antônio M Rezende[4,5#], Nesley JD Oliveira[4,5,6], Izabella CA Batista[2], Rodrigo Corrêa-Oliveira[2], Alexandre B Reis[2,3], Jeronimo C Ruiz[4]*

[1]Departamento de Análises Clínicas, Laboratório de Pesquisas Clínicas, Escola de Farmácia, Universidade Federal de Ouro Preto, Campus Morro do Cruzeiro, 35400-000, Ouro Preto – MG, Brasil

[2]Laboratório de Imunologia Celular e Molecular, Instituto René Rachou, Av. Augusto de Lima, 1715, Barro Preto, 30190-002, Belo Horizonte – MG, Brasil

[3]Laboratório de Imunopatologia, Núcleo de Pesquisas em Ciências Biológicas, Universidade Federal de Ouro Preto, Campus Morro do Cruzeiro, ICEB II, 35400-000, Ouro Preto – MG, Brazil

[4]Laboratório de Parasitologia Celular e Molecular, Instituto René Rachou, Av. Augusto de Lima, 1715, Barro Preto, 30190-002, Belo Horizonte – MG, Brazil

[5]Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Av. Antônio Carlos, 6627, Pampulha, 31270-901, Belo Horizonte – MG, Brazil

[6]Pontifícia Universidade Católica, R. Rio Comprido, 4580, Monte Castelo, 32285-040, Contagem – MG, Brazil

*Corresponding author

#these authors have equally contributed to this work

Email addresses:

DMR: dmresende@ef.ufop.br

AMR: am.rezende@cpqrr.fiocruz.br

NJDO: nesdaher@gmail.com

ICAB: izabellacristinab@hotmail.com

RCO: correa@cpqrr.fiocruz.br

ABR: alexreis@nupeb.ufop.br

JCR: jeronimo@cpqrr.fiocruz.br, jero.ruiz@gmail.com

Abstract

Background

Epitope prediction using computational methods represents one of the most promising approaches to vaccine development. Reduction of time, cost, and the availability of completely sequenced genomes are key points and highly motivating regarding the use of reverse vaccinology. Parasites of genus *Leishmania* are widely spread and they are the etiologic agents of leishmaniasis. Currently, there is no efficient vaccine against this pathogen and the drug treatment is highly toxic. The lack of sufficiently large datasets of experimentally validated parasites epitopes represents a serious limitation, especially for trypanomatids genomes. In this work we highlight the predictive performances of several algorithms that were evaluated through the development of a MySQL database built with the purpose of: a) evaluating individual algorithms prediction performances and their combination for CD8+ T cell epitopes, B-cell epitopes and subcellular localization by means of AUC (Area Under Curve) performance and a threshold dependent method that employs a confusion matrix; b) integrating data from experimentally validated and *in silico* predicted epitopes; and c) integrating the subcellular localization predictions and experimental data. NetCTL, NetMHC, BepiPred, BCPred12, and AAP12 algorithms were used for *in silico* epitope prediction and WoLF PSORT, Sigcleave and TargetP for *in silico* subcellular localization prediction against trypanosomatid genomes.

Results

A database-driven epitope prediction method was developed with built-in functions that were capable of: a) removing experimental data redundancy; b) parsing algorithms predictions and storage experimental validated and predict data; and c) evaluating algorithm performances. Results show that a better performance is

achieved when the combined prediction is considered. This is particularly true for B cell epitope predictors, where the combined prediction of AAP12 and BCPred12 reached an AUC value of 0.77. For T CD8+ epitope predictors, the combined prediction of NetCTL and NetMHC reached an AUC value of 0.64. Finally, regarding the subcellular localization prediction, the best performance is achieved when the combined prediction of Sigcleave, TargetP and WoLF PSORT is used.

Conclusions

Our study indicates that the combination of B cells epitope predictors is the best tool for predicting epitopes on protozoan parasites proteins. Regarding subcellular localization, the best result was obtained when the three algorithms predictions were combined. The developed pipeline is available upon request to authors.

Background

Reverse vaccinology uses the genome sequences of viral, bacterial or parasitic pathogens of interest rather than the cells as starting material for the identification of novel antigens, whose activity should be subsequently confirmed by experimental biology approaches [1]. In general, the aim of this approach is the identification of genes potentially encoding pathogenicity factors and secreted or membrane-associated proteins. In this context, specific algorithms suitable for the *in silico* identification of novel surface-exposed and, thus, antibody accessible proteins mediating a protective immune response are used [2].

Pizza and co-workers in collaboration with The Institute for Genomic Research (TIGR) provided the first example of a successful application of the reverse vaccinology approach [3]. They described that *in silico* identification of vaccine

candidates against *Neisseria meningitides* serogroup B, which is the major cause of sepsis and meningitis in children and young adults, could be effective, while conventional approaches to obtain a vaccine had failed for decades.

New powerful genomic technologies have increased the number of diseases that can be addressed by vaccination, and have reduced the time of discovery research and vaccine development [1]. Nowadays, it costs US$ 200-400 million to research, develop, manufacture and launch a new vaccine on the global market [4]. With the use of reverse vaccinology, time and cost spent on the search of new vaccine targets are significantly reduced.

Immunoinformatics is an emerging application of bioinformatics techniques that focuses on the structure, function, and interactions of the molecules involved in immunity. One of its main goals is the *in silico* prediction of immunogenicity at epitope level. Recently developed *in silico* tools and databases can be used to identify, characterize or predict antigen epitopes recognized by T- and B-lymphocytes, cells that play significant roles in infection and protective immunity [5].

Epitopes are the minimal essential units of information derived from self and nonself proteins that stimulate cellular (T-cell) and humoral (B-cell) immune responses. T-cells recognize T-cell epitopes that are derived from endogenous and exogenous proteins and presented in the cleft of MHC class I or MHC class II molecules at the surface of antigen presenting cells to the T-cell receptor. After the activation of CD8+ T cells or CD4+ T cells, respectively, cellular events, such as citotoxicity and cytokine secretion, will occur. B-cells also recognize epitopes, but generally intact

proteins. B-cell epitopes can be linear, contiguous amino acids, or discontinuous amino acids that are brought together in folded proteins. After activation, B-cells differentiate into plasmocytes and start secreting antibodies. B- and T-cell responses are called humoral and cellular adaptive immune responses, respectively, and they inform the immune system that a bacteria, virus, or parasite is present [6].

The subcellular localization of the protein is also important to investigate, as immunogenic proteins have to be in contact with T- and B-cells in order to elicit a protective immune response. In other words, correct subcellular localization is of great significance to the functional analysis of proteins [7]. Therefore, various prediction methods have been developed to predict proteins' subcellular location in the recent decades [8]. Prediction methods to identify the subcellular location of proteins can be classified into two categories: one is based on the recognition of protein N-terminal sorting signals [9] and the other is based on amino acid composition [10]. The predictors then combine these features with machine-learning techniques to decide which is the most probable location [11, 12].

A large variety of machine-learning techniques are commonly used in bioinformatics, including artificial neural networks (ANNs) [13], hidden Markov models (HMMs) [14] and support vector machines (SVMs) [15]. ANNs and SVMs are ideally suited to recognize non-linear patterns, which are believed to contribute to, for instance, peptide-HLA-I interactions [16]. In an ANN, information is trained and distributed into a computer network with an input layer, hidden layers and an output layer all connected in a given structure through weighted connections [13]. Finally, HMMs are well suited to characterized biological motifs with an inherent structural composition,

and have been used in the field of immunology to predict peptide binding to major histocompatibility complex (MHC) class I molecules [17].

The use of database system has been constant in the life of researchers and professionals in several fields. Conceptually, a database should be able to provide an easy access to experimental results and lexical surveys, preventing redundancy and wasteful duplication of research data. A well-designed database should also be able to provide support to researchers, facilitating guided searches for novel correlations in data. On the other hand, a poorly designed database makes the data mining process difficult and the new data integration infeasible for regular users [18] and in this perspective, the rebuilding and redesigning processes are frequent [18, 19].

The current challenge of modern biology is to unravel and understand the complex system of biological organization and to signal in all of its details at a molecular level. An essential part of this process goes through bioinformatics, particularly the use of management systems, and relational databases applied to biological data. Biological data reside in specialized databases that represent different data interpretation stages or different facets of biological phenomena [20]. Also, biological data present a particularity: they are highly complex when compared with data from most of other applications. Thus, definitions of such biological data must be able to represent a complex substructure of data as well as their relationships, and also ensure that no information is lost during the biological data modeling. The data model must be able to represent any level of complexity in any data schema, relationship, or schema substructure and not just in a hierarchical, binary, or tabular data format [21].

The main objective of this present work was to build a database-driven epitope prediction method capable of accurately predicting parasite B- and T-cell epitopes, as well as subcellular localization of parasites proteins (Figure 1). The interface language used was standard SQL (Standard Query Language) and several built-in functions were implemented, but are not limited to, the following: a) parse algorithms predictions and storage of experimental validated and predicted data; and b) evaluation of algorithm performances.

Results

MHC-I epitopes prediction

Several approaches that predict peptide binding to MHC molecules have been published [22]. In this study we opt for two currently available algorithms, NetCTL [22] and NetMHC [23]. Our choice for these MHC class I epitope prediction algorithms was made in terms of ISI citation indexes and regarding their availability for download and local machine implementation.

When possible, in order to establish the ideal settings for protozoan epitope prediction, the algorithms parameters were scanned and evaluated in terms of AUC values. In this framework, the NetCTL score threshold parameter was ranged from 0.50 to 0.90. The NetCTL and NetMHC algorithms outputs were parsed and the data of 3,906 *in silico* predicted epitopes loaded into MySQL database. In order to evaluate the algorithm performances, predicted epitopes [see Additional File 1 and Additional File 2] were aligned against the consensus experimentally validated dataset for MHC class I epitopes. Figure 2 presents an overview of the benchmark approach undertaken in

this study (see Methods Section for more details). In addition, we carried out a combined performance analysis using the best score threshold found for each methodology (Table 1).

The AUC performance measure obtained for NetCTL was 0.66 (for a score threshold of 0.50) and for NetMHC was 0.60 (score thresholds cannot be modified by the user). On the other hand, the combined performance of these algorithms produced an AUC value of 0.64 (Table 1, Figure 3).

B-cell epitopes prediction

Following the same rational described above for MHC I algorithm selection, three currently available algorithms were chosen, BepiPred [24], BCPred12 [25] and AAP12 [26]. When possible, in order to establish the ideal settings for protozoan epitope prediction, the algorithms parameters were scanned and then evaluated in terms of AUC values. In this framework, the score thresholds parameter ranged from 0.15 to 0.90 for BepiPred and from 0.50 to 0.90 for AAP12 and BCPred12.

Using the developed pipeline, the default algorithms outputs were parsed and the data of 187,187 *in silico* predicted epitopes [see Additional File 3, Additional File 4 and Additional File 5] loaded into the MySQL database. In order to evaluate the algorithm performances, predicted epitopes were aligned against the consensus experimentally validated dataset for B cell epitopes. Furthermore, we carried out a combined performance analysis using the best score threshold found for each methodology (Table 1, Figure 4).

The AUC performance measure obtained was: 0.53 for BepiPred using a threshold of 0.40; 0.52 for AAP12 using a threshold of 0.80; and 0.62 for BCPred12 using a threshold of 0.90 (Table 1). Regarding the combined performance analysis performed for these algorithms, the following results were found: 0.77 for AAP12 and BCPred12; 0.49 for AAP12 and BepiPred; 0.58 for BCPred12 and BepiPred; and 0.57 for AAP12, BCPred12 and BepiPred.

Subcellular localization of proteins prediction

Regarding prediction of subcellular localization of proteins, three currently available algorithms were selected, WoLF PSORT [27], Sigcleave [28] and TargetP [11]. Using the developed pipeline, the default algorithms outputs were parsed and the data of 538 *in silico* predictions loaded into the MySQL database. In order to evaluate the algorithms performances, an experimental validated dataset of 180 proteins with described subcellular localization was loaded in addition to *in silico* predictions. Results show that WoLF PSORT was capable of correctly predicting 27/44 (61.36%) secreted proteins, Sigcleave, 30/44 (68.18%), and TargetP, 32/44 (72.73%), showing that the proportion of correctly predicted binders (sensitivity) was similar between the three algorithms (Table 2). Files containing predictions made by each algorithm are available as Additional Files [see for WoLF PSORT, Additional File 6; for Sigcleave, Additional File 7; for TargetP, Additional File 8].

The evaluation of the intersecting portion of predictions made by the tested algorithms showed that, from 40 protozoan proteins with extracellular localization experimentally determined, 19 (~48%) were correctly predicted by all three algorithms (Figure 5).

Discussion

Despite of being a major public health problem in several countries, the life-threatening diseases caused by protozoan parasites represent a challenge in terms of vaccine development and nowadays there is no efficient vaccine against these parasites.

Epitope prediction by computational methods represents one of the most promising approaches to vaccine development, but there are several drawbacks in the process regarding trypanosomatid genomes. In this context, the lack of sufficiently large datasets of experimentally validated protozoan epitopes represents a serious limitation for validation of parasite *in silico* epitope prediction.

Several prediction methods were developed, but none of them had protozoan parasites data as training dataset (for some of them, protozoan parasites proteins represent only about 10% of the training dataset [24, 25, 29-32]) and, consequently, these results can be biased and should be treated with a grain of salt. The general wisdom is that the performance of epitope prediction methods critically depends on the dataset used for training and also on protein compositional bias. In addition, it is influenced by the evaluation criteria. Regarding epitope prediction in parasite genomes, these drawbacks are noteworthy considering that these organisms have a genome content that reflects proteins with a particular physicochemical profile and that are underrepresented in training datasets.

For this reason, we do not try to rank various prediction methods. Rather, we focus on the key concepts and ideas in the field. Thus, we evaluated algorithm performances focusing on parasites genomes. Comparison between algorithms was made in the basis of AUC (area under a ROC curve) values, which represent the probability that a randomly selected positive instance will score higher than a randomly selected negative instance [33].

Aiming at identifying a good set of tools for protozoan parasites epitope prediction and subcellular localization of proteins, we developed, in this work, a database approach in order to integrate and evaluate the combined performances of some open source currently available algorithms for MHC class I and B-cell epitope prediction, as well as for subcellular localization using protozoan parasites proteins and epitopes experimentally identified.

Concerning the epitope prediction, a database schema was developed and implemented integrating experimental validated data together with the information related to MHC I prediction (NetCTL and NetMHC algorithms) and B-cells prediction (BepiPred, AAP12 and BCPred12 algorithms).

The main source of experimental data was "Immune Epitope Database and Analysis Resource" (IEDB) (http://www.immuneepitope.org/) [34], that currently represents the main source of linear and conformational epitopes data. Besides, IEDB uses a metric that takes into account the number of references, number of positive assays, and total number of assays for each epitope which is crucial to extract an

experimentally validated epitope subset with a high level of confidence for the benchmark.

Regarding MHC I prediction, our AUC results indicate a little difference in the performances related with NetCTL and NetMHC algorithms, 0.66 and 0.60 respectively. If we consider that it is reported that the MHC class I prediction methods have achieved an accuracy that in many cases allows for AUC values in the range 0.95-0.99 [22], both algorithms didn't achieve the expected performance. In fact, this is not the first time that underperformance of prediction algorithms is reported in literature. In a recent study, 167 9mer peptides from *Influenza A virus* were predicted as potential binders by NetMHC, and just 89 of them (53% of the pool) were confirmed as real binders [35]. Furthermore, the underrepresentation of protozoan proteins in the training datasets in general and the compositional bias certainly have a deep impact on epitope prediction methods and also in the benchmark. In fact, to highlight the different performances of tested algorithms in front of different datasets and exclude the influence of approach undertaken, we evaluated the algorithm performances under the same framework but with the human proteins dataset available for download from NetCTL website [36-38]. The results for both NetCTL and NetMHC algorithms were considerably better than the results obtained for protozoan dataset. The AUC value for NetCTL was 0.80 and for NetMHC was 0.77 (Figure 3). In addition, our performance evaluation does not include MHCII prediction since experimental data was insufficiently represented (data not shown). In practice, the prediction of MHC-peptide binding is far from perfect, but this fact does not preclude all the advances made in the last years in the field [39].

Regarding B-cell epitope prediction, our AUC results indicate a better performance for BCPred12 algorithms when compared to AAP12 and BepiPred (Table 1). Again the observed performances were inferior from those currently observed for B-cell epitope predictions [24]. This difference might be explained by same reasons which were just discussed for MHCI prediction. Also for B-cell epitope prediction, this is not the first report in literature of low epitope prediction performance [40].

Lafuente and Reche (2009) believe launching a Critical Assesment of Techniques for Epitope Prediction will benefit the field. Under this program, computational methods will be used for blind *de novo* prediction of peptides that are immunogenic from query proteins that, for evaluation purposes, has been experimentally screened [39]. Considering that and the results obtained by us, we do believe this approach will be useful to bring advances to epitope prediction area.

Despite of the shortcomings cited above, the combined performance analysis seems to be a promising approach. For B-cell algorithms, when the combined performance analysis was made, the best combination performance was found for AAP12 and BCPred12 that reached an AUC value of 0.77, which is within the expect range reported [24].

Seen in the light of the results obtained, the developed approach calls attention to several points: a) The general prediction models used by currently available algorithms cannot be used with the same performance for different protein subsets (especially true for protozoan parasites); b) The need for studies in which the algorithm performances are evaluated for underrepresented and compositional biased

proteins subsets; and c) The combinatorial prediction approach can improve the epitope prediction performance.

Concerning the subcellular localization prediction, the database schema developed also integrated experimental and predicted data for subcellular localization of proteins. Experimental data was obtained from UniProt (http://www.uniprot.org), and the *in silico* predictions made by WoLF PSORT, Sigcleave and TargetP algorithms. The result shows that there is not much difference, in terms of percentage of matches, between the tested algorithms. Nevertheless, the Venn diagram analysis related to true positives (extracellular localization) result shows that the tested algorithms match different proteins in the dataset, and the consensus prediction of the three algorithms would better define a protein located in the extracellular compartment.

Conclusions

Considering the public health importance of the studied organisms and the lack of studies specifically addressing epitope and subcellular localization prediction in these parasites, our results suggest that the algorithm combinatorial approach employed in the developed database-driven epitope prediction methodology is capable of proposing the best set of tools for *in silico* epitope prediction in protozoan parasite genomes. Several drawbacks exist, but the present work will certainly speed up the process of data mining analysis and prediction of potential candidates for vaccine development.

Methods

Databases of experimentally tested epitopes

Two datasets of experimentally tested epitopes were built, one of B-cell epitopes and another of MHC-I epitopes. Parasite proteins experimental datasets where extracted from Immune Epitope Database and Analysis Resource (IEDB) [34, 41, 42]. The following criteria were adopted for epitope selection: a) from protozoan parasites; b) host organism must be mice or human. Selected epitopes were minimal epitopes, experimentally validated as immunogenic [see Additional File 9 and Additional File 10, for B and T-cell epitopes, respectively] or non-immunogenic [see Additional File 11 and Additional File 12, for B and T-cell non-immunogenic regions, respectively].

Furthermore, several overlapping epitopes anchored in the same protein region were observed. In order to have a non-redundant set of experimentally validated regions for each protein from dataset, make sense for us to use what we call "the consensus validated region", which consists to cluster the overlapping epitopes in a unique consensus region called "experimentally validated consensus region".

Thus, the dataset of B-cell experimental epitopes ended up with 312 proteins and 866 experimentally validated consensus regions including immunogenic and non-immunogenic [see Additional File 13 and Additional File 14, respectively]; for MHC-I epitopes, 81 proteins and 224 experimentally validated consensus regions including immunogenic and non-immunogenic assignments [see Additional File 15 and Additional File 16, respectively]. Furthermore, these data were used as input for the formatdb program (BLAST package) which prepares the sequences to be aligned as subject by the BLAST algorithm [43].

Database of proteins with experimentally validated subcellular localization

Proteins with experimentally validated subcellular localization were obtained from UniProt. The search was done with the term "trypanosomatidae", with the field "subcellular location" set to the confidence "experimental", which retrieved 180 proteins with subcellular localization described experimentally [see Additional File 17 and Additional File 18]. This dataset was used to evaluate the three selected algorithms for subcellular localization prediction.

Selection of prediction tools

The prediction algorithms were selected taking into account the possibility of being installed locally, and the reliability of their predictions reported on literature. The predictions of the following algorithms were evaluated: a) for MHCI epitope prediction: NetCTL [22, 32, 44] and NetMHC [23, 31, 45]; b) for B-cell epitope prediction: BepiPred [24] and BCPreds [25, 26, 29], which included two methodologies, AAP12 and BCPred12; c) for protein subcellular localization prediction, WoLF PSORT [27], Sigcleave [28] and TargetP [11].

Score thresholds and allele used

Score thresholds used for CD8+ T cell epitope predictors ranged from 0.50 to 0.90 for NetCTL. We did not set the threshold value for NetMHC because this parameter is not variable in the NetMHC command line mode, so a unique value (0.426) was used for NetMHC. Concerning B cell epitope predictors, the score thresholds ranged from 0.50 to 0.90 for AAP12 and BCPred12 and from 0.15 to 0.90 for BepiPred.

For CD8+ T cell epitopes prediction, the human supertype A2 was the allele model used to scan MHC binding affinity. Other HLA alleles are present in IEDB, but

regarding the protozoan proteins they are underrepresented in the non-redundant database used (see previous section). In addition, the supertype HLA-A2 is included in a group which is expressed in 88% of the population, what illustrates the relevance of the supertype used in this work.

Development of parsers and algorithms

Parsers and algorithms were developed in PERL and SQL languages in order to extract the results obtained after running the programs and help to integrate all the results in the relational database.

Construction of relational database

To aggregate all information generated during the development of the project we used MySQL as a Relational Database Management System (RDBMS) (http://www.mysql.com). The use of a database system in this work represents a manner of getting a data receptacle or conceptual repository from which was possible to extract data correlation and results. The MySQL GUI Tools (http://dev.mysql.com/downloads/gui-tools/5.0.html) were used as a graphical user interface for our MySQL database. The entity-relational model (ERM) was built using MySQL Workbench (http://wb.mysql.com). For automatic data parsing and to load information into database, Perl scripts using DBI and BioPerl modules were developed. An overview of implemented workflow is presented in Figure 1.

Analysis of the epitopes results

The developed methodology for result analyses was based on threshold dependent parameters and also on AUC performance analyses. After the identification of the regions of the source proteins which were assigned as consensus experimentally validated regions (immunogenic or non-immunogenic), and with the results of the predictions made by the tested algorithms stored on the constructed relational database, we identified the True Positives (TP) and False Positives (FP) hits. This information was employed in the AUC performance analysis.

In our methodology, we classified the predicted epitopes as TP or FP using the results produced from BLAST algorithm [43]. The following parameters were utilized to decide if a prediction was going to be considered as TP: 1) the local alignment between an epitope prediction (query) and an immunogenic consensus experimentally validated region (subject) had to have minimal query coverage (50% for B cell prediction and 87% for CD8+ T cell). In addition, the coverage cutoff was established based on a minimum size of B cell and CD8+ T cell found in the experimental epitopes database used (IEDB) which are 6 and 8 amino acids respectively. The NetCTL and NetMHC algorithms predict epitopes with 9 amino acids, thus 87% of coverage guarantees the minimum epitope alignment size of 8 amino acids. By the other hand, the AAP12 and BCPred12 algorithms predict epitopes with 12 amino acids, therefore 50% of coverage guarantees the minimum epitope alignment size of 6 amino acids. Regarding the BepiPred algorithm, since it predicts epitopes with variable sizes, only those with at least 6 amino acids were considered in the analysis. Specifically for predictions ranging from 6 to 11 amino acids, the coverage cutoff varied to guarantee a minimal amino acid alignment length of 6 residues. 2) The local alignment between an epitope prediction and a consensus experimentally validated

region (immunogenic) had to have 100% of identity. This parameter was used to confirm that a given subject extracted from an alignment exactly matches with the real query. 3) Finally, in order to guarantee subject and query reciprocity the query name and the subject name must be the same. Using the same rational a prediction was considered FP, but the alignment analyses were made using as subject the non-immunogenic consensus experimentally validated region. Predicted epitopes that did not align with the parameters cited just above or if they aligned with both immunogenic and non-immunogenic consensus experimentally validated region were not considered for further analysis.

Algorithms combined predictions

To perform the combined prediction we adopted the following rational: 1) for a given protein, the experimental regions are indexed and so, considering a protein (P) with three experimental validated regions, they would be named P1, P2 and P3; 2) if a given algorithm A predicts an epitope that matches with P2 for instance and another algorithm B predicts an epitope that also matches with P2 they are considered a combined prediction; 3) if a given algorithm A predicts an epitope that matches with P2 and another algorithm B predicts an epitope that matches with P1 or P3 they are not considered as a combined prediction.

Based on the above rules, the combined prediction score was calculated as the mean of the individual normalized scores of the original predictions. The score normalization was done as follows:

NS=(PS-MLS)/((MHS-MLS)/100), where:

NS = normalized score;
PS = prediction score;
MLS = methodology lowest score;
MHS = methodology highest score.

Accuracy evaluation

A non-parametric performance measure was used to avoid the influence of arbitrary thresholds. In order to carry out an accuracy evaluation, we used the area under the ROC curve, or simply AUC, that aggregates the model's behavior for all possible decision thresholds. The nonparametric estimate of the AUC [46] was calculated through an implemented GNU R package called ROCR [33].

Analysis of the subcellular localization results

In order to analyze the subcellular localization prediction results, we determined the True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). This information was incorporated into a confusing error matrix that allowed the determination of parameters: sensitivity (Sn), specificity (Sp), positive predictive value (PPV), negative predictive value (NPV) and accuracy (Table 3).

Authors' Contributions

DMR obtained the experimental tested epitopes and source proteins, selected the algorithms to be tested, run the selected algorithms, participated in the analysis of the results, and wrote the manuscript. AMR developed the algorithms for prediction performance analyses, participated in the analysis of the results and in the writing process of the manuscript. NJDO constructed and populated the relational database, developed algorithms in SQL language and participated in the analysis of the results; DMR, AMR and JCR developed parsers in PERL language that made possible results

extraction and load the data into the relational database. ICAB monitored the analysis of the selected algorithms. ABR was involved with the conceptual design the study and RCO provided useful discussions for this work. JCR was involved in the drafting process of the manuscript and in the experimental and conceptual design of the study. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Patrícia Cássia Ruy and Raul Torrieri for the comments given at the beginning of this work.

Instituto René Rachou (IRR/FIOCRUZ Minas), Universidade Federal de Ouro Preto (UFOP), FAPEMIG (PRONEX 503/07 and PPP APQ-04554-10), CNPq (GENOPROT 560943/2010-5 and Universal 478100/2011-6) and CAPES (PNPD 2009) supported this work. JCR is supportted by the following grants: CNPq 476898/2008-0 and 476539/2010-2; FAPEMIG APQ-02382-10 and APQ-01835-10. Fellowships were provided by CNPq to ABR, RCO, RT and PCR, by CAPES to DMR and by FAPEMIG to ICAB.

References

1. Bambini S, Rappuoli R: **The use of genomics in microbial vaccine development**. *Drug DiscovToday* 2009, **14**(5-6):252-260.
2. Rinaudo CD, Telford JL, Rappuoli R, Seib KL: **Vaccinology in the genome era**. *JClinInvest* 2009, **119**(9):2515-2525.
3. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B *et al*: **Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing**. *Science* 2000, **287**(5459):1816-1820.

4. Andre FE: **Overview of a 5-year clinical experience with a yeast-derived hepatitis B vaccine**. *Vaccine* 1990, **8 Suppl**:S74-S78.

5. Korber B, LaBute M, Yusim K: **Immunoinformatics comes of age**. *PLoSComputBiol* 2006, **2**(6):e71.

6. Borja-Cabrera GP, Cruz Mendes A, Paraguai de Souza E, Hashimoto Okada LY, de ATFA, Kawasaki JK, Costa AC, Reis AB, Genaro O, Batista LM *et al*: **Effective immunotherapy against canine visceral leishmaniasis with the FML-vaccine**. *Vaccine* 2004, **22**(17-18):2234-2243.

7. Emanuelsson O, von Heijne G: **Prediction of organellar targeting signals**. *BiochimBiophysActa* 2001, **1541**(1-2):114-119.

8. Feng ZP: **An overview on predicting the subcellular location of a protein**. *In SilicoBiol* 2002, **2**(3):291-303.

9. Bannai H, Tamada Y, Maruyama O, Nakai K, Miyano S: **Extensive feature detection of N-terminal protein sorting signals**. *Bioinformatics* 2002, **18**(2):298-305.

10. Cedano J, Aloy P, Perez-Pons JA, Querol E: **Relation between amino acid composition and cellular location of proteins**. *JMolBiol* 1997, **266**(3):594-600.

11. Emanuelsson O, Nielsen H, Brunak S, von Heijne G: **Predicting subcellular localization of proteins based on their N-terminal amino acid sequence**. *JMolBiol* 2000, **300**(4):1005-1016.

12. Hua S, Sun Z: **Support vector machine approach for protein subcellular localization prediction**. *Bioinformatics* 2001, **17**(8):721-728.

13. Baldi P, Atiya AF: **How delays affect neural dynamics and learning**. *IEEE TransNeural Netw* 1994, **5**(4):612-621.

14. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: extension and analysis of the basic method**. *ComputApplBiosci* 1996, **12**(2):95-107.

15. Vapnik V, Vashist A: **A new learning paradigm: learning using privileged information**. *Neural Netw* 2009, **22**(5-6):544-557.

16. Lundegaard C, Lund O, Kesmir C, Brunak S, Nielsen M: **Modeling the adaptive immune system: predictions and simulations**. *Bioinformatics* 2007, **23**(24):3265-3275.

17. Mamitsuka H: **Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models**. *Proteins* 1998, **33**(4):460-474.

18. Shekhar S, Chawla S: **Spatial Databases: A Tour**: Prentice-Hall; 2002.

19. Elmasri RA, Navathe SB: **Fundamentals of Databases Systems**: Addision-Wesley Pubishing; 2000.

20. Markowitz VM: **Biological Data Management in a Dataspace Framework**: Biological Data Management and Technology Center, Lawrence Berkeley National Laboratory; 2000.

21. Ostell JM, Wheelan SJ, Kans JA: **The NCBI Data Model in Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins**: John Wiley & Sons Publishing; 2001.

22. Larsen MV, Lundegaard C, Lamberth K, Buus S, Lund O, Nielsen M: **Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction**. *BMC Bioinformatics* 2007, **8**:424.

23. Nielsen M, Lundegaard C, Worning P, Hvid CS, Lamberth K, Buus S, Brunak S, Lund O: **Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach**. *Bioinformatics* 2004, **20**(9):1388-1397.

24. Larsen JE, Lund O, Nielsen M: **Improved method for predicting linear B-cell epitopes**. *ImmunomeRes* 2006, **2**:2.

25. El Manzalawy Y, Dobbs D, Honavar V: **Predicting linear B-cell epitopes using string kernels**. *JMolRecognit* 2008, **21**(4):243-255.

26. Chen J, Liu H, Yang J, Chou KC: **Prediction of linear B-cell epitopes using amino acid pair antigenicity scale**. *AminoAcids* 2007, **33**(3):423-428.

27. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W585-W587.

28. von Heijne G: **A new method for predicting signal sequence cleavage sites**. *Nucleic Acids Res* 1986, **14**(11):4683-4690.

29. El Manzalawy Y, Dobbs D, Honavar V: **Predicting flexible length linear B-cell epitopes**. *ComputSystBioinformatics Conf* 2008, **7**:121-132.

30. Nielsen M, Lund O: **NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction**. *BMC Bioinformatics* 2009, **10**:296.

31. Nielsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O: **Reliable prediction of T-cell epitopes using neural networks with novel sequence representations**. *Protein Sci* 2003, **12**(5):1007-1017.

32. Peters B, Bulik S, Tampe R, Van Endert PM, Holzhutter HG: **Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors**. *JImmunol* 2003, **171**(4):1741-1749.

33. Wu S, Flach P: **A scored AUC metric for classifier evaluation and selection**. In: *ROCML workshop at ICML.* Citeseer; 2005.

34. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B: **The immune epitope database 2.0**. *Nucleic Acids Res*, **38**(Database issue):D854-D862.

35. Wang M, Lamberth K, Harndahl M, Roder G, Stryhn A, Larsen MV, Nielsen M, Lundegaard C, Tang ST, Dziegiel MH *et al*: **CTL epitopes for influenza A including the H5N1 bird flu; genome-, pathogen-, and HLA-wide screening**. *Vaccine* 2007, **25**(15):2823-2831.

36. **NetCTL** [http://www.cbs.dtu.dk/suppl/immunology/CTL-1.2.php]

37. **NetCTL SYFPEITHI dataset** [http://www.cbs.dtu.dk/suppl/immunology/CTL-1.2/syf.data.fsa]

38. Larsen MV, Lundegaard C, Lamberth K, Buus S, Brunak S, Lund O, Nielsen M: **An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions**. *Eur J Immunol* 2005, **35**(8):2295-2303.

39. Ostell JM, Kans JA: **The NCBI data model**. *Bioinformatics* 1998:121-144.

40. Sebatjane S, Pretorius A, Liebenberg J, Steyn H, Van Kleef M: **In vitro and in vivo evaluation of five low molecular weight proteins of Ehrlichia ruminantium as potential vaccine components**. *Veterinary Immunology and Immunopathology* 2010, **137**(3-4):217-225.

41. Vita R, Peters B, Sette A: **The curation guidelines of the immune epitope database and analysis resource**. *Cytometry A* 2008, **73**(11):1066-1070.

42. Vita R, Vaughan K, Zarebski L, Salimi N, Fleri W, Grey H, Sathiamurthy M, Mokili J, Bui HH, Bourne PE *et al*: **Curation of complex, context-dependent immunological data**. *BMC Bioinformatics* 2006, **7**:341.

43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *JMolBiol* 1990, **215**(3):403-410.

44. Nielsen M, Lundegaard C, Lund O, Kesmir C: **The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage**. *Immunogenetics* 2005, **57**(1-2):33-41.

45. Buus S, Lauemoller SL, Worning P, Kesmir C, Frimurer T, Corbet S, Fomsgaard A, Hilden J, Holm A, Brunak S: **Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach**. *Tissue Antigens* 2003, **62**(5):378-384.

46. Sing T, Sander O, Beerenwinkel N, Lengauer T: **ROCR: visualizing classifier performance in R**. *Bioinformatics* 2005, **21**(20):3940-3941.

FIGURES

Figure 1 – The methodology flowchart used to develop this work. A - Construction of the experimentally tested epitopes database, obtained on the curated Immune Epitope Database and Analysis Resource (IEDB) and on Uniprot. B - Algorithms used to predict B-cell epitopes, MHC-I epitopes, MHC-II epitopes and subcellular location of proteins. C - Parsers developed to extract the results. D - Construction of the relational database to integrate the results. E - Analysis of the results in the framework of AUC and confusion error matrix. PERL, Practical and Extraction Report Language; SQL, Structured Query Language.

Figure 2 – Strategy employed to assess the predictors' performance. The scale bar represents the sequence of a theoretical protein of 120 amino acids. Dark blue rectangles represent one single epitope or a consensus of overlapping epitopes that were experimentally validated according to IEDB (Positive exp.); light blue rectangles represent one single non-immunogenic region or a consensus of overlapping non-immunogenic regions that were experimentally validated according to IEDB (Negative exp.); red rectangles represent the predicted epitopes from evaluated algorithms. For B cell prediction the predicted epitopes were considered true positive if they aligned with minimum coverage of 50% and 100% of identity with a Positive Exp. Region; for CD8+ T cell prediction the predicted epitopes were considered true positive if they aligned with minimum coverage of 87% and 100% of identity with a Positive Exp. region (Prediction 1, Prediction 2, Prediction 3 and Prediction 11). For B cell prediction the predicted epitopes were considered false

positive if they aligned with minimum coverage of 50% and 100% of identity with a Negative Exp. Region; for CD8+ T cell prediction the predicted epitopes were considered false positive if they aligned with minimum coverage of 87% and 100% of identity with a Negative Exp. region (Prediction 6, Prediction 7, Prediction 8 and Prediction 13). Predictions were not considered during the analysis if they did not align with the parameters cited above (Prediction 4, Prediction 5, Prediction 9 and Prediction 10) or if they aligned with both Positive exp. and Negative exp. (Prediction 12).

Figure 3 – ROC (Received Operating Characteristics) curve representing the performance of CD8+ T cells epitope predictors analyzed (NetCTL and NetMHC), from protozoan proteins database, and their combination. In addition, the black and grey curves represent the performance of the same algorithms for human proteins database (SYFPEITHI) extracted from NetCTL homepage (http://www.cbs.dtu.dk/suppl/immunology/CTL-1.2/syf.data.fsa).

Figure 4 – ROC (Received Operating Characteristics) curve representing the performance of B cells epitope predictors analyzed (AAP12, BCPred12 and BepiPred), from protozoan proteins database, and their combination.

Figure 5 – Venn diagram showing the evaluation of the intersecting portion of predictions made by the tested algorithms for subcellular localization of proteins. The evaluation considered just the proteins experimentally determined as extracellular (n=40).

TABLES

Table 1 – Algorithms performance evaluation for B cell epitope prediction and CD8+ T cells epitope prediction.

Algorithms	AUC values
CD8+ T cells epitope prediction	
NetCTL	0.66
NetMHC	0.60
NetCTL and NetMHC	0.64
B cell epitope prediction	
AAP12	0.52
BCPred12	0.62
BepiPred	0.53
AAP12 and BCPred12	0.77
AAP12 and BepiPred	0.49
BCPred12 and BepiPred	0.58
AAP12, BCPred12 and BepiPred	0.57

Table 2 – Analyzed parameters for subcellular location of proteins.

	Subcellular location of proteins		
	WoLF PSORT	Sigcleave	TargetP
Sensitivity	61.36% (27/44)	68.18% (30/44)	72.73% (32/44)
Specificity	93.38% (127/136)	76.47% (104/136)	79.41% (108/136)
PPV	74.29% (26/35)	48.39% (30/62)	53.33% (32/60)
NPV	88.11% (126/143)	88.14% (104/118)	90% (108/120)
Accuracy	85.39% (152/178)	74.44% (134/180)	77.78% (140/180)

Table 3 – Parameters used in the analysis of the results.

Parameter	Brief description	Formula
Sensitivity (Sn)	The proportion of correctly predicted binders	(a/(a + c))*100

Specificity (Sp)	The proportion of correctly predicted non-binders	(d/(b + d))*100
Positive Predictive Value (PPV)	The probability that a predicted binder will actually be a binder	(a/(a + b))*100
Negative Predictive Value (NPV)	The probability that a predicted non-binder will actually be a non-binder	(d/(c + d))*100
Accuracy	The proportion of correctly predicted peptides (both binders and non-binders)	((a + d)/(a+b+c+d)*100

a = True positive observations

b = False positive observations

c = False negative observations

d = True negative observations

ADDITIONAL FILES

Additional File 1 (*.txt)

Title: Epitopes predicted by NetCTL.

Description: This file in fasta format contains 2,657 epitopes predicted by NetCTL.

Additional File 2 (*.txt)

Title: Epitopes predicted by NetMHC.

Description: This file in fasta format contains 1,249 epitopes predicted by NetCTL.

Additional File 3 (*.txt)

Title: Epitopes predicted by BepiPred.

Description: This file in fasta format contains 5,450 epitopes predicted by BepiPred.

Additional File 4 (*.txt)

Title: Epitopes predicted by AAP12.

Description: This file in fasta format contains 138,987 epitopes predicted by AAP12.

Additional File 5 (*.txt)

Title: Epitopes predicted by BCPred12

Description: This file in fasta format contains 42,750 epitopes predicted by BCPred12.

Additional File 6 (*.txt)

Title: Predictions made by WoLF PSORT.

Description: This file contains predictions made by WoLF PSORT.

Additional File 7 (*.txt)

Title: Predictions made by Predictions made by Sigcleave.

Description: This file contains predictions made by Sigcleave.

Additional File 8 (*.txt)

Title: Predictions made by TargetP.

Description: This file contains predictions made by TargetP.

Additional File 9 (*.txt)

Title: B-cell minimal epitopes experimentally validated extracted from IEDB

Description: This file in fasta format contains 3,021 B-cell minimal epitopes from parasite proteins experimentally validated as immunogenic extracted from IEDB.

Additional File 10 (*.txt)

Title: CD8+ T cell minimal epitopes experimentally validated extracted from IEDB

Description: This file in fasta format contains 228 CD8+ T cell minimal epitopes from parasite proteins experimentally validated as immunogenic extracted from IEDB.

Additional File 11 (*.txt)

Title: B-cell non-immunogenic regions experimentally validated extracted from IEDB

Description: This file in fasta format contains 3,039 B-cell non-immunogenic regions from parasite proteins experimentally validated extracted from IEDB.

Additional File 12 (*.txt)

Title: CD8+ T cell non-immunogenic regions experimentally validated extracted from IEDB

Description: This file in fasta format contains 166 CD8+ T cell non-immunogenic regions from parasite proteins experimentally validated extracted from IEDB.

Additional File 13 (*.txt)

Title: B-cell immunogenic consensus regions experimentally validated

Description: This file in fasta format contains 607 B-cell immunogenic consensus regions from parasite proteins experimentally validated.

Additional File 14 (*.txt)

Title: B-cell non-immunogenic consensus regions experimentally validated

Description: This file in fasta format contains 243 B-cell non-immunogenic consensus regions from parasite proteins experimentally validated.

Additional File 15 (*.txt)

Title: CD8+ T cell immunogenic consensus regions experimentally validated

Description: This file in fasta format contains 140 CD8+ T cell immunogenic consensus regions from parasite proteins experimentally validated.

Additional File 16 (*.txt)

Title: CD8+ T cell non-immunogenic consensus regions experimentally validated

Description: This file in fasta format contains 84 CD8+ T cell cell non-immunogenic consensus regions from parasite proteins experimentally validated.

Additional File 17 (*.txt)

Title: Trypanosomatid proteins with experimentally validated subcellular localization extracted from Unitprot

Description: This file contains a list of 180 trypanosomatid proteins with its subcellular localization experimentally validated extracted from Uniprot.

Additional File 18 (*.txt)

Title: Trypanosomatid proteins in fasta format with experimentally validated subcellular localization

Description: This file in fasta format contains 180 trypanosomatid proteins with its subcellular localization experimentally validated.

9.2 – Anexo II

ANEXO II - Análises das 20 proteínas mais bem ranqueadas para cada espécie de leishmania presente no trabalho

Proteína	MCC	Degree	CLF	MMU	HSA	Grupo Ortólogo	Average_identity	BCPred12	NetCTL	NetMHCII
LbrM31_V2.0300	3623677	94				OG5_126796	73.00%	116	12	5
LbrM21_V2.2070	2824607	126				OG5_126700	90.00%	45	12	2
LbrM20_V2.5830	2766959	110				OG5_127903	77.00%	58	12	7
LbrM26_V2.1840	2415074	83				OG5_128124	74.00%	57	12	3
LbrM33_V2.2140	2161529	76				OG5_127411	86.00%	121	12	3
LbrM26_V2.0980	2032439	83				OG5_127754	75.00%	55	12	3
LbrM35_V2.4900	1911757	97				OG5_127990	80.00%	13	12	2
LbrM20_V2.0580	1903166	83				OG5_127757	81.00%	24	12	4
LbrM09_V2.0920	1797962	90				OG5_127677	68.00%	51	12	4
LbrM32_V2.3060	1731661	61				OG5_127957	79.00%	36	12	3
LbrM34_V2.3470	1599336	74				OG5_127464	84.00%	11	11	2
LbrM17_V2.1410	1506901	92				OG5_127380	73.00%	87	12	7
LbrM02_V2.0400	1418627	40				OG5_127922	72.00%	83	12	3
LbrM26_V2.0030	1386992	73				OG5_127909	74.00%	58	12	2
LbrM30_V2.3680	1382873	68				OG5_127033	94.00%	36	12	3
LbrM22_V2.0510	1331220	24	1	1	1	OG5_128056	89.00%	62	12	2
LbrM34_V2.3820	1328258	34				OG5_127705	89.00%	48	11	3
LbrM34_V2.2090	1258905	64				OG5_127245	90.00%	38	12	1
LbrM21_V2.2010	1202576	34				OG5_127791	85.00%	33	12	1
LbrM07_V2.1200	1146835	42				OG5_127854	65.00%	37	12	3
LinJ34_V3.0670	33052390694	91				OG5_127757	81.00%	19	12	3
LinJ21_V3.2070	33046922208	35				OG5_127791	85.00%	32	12	1
LinJ06_V3.0140	33016962826	32				OG5_127834	78.00%	23	12	1
LinJ36_V3.0340	32891974150	50				OG5_127700	78.00%	32	12	1
LinJ34_V3.4390	32877190946	41				OG5_127562	87.00%	5	12	3
LinJ11_V3.0240	32563302555	41				OG5_127424	83.00%	26	12	1
LinJ32_V3.2960	30957510497	58				OG5_127957	79.00%	25	12	3
LinJ21_V3.2200	29733340525	37				OG5_127593	83.00%	39	11	2
LinJ32_V3.0400	29666718150	34				OG5_127933	76.00%	67	12	4
LinJ02_V3.0340	26526043326	54				OG5_127922	72.00%	71	12	2
LinJ36_V3.1730	25185041573	33	1	1	1	OG5_127100	81.00%	40	12	2
LinJ22_V3.0490	21249059971	30	1	1	1	OG5_127786	81.00%	37	12	1
LinJ21_V3.0840	20273109403	47				OG5_127889	76.00%	48	12	5
LinJ27_V3.1370	16699368581	36				OG5_128135	82.00%	17	12	3
LinJ32_V3.1260	12546787626	58				OG5_128073	73.00%	12	12	5
LinJ28_V3.0110	10084774961	31				OG5_127761	88.00%	13	12	1
LinJ36_V3.1670	6017388278	32				OG5_127884	82.00%	25	12	2
LinJ14_V3.0310	4882401966	28				OG5_127759	78.00%	51	12	1
LinJ35_V3.3880	4748881680	31				OG5_127705	89.00%	45	11	4
LinJ12_v4.0030	3403026437	35				OG5_127762	82.00%	22	12	0
LmjF32.1200	32217730	54				OG5_128073	73.00%	8	12	5
LmjF32.0390	32190849	33				OG5_127933	76.00%	80	12	4
LmjF32.2820	32183892	58				OG5_127957	79.00%	33	12	3
LmjF21.1700	27799950	46				OG5_127791	85.00%	37	12	1
LmjF35.3840	27667880	34				OG5_127705	89.00%	38	11	3
LmjF34.4520	26008612	38				OG5_127562	87.00%	12	12	3
LmjF27.1460	22070715	25				OG5_128135	82.00%	17	12	3
LmjF02.0370	21344704	47				OG5_127922	72.00%	77	12	2
LmjF36.0320	19956093	48				OG5_127700	78.00%	26	12	1
LmjF36.1650	17725710	36	1	1	1	OG5_127100	81.00%	38	12	2
LmjF36.1600	17401520	27				OG5_127884	82.00%	28	12	2
LmjF28.1730	12501630	33				OG5_127160	73.00%	128	12	6
LmjF21.0760	12087106	31				OG5_127889	76.00%	42	12	4
LmjF26.0180	10919813	83				OG5_126953	88.00%	51	12	5
LmjF06.0140	10094162	35				OG5_127834	78.00%	14	12	2
LmjF35.4850	8966995	32				OG5_127623	79.00%	45	12	2
LmjF35.0600	8434330	47				OG5_127130	91.00%	11	11	4
LmjF35.1890	7977876	85				OG5_127090	88.00%	48	12	4
LmjF30.3650	7854772	99				OG5_127033	94.00%	34	12	3
LmjF33.1350	7365193	94				OG5_127869	88.00%	18	12	1