Ministério da Saúde
Fundação Oswaldo Cruz
Centro de Pesquisas René Rachou
Programa de Pós-graduação em Ciências da Saúde

Integração de dados de expressão gênica e proteômica em redes de interação proteína-proteína de *Trypanosoma cruzi*

por

Frederico Gonçalves Guimarães

Belo Horizonte
2017
DISSERTAÇÃO-MCS-CPqRR F.G.GUIMARÃES 2017

FREDERICO GONÇALVES GUIMARÃES

Integração de dados de expressão gênica e proteômica em redes de interação proteína-proteína de *Trypanosoma cruzi*

Dissertação apresentada ao Programa de Pós-Graduação em Ciências da Saúde do Centro de Pesquisas René Rachou como requisito parcial para a obtenção do título de Mestre em Ciências – área de concentração, Biologia Celular e Molecular, Genética e Bioinformática.

Orientação: Dr. Jeronimo C. Ruiz
Coorientação: Dr. Douglas Eduardo Valente Pires

Belo Horizonte
2017

Catalogação-na-fonte
Rede de Bibliotecas da FIOCRUZ
Biblioteca do CPqRR
Segemar Oliveira Magalhães CRB/6 1975

G963i 2017

Guimarães, Frederico Gonçalves.

Integração de dados de expressão gênica e proteômica em redes de interação proteína-proteína de *Trypanosoma cruzi* / Frederico Gonçalves Guimarães. – Belo Horizonte, 2017.

XX, 133 f.: il.: 210 x 297 mm.

Bibliografia: 96 - 100

Dissertação (mestrado) – Dissertação para obtenção do título de Mestre em Ciências pelo Programa de Pós-Graduação em Ciências da Saúde do Centro de Pesquisas René Rachou. Área de concentração: Biologia Celular e Molecular, Genética e Bioinformática.

1. Doença de Chagas/genética 2. *Trypanosoma cruzi*/genética 3. Biologia Computacional/instrumentação I. Título. II. Ruiz, Jeronimo Conceição (Orientação). III. Pires, Douglas Eduardo Valente (Coorientação)

CDD – 22. ed. – 616.936 4

FREDERICO GONÇALVES GUIMARÃES

Integração de dados de expressão gênica e proteômica em redes de interação proteína-proteína de *Trypanosoma cruzi*

Dissertação apresentada ao Programa de Pós-Graduação em Ciências da Saúde do Centro de Pesquisas René Rachou como requisito parcial para a obtenção do título de Mestre em Ciências – área de concentração, Biologia Celular e Molecular, Genética e Bioinformática

Banca Examinadora:
Prof. Dr. Jeronimo Conceição Ruiz (CPqRR/FIOCRUZ) Presidente
Prof. Dr. Douglas Eduardo Valente Pires (CPqRR/FIOCRUZ)
Profa. Dra. Cristiana Ferreira Alves de Brito (CPqRR/FIOCRUZ) Titular
Profa. Dra. Hélida Monteiro de Andrade (UFMG) Titular
Prof. Dr. Rubens Lima do Monte Neto (CPqRR/FIOCRUZ) Suplente

Dissertação defendida e aprovada em Belo Horizonte, 21/02/2017

Este trabalho é dedicado aos pesquisadores e técnicos, muitas vezes anônimos, envolvidos na construção e manutenção de bancos de dados biológicos públicos. A Ciência ganha muito quando a informação está disponível para o mundo.

AGRADECIMENTOS

“Se vi mais longe, foi por estar sobre o ombro de gigantes”, famosa citação de Isaac Newton, se aplica bem minha situação neste trabalho. Longa foi a caminhada e enorme o meu aprendizado. Mas isso só foi possível por poder eu contar com o apoio de seres muito especiais. Eu não teria chegado aonde estou sem esse suporte. Tentarei listá-los abaixo, correndo o risco de deixar alguém de fora, pelo qual eu já peço desculpas. Esses são meus agradecimentos, distribuídos de acordo com seu coeficiente de agrupamento.

A Deus e aos amigos espirituais, pela ajuda imaterial, nem sempre percebida, mas sempre recebida.

À minha mãe Joana D’arc e meus irmãos Fabiano, Felipe e Camila, pelo amor, pela diversão, pela minha orientação moral desde a infância, pelos sacrifícios que garantiram meus primeiros estudos e pelo amparo ao longo do caminho. E, também por tudo isso, um agradecimento especial ao meu pai, Pio, agora ausente em nosso plano, mas sempre presente em meu coração. Saudades...

À minha querida Cris e ao meu querido João. A ela, por ter plantado a ideia da bioinformática na minha cabeça. A ele, por ter revirado a terra e podado os ramos. Espero que estejam satisfeitos com esses primeiros frutos.

À Donana, pelo carinho, cuidado e suporte emocional, tanto pessoalmente quanto à distância. Pode ficar tranquila agora, que essa etapa já passou.

À minha esposa e companheira Cláudia, pelo amor, pelo suporte em minhas decisões (mesmo as mais estranhas, como, por exemplo, fazer um mestrado) e por ter mudado vários detalhes em minha vida que fazem, hoje, uma grande diferença e me tornaram uma pessoa muito melhor.

À princesa Alice, ao anjinho Miguel e ao tigrão Chico, companheiros de quatro patas, presentes e fundamentais em minha vida, sempre prontos a zombar do meu mestrado, dormindo tranquilamente no meu colo ou me chamando pra brincar enquanto eu me debatia com o meu trabalho.

À Eleonora, minha gerente predileta, por todo o apoio burocrático com a flexibilização dos horários de trabalho e pelo estímulo para que eu seguisse com essa ideia. Não pense que acabou. O doutorado vem aí.

Aos amigos e às amigas da SMED, por segurarem as pontas quando a coisa apertava e por ouvirem minhas queixas, o que é muita coisa, pois eu sei o quanto posso ser chato.

Ao Centro de Pesquisas René Rachou, por deixar de ser um prédio no meio do Barro Preto e se tornar um refúgio de estímulo intelectual em um mundo que cada vez pensa menos.

Ao Jeronimo, orientador e amigo, que mostrou que essas duas categorias podem compartilhar significados e tornam-se mais ricas quando o fazem. Obrigado pela acolhida, por acreditar em mim, por apontar o caminho e me permitir trilhá-lo.

Ao Douglas, pelas ideias para a construção deste trabalho.

À Dani, pelos insights científicos, por ouvir minhas ideias mirabolantes (e até acreditar em algumas) e minhas queixas, sempre com algum retorno.

À Laila, pelas discussões científicas e meta-científicas, por me estimular a pensar fora da caixinha e por toda a atenção dispensada ao meu trabalho. E por ter uma singularidade espaço-temporal em sua sala, que te prende gravitacionalmente e faz o tempo se comportar de forma estranha. Não entrem na sala da Laila se tiverem algum compromisso próximo. Fica o aviso.

Aos companheiros e companheiras do grupo Informática de Biossistemas e Genômica pelas dicas e discussões.

Às amigas Grace e Elvira, por me receberem com tanto carinho no laboratório, em uma época em que o nosso espaço físico era muito menor que a Ciência que discutíamos. Obrigado por todo o aprendizado e pelos compartilhamentos.

À Leilane, pela contagiante paixão pela ciência, pela parceria nas tarefas do laboratório e pela participação fundamental na elaboração e execução desse trabalho. E por manter a garrafa de água sempre cheia.

Ao Jader, pela prontidão em sempre ajudar com os problemas técnicos (e resolvê-los!) nos mais diversos momentos.

Aos novos amigos e amiga do laboratório (em ordem alfabética pra não dar briga): Amanda e Carlos, os irmãos Henrique e João, Ludmila, Luiz e Paul. Obrigado pelas sugestões, pelos patuás, pelos memes, pelas gambiarras, pelos biscoitos e pelas risadas, essenciais à manutenção da sanidade mental.

Às Dras. Cristiana e Hélida e ao Dr. Rubens, membros da banca examinadora, pela prontidão em aceitar esta tarefa e dispor de seu precioso tempo para avaliar e colaborar com este trabalho.

RESUMO

Paradoxalmente vivemos um momento da pesquisa científica em que possuímos um volume abundante de dados, mas com dificuldades cada vez maiores de se obter informações a partir deles. Diversidade de formatos, dificuldade na construção de uma forma de acesso simples, mas não superficial, ausência de um identificador único para as unidades biológicas de estudo (proteínas, RNAs, genes, etc.) e falta de integração entre os bancos de dados são alguns dos desafios enfrentados cotidianamente na tarefa de mineração de informações a partir dessas diversas fontes. Com o objetivo de contribuir na tarefa de extração de informações a partir de fontes públicas, mediante a integração de dados de enriquecimento funcional, construímos uma metodologia de trabalho que permite a obtenção, filtragem e tratamento de dados oriundos do banco *STRING* v.10 e de análises massivas de RNA e proteínas, integrando-os em redes de interação proteína-proteína através do software *Cytoscape*. Como organismo modelo, trabalhamos com dois clones de *Trypanosoma cruzi*, apresentando diferenças relacionadas aos perfis de infectividade (alta e baixa infectividade). Utilizamos dados de genes diferencialmente expressos identificados em experimentos de *RNA-Seq* e *shotgun proteomics*. Durante o estudo foram construídos 11 scripts e 3 programas, parte integrante de uma metodologia modular aplicável a outros organismos e modelos experimentais, tanto em sua totalidade quando parcialmente. Como resultado, além da metodologia, obtivemos também o resultado de sua implementação, que consistiu de uma série de redes de interação proteína-proteína do organismo estudado, onde foram destacadas características de interesse biológico, tais como informações de *EC number,* agrupamentos funcionais, tipo de interação entre as proteínas e importância das proteínas segundo métricas de teoria de grafos. Concluímos, então, que a utilização de redes de interação proteína-proteína pode ser uma ótima estratégia tanto para a realização de novos estudos quanto para a revisão de estudos anteriores, uma vez que podemos extrair novas informações a partir de dados já existentes publicamente. Além disso as redes nos fornecem uma visão sistêmica do organismo, o que pode desvelar novos olhares sobre a sua biologia dos organismos de estudo.

Palavras-chave: bioinformática; redes biológicas; redes de interação proteína-proteína; proteínas; *Trypanosoma cruzi*; integração de dados

ABSTRACT

The present epistemological moment in scientific research is characterized by a paradox: there is a considerable amount of data, but the odds in the process of obtaining information from them is overwhelming and growing. The differences between data formats and the difficulties in their handling, the absence of a single identifier for the biological units of investigation (proteins, RNAs, genes etc.) and the lack of integration between the existing databases are some of the challenges researchers face constantly in the task of information mining from this multiple sources. We have built, with the main purpose of contributing to better extracting information from public databases and through the integration of functional enrichment data, a procedural methodology that leads to the obtaining, filtering and handling of data originally contained in the *STRING* v.10 public database and massive analysis of RNA and proteins, integrating them in protein-protein interaction networks with *Cytoscape*. As model organism, we used two clones of *Trypanosoma cruzi*, with different infectivity profiles (high and low infectivity). We used differentially expressed gene data from RNA-Seq and shotgun proteomics experiments. Along our study we built 11 scripts and 3 programs, integrating a modular methodology applied to other organisms and experimental models, as a whole or partially. This work also comes out with the results of the implementation of such a methodological innovation, which consists of a series of protein-protein interaction networks that emphasize characteristics of biological interest, as *EC number* information, functional grouping, protein interaction type and the relevance of protein according to graph theory metrics. We come to a conclusion that the use of protein-protein interaction networks can be an excellent strategy even to produce new researches as to review existing ones, once it's possible to mine new informations from data previously published. Such a procedure becomes especially relevant under the consideration that the strategy of using networks makes it possible to cast a new perspective on current scientific research, usually centered in the study of individual components and not in the systemic aspects of an organism's interactions.

Keywords: bioinformatics; interaction network; protein; *Trypanosoma cruzi*; data integration

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADT – Difosfato de adenosina

ATP – Trifosfato de adenosina

BLAST – Basic Local Alignment Search Tool

CSV – Comma-Separated Values

DC – Doença de Chagas

DNA – Ácido desoxirribonucléico

DNDi – Drugs for Neglected Diseases initiative

DTU – Discrete typing unit

EC number – Enzyme commission number

FTP – File Transfer Protocol

IUBMB – Nomenclature Committee da *International Union of Biochemistry and Molecular Biology*

KEGG – Kyoto Encyclopedia of Genes and Genomes

LogFC – *Log Fold Change*

MAPK – Mitogen Activated Protein Kinases

NCBI – National Center for Biotechnology Information

NR – Não redundante

NTD – Neglected tropical diseases

OMS – Organização Mundial de Saúde

Perl –Practical Extraction and Reporting Language

PNG – Portable Network Graphics

PSI-MI – Proteomics Standards Initiative - Molecular Interaction

ptmod – Post-translational modifications

RefSeq – *Reference Sequence Database*

RGB – Escala de cores vermelha (*Red*), verde (*Green*) e azul (B*lue*)

RIPP – Redes de interação proteína-proteína

RNA – Ácido ribonucléico

RNA Seq – Sequenciamento de *RNA*

rRNA – *RNA* ribossomal

SRPK - serine/arginine-rich protein specific kinase

STRING – Search Tool for the Retrieval of Interacting Genes

SVG – Scalable Vector Graphics

tRNA – *RNA* transportador

TSV – Tab-Separated Values

XML – eXtensible Markup Language

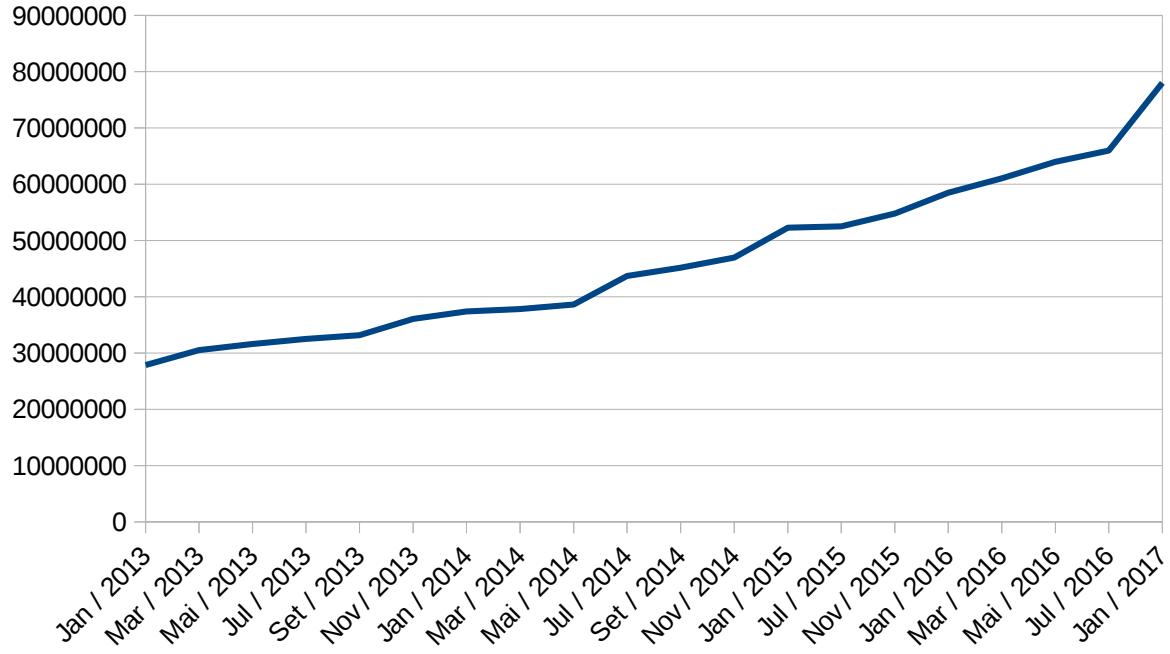
SUMÁRIO

1 INTRODUÇÃO

1.1 O desafio da integração de dados biológicos

A avalanche de dados genômicos que inunda os bancos de dados de domínio público é um fenômeno corriqueiro nos dias de hoje. Somente no *site* do *National Center for Biotechnology Information* (*NCBI*), estão disponíveis 78.028.152 proteínas e 17.862.608 transcritos de RNA, de 66.224 organismos, segundo informações do *Release* 80 do banco de dados de *RefSeq* do *NCBI* ("NCBI Reference Sequence (RefSeq) Database - Release 80", 2017). A evolução da quantidade de proteínas com informações disponíveis no *NCBI* pode ser vista no GRÁFICO 1.

GRÁFICO 1: Evolução da quantidade de proteínas com informações disponíveis no *NCBI*, (janeiro de 2013 a janeiro de 2017).



Entretanto, o processo de extração de informações biológicas úteis desses dados enfrenta uma série de dificuldades. E esse não é um problema novo. Questões envolvendo as diversas fontes de dados biológicos e sua integração já eram discutidas em 1995 (DAVIDSON; OVERTON; BUNEMAN, 1995). Os autores chegam a citar um artigo de 1985, onde Morowitz *et al*. afirmam que: "*new generalizations and higher order biological laws are being approached but may be obscured by the simple mass of data*". Ainda nesse artigo, os autores destacam três pontos que

dificultam a organização dos dados: a variedade dos formatos das fontes de dados, muitas vezes incompatíveis entre si, a necessidade de se desenvolver interfaces simples para filtragens, combinações e tratamentos de dados em condições cada vez mais elaboradas e, por fim, o fato do controle de atualizações ser restrito às equipes locais, de manutenção desses bancos.

Mesmo com diversos avanços tecnológicos, verificamos que as questões básicas de 1995 continuaram as mesmas com o passar dos anos: diversidade de formatos, ausência de interfaces simples e unificadas e a manutenção local dos bancos de dados (GOBLE; STEVENS, 2008; PHILIPPI, 2008; STEIN, 2002, 2003). Isso serviu para reforçar a ideia da necessidade de se trabalhar um conceito de “integração de dados”, o que acabou levando a uma série de publicações sobre o assunto. Segundo (GOMEZ-CABRERO et al., 2014), em 2013, o número de artigos mencionando “data integration” chegou a 2.365. E mesmo assim ainda não existia um consenso sobre a melhor definição para o termo nem a melhor forma de implementar tal tarefa. Ainda segundo esse autor, reforçando o que foi dito anteriormente, existe uma aparente dicotomia entre os dois objetivos das ciências da vida, que são identificar os componentes constituintes dos seres vivos e entender como eles se relacionam. Podemos exemplificar isso apontando que temos um conhecimento bem completo de quais são os elementos constituintes de uma célula, mas ainda existem várias lacunas em entender os detalhes nas interações de todos esses elementos. Tal conhecimento demanda a integração de dados de várias áreas tanto dentro quanto fora da Biologia. E uma forma de se tratar essa integração de dados é através de redes biológicas.

Isso porque a abordagem reducionista, que dominou grande parte da história da pesquisa biológica, foi importante para o entendimento dos componentes individuais das células, mas é insuficiente para explicar a complexidade das interações entre as células e do organismo como um todo. Considerando-se o que foi dito até agora, nos deparamos com uma situação em que temos um grande volume de dados, mas, muitas vezes, isolados de outras análises ou mesmo com uma deficiência em sua caracterização. Uma forma de resolver esse problema seria associar dados ômicos de diversas origens em redes de interação biológicas, incluindo aí redes de

interação proteína-proteína (RIPPs), metabólicas, de sinalização e de regulação transcricional. Estamos focando, neste trabalho, nas RIPPs.

Para a construção dessas redes, partimos de um conjunto de dados com informações de interação entre as proteínas e acrescentamos uma série de camadas de dados, visando a mineração de novas informações que porventura surjam da associação entre eles. E aqui nos deparamos novamente com a questão da diversidade de dados, dessa vez relacionada aos bancos de dados.

Para alguns autores, em particular, STEIN (2002), a diversidade de fontes de dados é importante, pois grupos específicos podem aprofundar o detalhamento sobre seus temas de estudo. Entretanto, GOBLE e STEVENS (2008) argumentam na direção contrária e questionam se realmente é necessária toda essa diversidade, usando, como exemplo, a existência de 231 bancos de dados de caminhos metabólicos (à época da publicação do artigo, em 2008). No caso específico de bancos de dados de interação de proteínas, um levantamento feito em janeiro de 2017 no site *Pathguide* (*http://www.pathguide.org/*), que contém informações de bancos de vias metabólicas e de interações moleculares, identificamos a existência de 256 bancos de dados que tratam desse assunto. Mas apesar dessa aparente diversidade, existe uma concentração nos organismos com maior abundância de dados, tais como *Saccharomyces cerevisiae*, *Drosophila melanogaster*, *Escherichia coli*, *Arabidopsis thaliana* e *Homo sapiens sapiens*.

Uma possível abordagem para contornar esse problema da diversidade de bancos seria o uso de “meta-bancos”, ou seja, bancos de dados que agreguem informações de outras fontes, mas, mesmo assim, o problema da ausência de diversidade de organismos persiste. Em particular no caso deste trabalho, que trata como organismo modelo o *Trypanosoma cruzi*.

Por fim, um outro ponto a ser considerado é a longevidade das fontes de informação. Ao longo de nosso trabalho encontramos referências a vários bancos de dados e ferramentas de agregação que não estão mais disponíveis ou que deixaram de ser atualizadas. Por isso, ao se escolher um banco de dados é importante verificar há quanto tempo ele está ativo e se é constantemente atualizado.

Por tudo isso, após extensa pesquisa entre as opções de bancos de dados, escolhemos o *Search Tool for the Retrieval of Interacting Genes* (*STRING*) como referência para este trabalho. Acessível a partir da URL *http://version10.string-db.org/*, o objetivo do *STRING* é organizar e disponibilizar dados de interação proteína-proteína, incluindo associações diretas (físicas) e indiretas (funcionais) (SZKLARCZYK et al., 2015). Esse banco de dados se diferencia dos outros por agregar em um único local tanto dados experimentais quanto preditos, possuir um escore de confiança associado a cada interação, permitir a transferência dos seus dados brutos e apresentar recursos flexíveis de criação online de redes, com a possibilidade de recuperação dos dados associados às redes criadas. Além disso, ele atende aos critérios de ser longevo (foi criado em 2001), passar por atualizações constantes e possuir dados sobre o *T. cruzi*.

O *STRING* obtém seus dados a partir de várias fontes e, conforme descrito acima, atribui escores de confiança para cada uma delas. As fontes de dados são: vizinhança genômica (*neighborhood*), fusão gênica (*fusion*), co-ocorrência entre espécies (*cooccurence*), co-expressão na mesma ou em outras espécies (*coexpression*), dados experimentais (*experimental*), bancos de dados (*database*) e mineração de dados na literatura (*textmining*). Ele também possui um escore combinado (*combined_score*), que é o resultado da ponderação entre os valores atribuídos a cada fonte.

O *STRING* disponibiliza, em sua página de downloads (*http://version10.string-db.org/cgi/download.pl*), dados na forma de arquivos em formato texto puro, tanto em formato tabular quanto arquivos de *dump*, que podem ser importados diretamente em bancos de dados. Como esses arquivos tendem a ser muito extensos, pela quantidade de organismos abarcada pelo projeto e o volume de dados associados a cada um deles, o site oferece a possibilidade de se transferir somente arquivos relacionados a determinada espécie, o que reduz consideravelmente o seu tamanho. A título de comparação, o arquivo que apresenta detalhes das interações de todos os organismos possui um tamanho compactado de 17,8 Gb. O mesmo arquivo, com dados somente de *T. cruzi*, possui apenas 17 Mb.

Além da possibilidade de transferir os arquivos de dados, o *STRING* também oferece o recurso de se construir a rede de interações a partir do próprio site, utilizando dados de entrada que podem ser os identificadores das proteínas ou as suas sequências de aminoácidos. Ao entrar com essa informação, o *STRING* procura, em seu banco de dados, pela existência de conexões diretas entre essas proteínas, exibido-as na forma de ligações entre os nós conectados. É possível definir parâmetros específicos para a construção dessa rede, tais como o escore de confiança mínimo desejado e o número de primeiros vizinhos das proteínas de entrada, que o *STRING* busca em seu banco de dados.

Quando se usa esse recurso, além do desenho da rede, ele também permite que sejam transferidos diversos arquivos de dados com os resultados da rede gerada. Os arquivos que podem ser obtidos são:

- Uma imagem *bitmap* da rede gerada, no formato*PNG*;
- uma imagem *bitmap*, em alta resolução (400 dpi), da rede gerada, no formato *PNG*;
- uma imagem vetorial da rede gerada, em formato SVG;
- dados tabulares das interações da rede, em formato *TSV*;
- dados das interações da rede, definidos segundo o padrão *PSI-MI*, em formato XML estruturado;
- dados de cores e coordenadas de cada um dos nós da rede, em formato texto puro;
- sequências de aminoácidos de todas as proteínas da rede, em formato *multi-fasta*;
- nomes, domínios e anotações funcionais de cada uma das proteínas da rede, organizados de forma tabular, em formato *TSV*.

Entretanto, o recurso de criação de redes dentro do próprio *STRING* apresenta algumas limitações, pois não permite uma análise mais detalhadas das métricas da rede, além de restringir o tamanho máximo da rede a ser analisada. Uma vez que a rede gerada nesse trabalho apresenta um número restrito de proteínas, essa segunda limitação não foi um problema, mas a primeira, sim. Por isso, utilizamos o

site somente para gerar a nossa rede básica e exportar os dados referentes a essa montagem, efetuando as análises mais detalhadas com o programa *Cytoscape*.

1.2 Redes biológicas

Uma informação relevante relativa às redes biológicas é que elas possuem as mesmas características funcionais e de arquitetura presentes em outras redes complexas, como a Internet ou redes sociais (BARABÁSI; OLTVAI, 2004). Essas redes complexas, conhecidas matematicamente como *grafos*, podem ser definidas como um conjunto de elementos, denominados *nós* (do inglês *node*) conectados entre si por uma série de ligações, denominadas *arestas* (*edge* em inglês). Em termos biológicos, por exemplo, os nós poderiam representar proteínas e, as arestas, as relações metabólicas entre elas. Na FIGURA 1 temos a representação de uma rede, com os seus nós e arestas identificados, respectivamente, por letras minúsculas e maiúsculas.

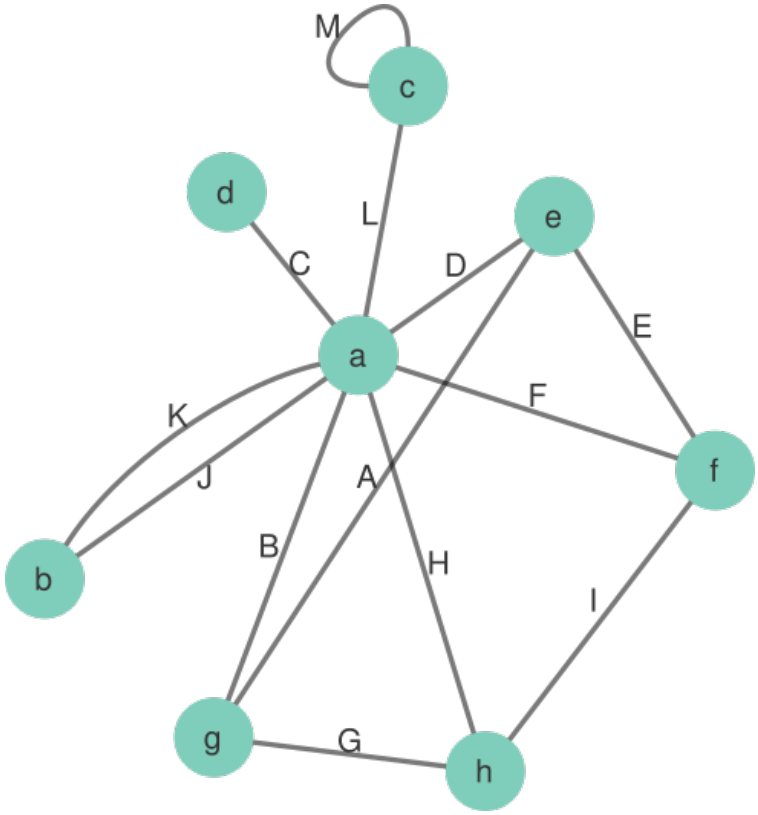


FIGURA 1: Representação de uma rede, identificando seus nós (em letras minúsculas) e arestas (em letras maiúsculas).

O número de arestas associadas a um nó determina o seu *grau* (*degree* em inglês). Ou seja, quanto maior o número de arestas de um nó (ou, falando de outro modo, quanto mais conexões esse nó possui), maior o seu grau. Usando a FIGURA 1 como exemplo, o grau do nó *a* é 8 e do nó *b* 2. Os grafos também possuem uma métrica, denominada *distribuição de graus*, que é a probabilidade de um determinado grau aparecer na rede, ou seja, representa a fração de nós que possuam determinado grau. Essas métricas são importantes na identificação da arquitetura das redes, conforme será detalhado adiante.

Para percorrermos uma rede, de um nó específico a outro, muitas vezes é possível seguirmos por vários caminhos distintos, percorrendo diferentes arestas. Usando a FIGURA 1 como exemplo, para irmos do nó *d* ao nó *h*, podemos percorrer as arestas C, *B* e *G* ou *C* e *H* ou *C*, *F* e *I* ou, por fim, *C*, *D*, *E* e *I*. Deriva daí uma métrica importante, denominada *caminho mínimo*, que corresponde ao menor número de arestas a serem percorridos para se chegar de um nó a outro. No exemplo apresentado anteriormente, o *caminho mínimo* entre *d* e *h* seria *C* e *H*. Essa métrica pode ser utilizada para definirmos mais outra, denominada *intermediação* (*betweenness* em inglês) que indica o número de menores caminhos que passam por um determinado nó. Ao contrário da medida do *grau*, que pode ser obtido a partir do próprio nó, a medida de *intermediação* só pode ser obtida analisando-se a rede como um todo, uma vez que ela é comparativa entre todos os nós da rede. Por isso ela é considerada uma medida global, em contraposição ao grau, que é uma medida local. O valor da intermediação varia entre 0 e 1.

Outra métrica utilizada neste trabalho é o coeficiente de agrupamento (*clustering coefficient*), que mede a tendência que os nós em uma rede têm de se agruparem. Em redes biológicas, é comum existirem agrupamentos funcionais que aparecem representados como também como grupos na rede (BARABÁSI; OLTVAI, 2004). O valor dessa métrica também varia entre 0 e 1

As redes possuem uma série de características de nomenclatura e arquitetura. Esse assunto foi muito bem apresentado por BARABÁSI e OLTVAI (2004) e está sintetizado nas definições a seguir.

Segundo a natureza da sua interação, uma rede pode ser *orientada*, quando a interação entre seus nós possui uma direção bem definida. Um exemplo seria uma rede formada a partir de dados de uma via metabólica, já que a interação de seus componentes segue sempre uma determinada direção. Já uma rede *não orientada* não possui direção definida na interação entre seus nós. Uma rede de interação de proteínas é comumente não orientada, uma vez que ela normalmente representa somente as interações físicas entre as proteínas, sem nenhum tipo de direcionamento pré-definido.

Em relação à arquitetura da rede, existem dois modelos básicos: *redes aleatórias* e *redes livres de escala*. Nas redes aleatórias, as arestas estão distribuídas aleatoriamente entre os nós da rede. Assim, de uma forma geral, existe uma distribuição razoavelmente homogênea na quantidade de ligações entre os nós e os seus graus seguem uma distribuição de Poisson (FIGURA 2a). Já nas redes livres de escala, encontramos alguns poucos nós que possuem muitas arestas e a maioria possui poucas. Ou seja, alguns poucos nós estão ligados a muitos outros nós, mas a maioria está ligada a poucos outros nós. Dessa forma, os graus dos nós dessa rede seguem uma distribuição segundo uma lei de potência (FIGURA 2b). Nesse modelo de rede, os nós que possuem muitas ligações são chamados *hubs*.

Uma constatação que surge ao analisarmos os processos biológicos é que, em sua grande maioria, as redes biológicas se encaixam na arquitetura “livre de escala” (BARABÁSI; OLTVAI, 2004).

Em particular nas redes de interação de proteínas, é comum encontrarmos proteínas *hub* que estabelecem conexões com várias outras. Essas proteínas são bons alvos para intervenções moleculares, especialmente quando associadas a vias metabólicas importantes, uma vez que a sua inativação pode desestruturar grande parte da rede ao seu redor, interferindo no metabolismo do organismo.

A

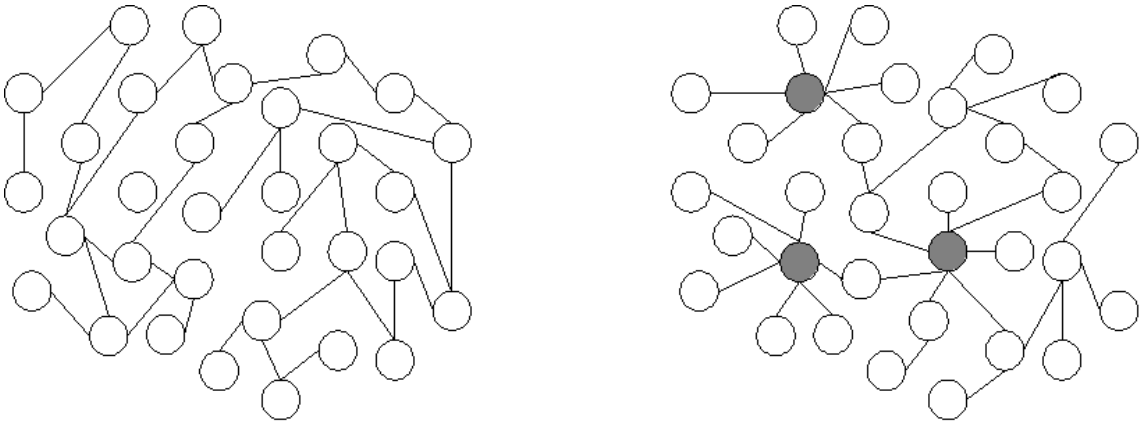


FIGURA 2: Modelos representando as duas principais arquiteturas de redes complexas. O modelo a representa uma rede aleatória, onde se percebe a distribuição não ordenada de arestas entre os nós, e o modelo b, uma rede livre de escala, marcada pela existência poucos nós associados a várias arestas. Esses nós, marcados em cinza, são denomindos *hubs*. Fonte: Wikipedia - Scale-free network (https://en.wikipedia.org/wiki/Scale-free_network).

1.3 O *Trypanosoma cruzi* no contexto das doenças negligenciadas

São classificadas como “negligenciadas” diversas doenças tropicais que afetam a vida de milhões de pessoas no mundo, mas que, mesmo assim, não são valorizadas pela indústria farmacêutica nem recebem o devido aporte de pesquisas e estudos, quando comparadas com outras enfermidades (SOUZA, 2010). Neste grupo estão as doenças negligenciadas tropicais (*NTD*, do inglês, *Neglected tropical diseases*) que, segundo a Organização Mundial de Saúde, são um conjunto de doenças com alta prevalência em regiões tropicais e subtropicais. Em conjunto, estima-se que elas afetem cerca de um bilhão de pessoas anualmente.

Em 2003, o *Drugs for Neglected Diseases initiative* (*DNDi*) reconheceu a necessidade de maiores iniciativas e investimentos em algumas das *NTD*s, das quais podemos citar a Doença de Chagas, as Leishmanioses e a Tripanossomíase Africana (FEASEY et al., 2010; WAINWRIGHT, 2017). Tendo-se em conta o impacto dessas doenças na saúde pública e a tendência geral da indústria farmacêutica em priorizar pesquisas de medicamentos pela ordem de lucro que geram, recai sobre o

poder público a responsabilidade de arcar com as pesquisas relacionadas a essas enfermidades e seus organismos de interesse. A título de exemplificação, apesar de mais de 500 milhões de pessoas estarem sob a ameaça dessas três doenças somente 5% do financiamento mundial de inovação para doenças negligenciadas foram investidos nesse grupo (“Por que negligenciadas?”, 2010).

Nesse contexto, a Doença de Chagas (DC), transmitida pelos protozoários parasitos da espécie *T. cruzi*, e também conhecida como *Tripanossomíase Americana*, afeta entre 6-7 milhões de pessoas mundialmente e, devido as taxas de morbidade e mortalidade, estima-se uma perda econômica de aproximadamente 600 milhões de dólares anuais (LEE et al., 2013). Como pode ser observado na FIGURA 3, na América Latina a DC é endêmica em cerca de 21 países. Em outros, como nos Estados Unidos, Austrália, Japão e alguns locais na Europa as transmissões autóctones são extremamente raras, no entanto, devido à imigração de pessoas infectadas de áreas endêmicas diversos casos têm sido reportados (CONNERS et al., 2016; GASCON; BERN; PINAZO, 2010; WHO, 2016).

Por apresentarem um ciclo de vida tanto em vetores invertebrados (formas epimastigotas e tripomastigotas metaciclicas) quanto em hospedeiros mamíferos (formas tripomastigotas sanguineas e amastigotas), com diferentes estágios de desenvolvimento em cada um deles, esses parasitos necessitam de uma maquinaria genética eficiente, capaz de promover uma rápida resposta morfológica e bioquímica decorrentes das mudanças de ambiente (FERNÁNDEZ-MOYA; ESTÉVEZ, 2010; HAILE; PAPADOPOULOU, 2007; MARTÍNEZ-CALVILLO et al., 2010).

E para serem capazes de se adaptar a todas essas mudanças estes parasitas apresentam características distintas como alta sintenia e mecanismos de expressão gênica como a transcrição policistrônica, *trans-splicing* e mecanismos de edição de RNA. Devido a esses mecanismos a regulação da expressão gênica é pós-transcricional (FERNÁNDEZ-MOYA; ESTÉVEZ, 2010; GAUDENZI et al., 2011).

Segundo Fernández-Moya e Estévez (2010) esse controle pós-transcricional possui um importante papel na alteração da modulação da expressão em resposta aos estímulos externos.

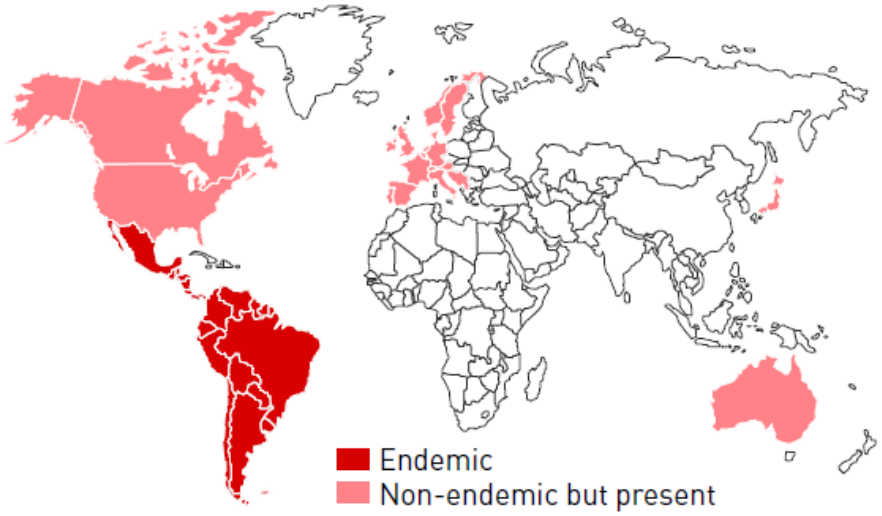


FIGURA 3: Distribuição da endemicidade da doença de Chagas. Essa doença é endêmica em países da América Latina, no entanto, devido a imigração de pessoas dessas áreas para outros países, diversos casos têm sido reportados. Fonte: Drugs for Neglected Diseases initiative (http://www.dndi.org/diseases-projects/chagas/)

Seu genoma possui um tamanho médio de 55Mb com mais de 50% de regiões de sequências repetidas, incluindo um grande número de famílias multigênicas, como por exemplo, as transialidases e mucinas, além de retrotransposons e repetições em regiões subteloméricas (GAUDENZI et al., 2011; MARTÍNEZ-CALVILLO et al., 2010).

Atualmente, no banco de dados de domínio público *TritrypDB*, estão disponíveis os genomas de 9 linhagens de *T. cruzi*. Essas linhagens apresentam uma grande variação intra específica e por isso são classificadas em seis diferentes *DTU*s (do inglês *discrete typing unit*), e segundo Zingales (2011), cada DTU pode ser definida como grupos de isolados que compartilham semelhanças genéticas e cuja identificação possa ser realizada utilizando marcadores moleculares e/ou imunológicos comuns (ZINGALES, 2011; ZINGALES et al., 2009). A TABELA 1 mostra as diferentes *DTU*s e algumas das linhagens incluídas em cada uma delas.

TcI é uma *DTU* caracterizada predominantemente por linhagens com ciclo de transmissão silvestre. Já *TcII* é uma *DTU* predominantemente doméstica nos países

do Cone Sul (Brasil, Argentina, Bolívia, Chile, Paraguai e Uruguai). A *TcII* foi, posteriormente subdividida nas demais *DTU*s (TEIXEIRA et al., 2012).

TABELA 1: Linhagens de *T. cruzi* representativas para cada *DTU*. Adaptado de: Teixeira et al., 2009.

DTU	**Linhagens**
TcI	Sylvio X-10, Dm28c
TcII	Esmeraldo, Y
TcIII	CM17
TcIV	CanIII
TcV	SO3
TcVI	CL Brener

1.4 Justificativa

Um dos grandes desafios da biologia atual está voltado à integração de dados genômicos. Dentro desse contexto, a existência de dados genômicos do organismo modelo de estudo em bancos de dados de domínio público representa um fator que catalisa a integração de dados biológicos. Por outro lado, dados de expressão massiva de RNA (*RNASeq*) e proteômica gerados dentro do grupo Informática de Biossistemas representam uma oportunidade ímpar para a construção de redes biológicas que tem o grande potencial de trazer à luz aspectos ainda pouco conhecidos da biologia do protozoário.

Ao integrarmos dados de genômica, transcriptômica e proteômica em uma rede biológica, acrescentamos camadas de informação que podem fazer emergir uma série de hipóteses biológicas não percebidas anteriormente. Esta estratégia é particularmente importante quando estamos em busca de pontos de conexão na relação patógeno-hospedeiro ou de possíveis alvos para fármacos ou terapias genéticas, uma vez que, nesses casos, normalmente lidamos com relações complexas entre diversas regiões do interatoma e/ou do proteoma intra e interespecíficos. Uma rede bem construída e enriquecida com dados de outras análises pode funcionar como um norteador na busca dessas relações.

Com a execução deste projeto teremos elementos suficientes para a construção de um protocolo de enriquecimento de uma rede de interação de proteínas com dados de expressão gênica, e este representa nosso principal objetivo A plataforma de integração de dados a ser desenvolvida tem o potencial de servir como base para futuras análises em diferentes projetos e ser extrapolada para o enriquecimento do nosso conhecimento biológico sobre *T. cruzi* e outros organismos.

2 OBJETIVOS

2.1 Objetivo geral

Integrar dados massivos de sequenciamento paralelo de RNA e de proteínas de *Trypanosoma cruzi* em redes de interação de proteínas, identificando propriedades topológicas distintas e categorizando aquelas com potenciais propriedades funcionais biológicas, visando a produção de um protocolo de integração de dados que possa ser utilizado também para outros organismos.

2.2 Objetivos específicos

- Construir redes de interação de proteínas, usando como referência a base de dados do *STRING* v.10;
- caracterizar e classificar as redes geradas através de métricas específicas (grau, intermediação e coeficiente de agrupamento) visando a identificação de características que definam uma rede biológica;
- enriquecer a informação funcional associada às redes geradas com dados oriundos do sequenciamento massivo de RNA e de proteínas de dois isolados de *T. cruzi* que apresentam diferenças de infectividade;
- construir e implementar um protocolo de construção de redes de interação proteína-proteína a partir de dados de transcriptoma e/ou proteoma.

3 METODOLOGIA

3.1 Obtenção dos dados

Os dados utilizados durante o desenvolvimento desse trabalho foram obtidos junto ao grupo **Informática de Biossistemas e Genômica** do **Centro de Pesquisas René Rachou**.

Além disso, obtivemos o proteoma predito de *T. cruzi* Sylvio X10, a partir da versão 25 do TriTrypDB. O arquivo obtido foi salvo com o nome de *TcruziSylvioX10.fasta*. Ele foi utilizado, posteriormente, como banco de dados do BLAST para a obtenção de identificadores das proteínas analisadas.

3.1.1 Dados de RNASeq e Shotgun Proteomics

Foram utilizados, neste estudo, dados relativos a dois clones de *T. cruzi*, um apresentando perfil de alta infectividade (*NM1-cl1*) e outro com perfil de baixa infectividade (*CI2-cl2*). Os dados de sequenciamento massivo de RNA e proteínas desses organismos foram gerados através de um projeto de pesquisa envolvendo os grupos liderados pelo Dr. Jeronimo Ruiz (CPqRR), Dra. Barbara Burleigh (Harvard T. H. Chan, School of Public Health, Estados Unidos), Dra. Rebeca Manning Cela (CINVESTAV, México), Dr. Najib M. El-Sayed (University of Maryland, Estados Unidos) e Dr. Rushdy Ahmad (Broad Institute, Estados Unidos).

3.1.2 Seleção de genes integrantes de famílias multigênicas

No desenvolvimento deste trabalho, considerou-se como família multigênica agrupamentos de genes correlatos que possuíssem mais de 10 representantes. O agrupamento foi realizado em função do resultado da anotação e classificação funcional dos genes. A anotação funcional e classificação foi realizada através de buscas por similaridade de sequência utilizando-se a suíte *Basic Local Alignment Search Tool* (*BLAST*) do *NCBI* (CAMACHO et al., 2009) e o banco de dados não redundante (NR) de sequências de proteínas do *NCBI*. Foram selecionadas proteínas e genes com maior valor de cobertura, combinadas à classificação funcional.

3.1.3 Seleção de genes diferencialmente expressos

Com o objetivo de selecionar RNAs e proteínas regulados positiva e negativamente, em relação à infectividade, os dados de expressão de RNA e proteína foram submetidos aos critérios de corte de logFoldChange (logFC) de 0,99 e p-valor ajustado menor que 0,05 encontramos um total de 2.129 genes diferencialmente expressos e 380 proteínas diferencialmente expressas.

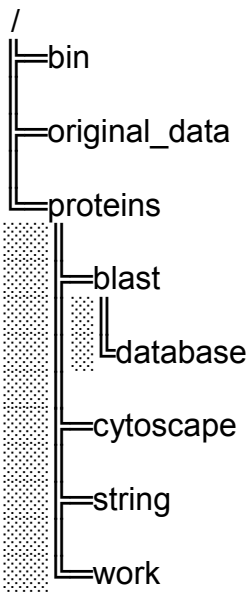
3.1.4 Consenso dos dados

Os procedimentos realizados até esta etapa, foram efetuados no projeto de pesquisa citado no item 3.1.1. Para a execução das análises efetuadas no escopo deste trabalho, utilizamos apenas o consenso desses dados, ou seja, genes diferencialmente expressos que codificavam proteínas também identificadas como diferencialmente expressas. O conjunto de dados inicial deste trabalho foi composto, portanto, de 233 proteínas. Esse consenso foi gerado agrupando-se os dados em quatro categorias, de acordo com a regulação (negativa ou positiva) de RNA e proteína: RNA neg/proteína neg, RNA neg/proteína pos, RNA pos/proteína neg e RNA pos/proteína pos.

Uma vez que foi realizado um agrupamento de dados de RNA e proteína, novamente nos deparamos com representantes de famílias multigênicas. Mais uma vez fizemos a seleção de um representante para cada família, conforme a metodologia descrita na seção 3.1.2, ou seja, maior cobertura.

3.2 Configuração do espaço de trabalho para o desenvolvimento do projeto

Para este trabalho, criamos uma estrutura de diretórios bem definida, visando organizar os dados em espaços distintos, associados ao tipo de tratamento e/ou sua origem. O diretório inicial do trabalho será referenciado, neste texto, como *raiz* (e será utilizado o símbolo “/”). Todos os outros estão contidos nele e serão referenciados a partir do símbolo “/”. Assim, a referência “*/proteins/cytoscape*” significa um diretório “*cytoscape*”, criado dentro de um “*proteins*”, criado no diretório inicial. A hierarquia de diretórios definida para este trabalho é a seguinte. Os diretórios criados foram os seguintes:


/
bin
original_data
proteins
blast
database
cytoscape
string
work

3.3 Construção dos arquivos utilizados como origem dos dados de análise

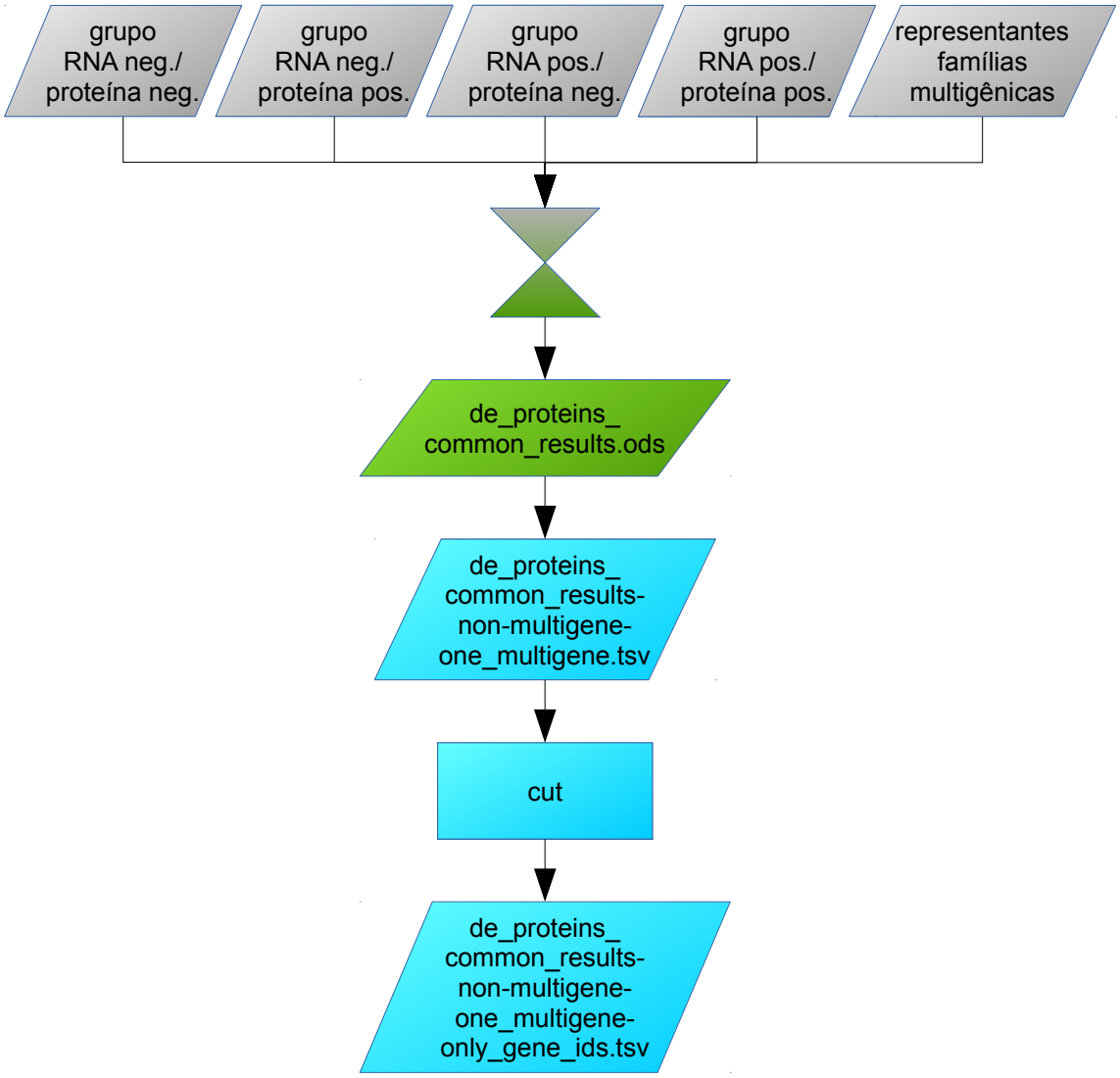


FIGURA 4: Fluxograma dos procedimentos utilizados na construção dos arquivos utilizados na origem das análises.

Reunimos os arquivos produzidos conforme descrição no tópico 3.1 em abas distintas de uma planilha eletrônica, de nome *de_proteins_common_results.ods*, que foi ordenada, formatada e salva dentro do diretório */original_data*.

Adicionamos a essa mesma planilha, mais uma aba, reunindo todas as proteínas não pertencentes a famílias multigênicas e mais aquelas selecionadas como representantes das famílias multigênicas. Em seguida, essa aba foi salva, isoladamente, como um arquivo à parte no diretório */proteins*, em formato *TSV* (*Tab-*

Separated Values), nomeado como *de_proteins_common_results-non-multigene-one_multigene.tsv*. Utilizamos tabulação como separador de campos.

Por fim, geramos mais um arquivo, denominado *de_proteins_common_results-non-multigene-one_multigene-only_gene_ids.tsv,* contendo somente a informação dos identificadores dos genes (*gene id*). Esse arquivo foi gerado, a partir do diretório */proteins*, com o comando abaixo:

```
cut -f1 de_proteins_common_results-non-multigene-
one_multigene.tsv > de_proteins_common_results-non-multigene-
one_multigene-only_gene_ids.tsv
```

3.4 Pesquisa pelos identificadores do *STRING* utilizando busca por similaridade de sequência

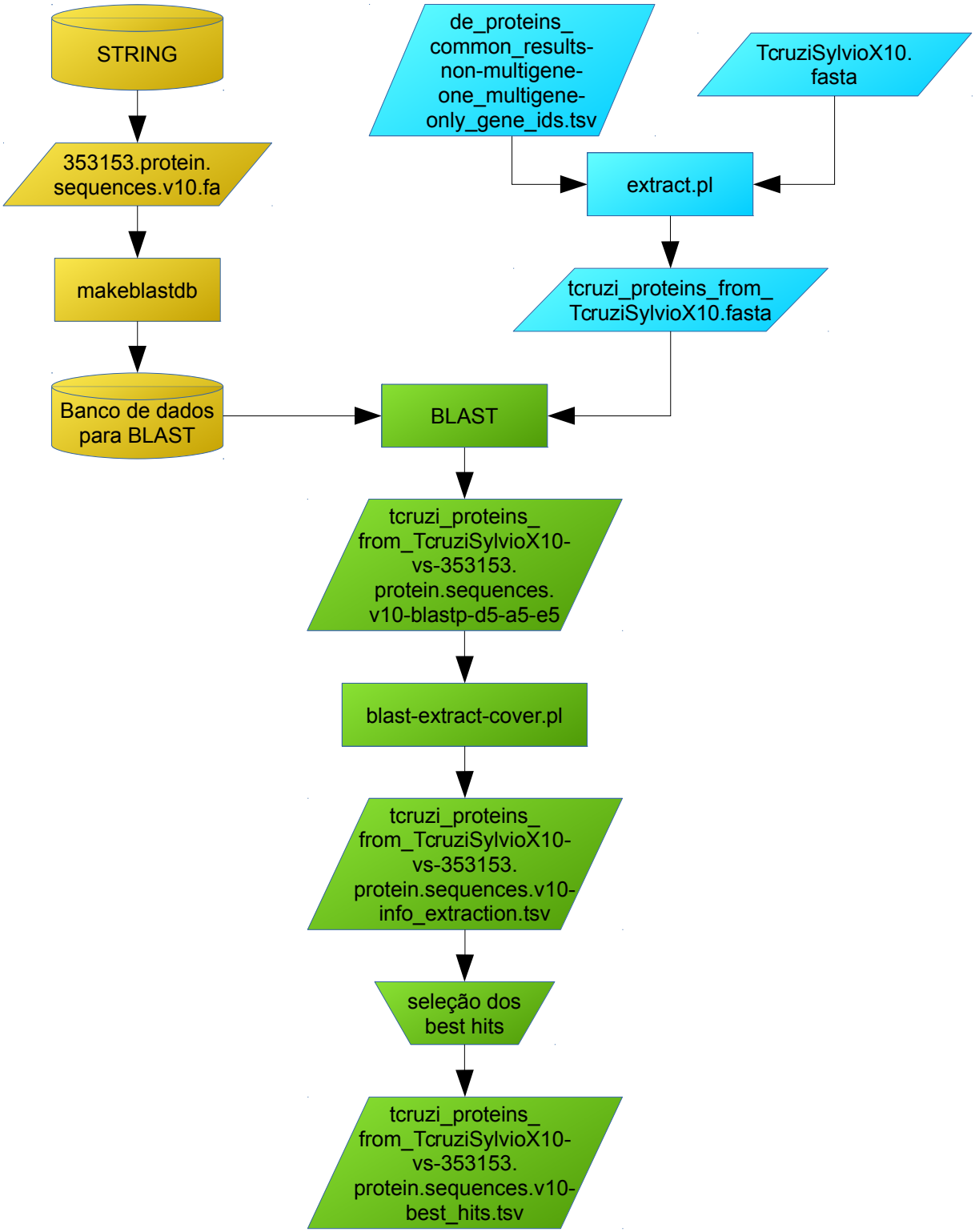


FIGURA 5: Fluxograma dos procedimentos utilizados na busca pelos identificadores do *STRING* associados às proteinas de *T.cruzi*.

Para montarmos a rede no *STRING* era necessário descobrir o identificador (denominado aqui como *STRING id)* de cada uma das proteínas estudadas, em seu banco de dados. O método utilizado para encontrar esses identificadores foi executar uma busca por similaridade de sequência contra as proteínas de *T. cruzi* disponíveis no *STRING*, a qual realizamos localmente, em nossos servidores, utilizando a suíte *BLAST* na versão 2.5.4.

O arquivo com todas as sequências de aminoácidos de cada uma das proteínas contempladas pelo *STRING* está disponível em seu link de *Downloads*. Lá, após selecionar o organismo de interesse – no caso deste trabalho, *Trypanosoma cruzi* –, fizemos a transferência do arquivo *353153.protein.sequences.v10.fa.gz* para o diretório */proteins/blast/database*. O arquivo foi descompactado e, após isso, construímos o banco de dados do *BLAST* dentro desse mesmo diretório, com o comando:

```
makeblastdb -in 353153.protein.sequences.v10.fa -dbtype prot
```

No diretório */proteins/blast*, fizemos a extração das sequências de aminoácidos das proteínas existentes no arquivo *de_proteins_common_results-non-multigene-one_multigene-only_gene_ids.tsv*, gerado anteriormente. Para isso, utilizamos o programa *extract.pl*, disponível para os membros do nosso grupo de pesquisas, com a seguinte linha de comando:

```
../../bin/extract.pl -i
../../original_data/TcruziSylvioX10.fasta -o
tcruzi_proteins_from_TcruziSylvioX10.fasta -l
../de_proteins_common_results-non-multigene-one_multigene-
only_gene_ids.tsv
```

Em seguida, nesse mesmo diretório, executamos o *BLAST* através da seguinte linha de comando:

```
blastp -db database/353153.protein.sequences.v10.fa -query
tcruzi_proteins_from_TcruziSylvioX10.fasta -out
tcruzi_proteins_from_TcruziSylvioX10-vs-
353153.protein.sequences.v10-blastp-d5-a5-e5 -num_descriptions
5 -num_alignments 5 -evalue 0.00001
```

Por fim, construímos um programa em *Perl* para extrair diversas informações do arquivo de resultados do *BLAST*, denominado *blast-extract-info.pl*. A sintaxe de execução do programa, ainda no mesmo diretório */proteins/blast*, é a seguinte:

```
blast-extract-info.pl -i tcruzi_proteins_from_TcruziSylvioX10-
vs-353153.protein.sequences.v10-blastp-d5-a5-e5 -o
tcruzi_proteins_from_TcruziSylvioX10-vs-
353153.protein.sequences.v10-info_extraction.tsv
```

Esse programa gera um arquivo tabular com um resumo dos resultados da busca por similaridade de sequência. Esse arquivo (*tcruzi_proteins_from_TcruziSylvioX10-vs-353153.protein.sequences-info_extraction.tsv*) foi utilizado para selecionar os melhores resultados (*best hits*) do *BLAST*, que foram salvos em um novo arquivo: *tcruzi_proteins_from_TcruziSylvioX10-vs-353153.protein.sequences.v10-best_hits.tsv*.

3.5 Estruturação da rede no *STRING*

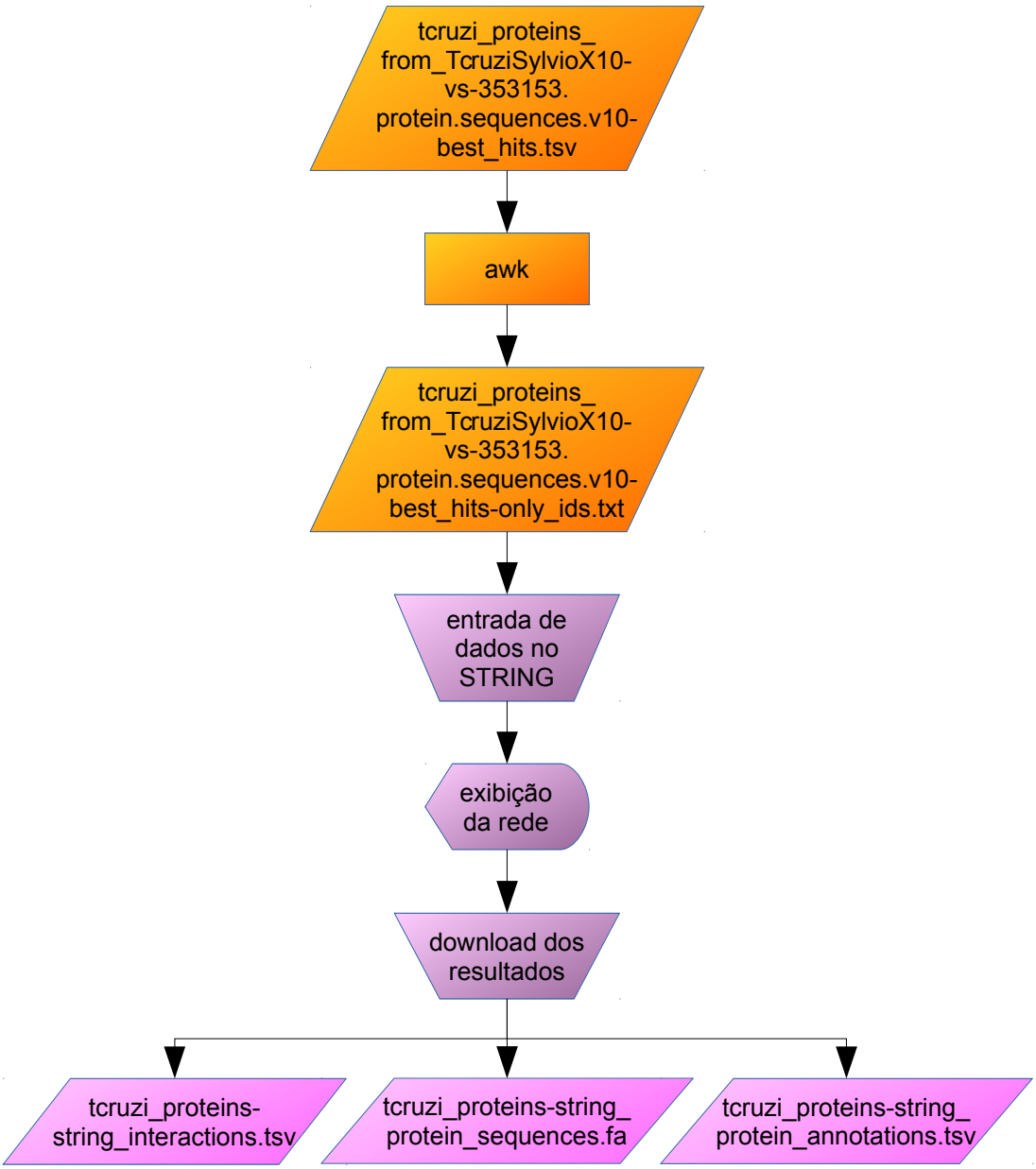


FIGURA 6: Fluxograma dos procedimentos utilizados na construção inicial da rede no STRING.

Na sequência, isolamos os *STRING ids* a partir do arquivo *tcruzi_non-multigenic_from_TcruziSylvioX10-vs.353153.protein.sequences.v10-best_hits.tsv* gerado anteriormente. Para isso, utilizamos o comando abaixo, a partir do diretório */proteins/string*:

```
awk '{if (substr($1,1,5)!="Query") print $4}'
../blast/tcruzi_proteins_from_TcruziSylvioX10-vs-
353153.protein.sequences.v10-best_hits.tsv >
tcruzi_proteins_from_TcruziSylvioX10-vs-
353153.protein.sequences.v10-best_hits-only_ids.txt
```

O arquivo gerado com os resultados (*tcruzi_non-multigenic_from_TcruziSylvioX10-vs.353153.protein.sequences.v10-best_hits-only_ids.txt*) foi utilizado como entrada de pesquisa no site do *STRING*, através do recurso de busca por múltiplas proteínas.

Nos parâmetros do STRING, adicionamos o valor 500 ao campo **1st shell** para que ele acrescentasse os primeiros vizinhos de cada proteína exibida inicialmente. Como fontes de interações, selecionamos apenas: *experiments e databases.* Foi também mantido o escore mínimo de interação de 400.

Uma vez montada a rede, utilizamos o recurso de exportação do site, disponível na aba "*Tables / Exports*". Criamos uma padronização nos nomes dos arquivos, acrescentando o prefixo "tcruzi_proteins-" a cada um dos nomes originais do *STRING*. Os links de onde os arquivos foram baixados estão listados abaixo, junto aos nomes utilizados para cada um deles:

- … *as simple tabular text output* – tcruzi_proteins-string_interactions.tsv;
- … *protein sequences* – tcruzi_proteins-string_protein_sequences.fa;
- … *protein annotations* – tcruzi_proteins-string_protein_annotations.tsv.

3.6 Obtenção e filtragem da rede com todas as interações de proteínas de *T. cruzi* disponíveis no *STRING*

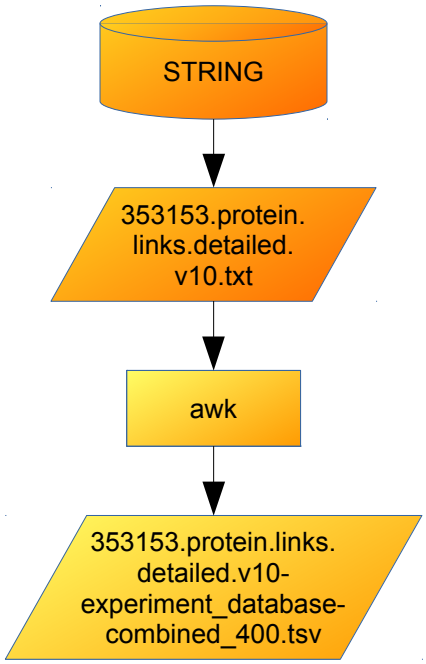


FIGURA 7: Fluxograma dos procedimentos utilizados na obtenção e filtragem dos dados da rede completa de *T. cruzi* disponível no *STRING*.

A partir do *site* do *STRING*, nós também transferimos, para o diretório */proteins/string*, o arquivo com a relação de todas as interações de proteínas disponíveis, em seu banco de dados, para *T. cruzi*. Esse arquivo foi obtido no link de *Downloads*. Selecionamos *Trypanosoma cruzi* como organismo de interesse e fizemos a transferência e posterior descompactação do arquivo *353153.protein.links.detailed.v10.txt.gz*.

Em seguida, para atender ao critério de corte utilizado no procedimento 3.5, separamos as interações oriundas de bancos de dados e experimentos e com um escore combinado de 400. Para isso utilizamos o seguinte comando *awk*, a partir do diretório */proteins/string*.

```
awk 'BEGIN {print "protein1 protein2 neighborhood fusion
cooccurence coexpression experimental database textmining
combined_score"} {if ($7 * $8 > 0) score = (1 - 0.95886979 *
((1 - ($7 / 1000)) / 0.95886979) * (1 - ($8 / 1000)) /
0.95886979); else score = ($7 + $8)/1000; if (score >= 0.400)
print $0}' 353153.protein.links.detailed.v10.txt >
353153.protein.links.detailed.v10-experiment_database-
combined_400.tsv
```

3.7 Preparação dos dados utilizados na análise de enriquecimento funcional da rede

Para trabalhar os dados a serem utilizados no enriquecimento funcional das redes, utilizamos o diretório */proteins/work*. Dependendo do tipo de informação funcional tratada, diferentes procedimentos foram adotados durante o processo de preparação dos dados.

3.7.1 Informação funcional: Categorias de regulação RNA/proteína

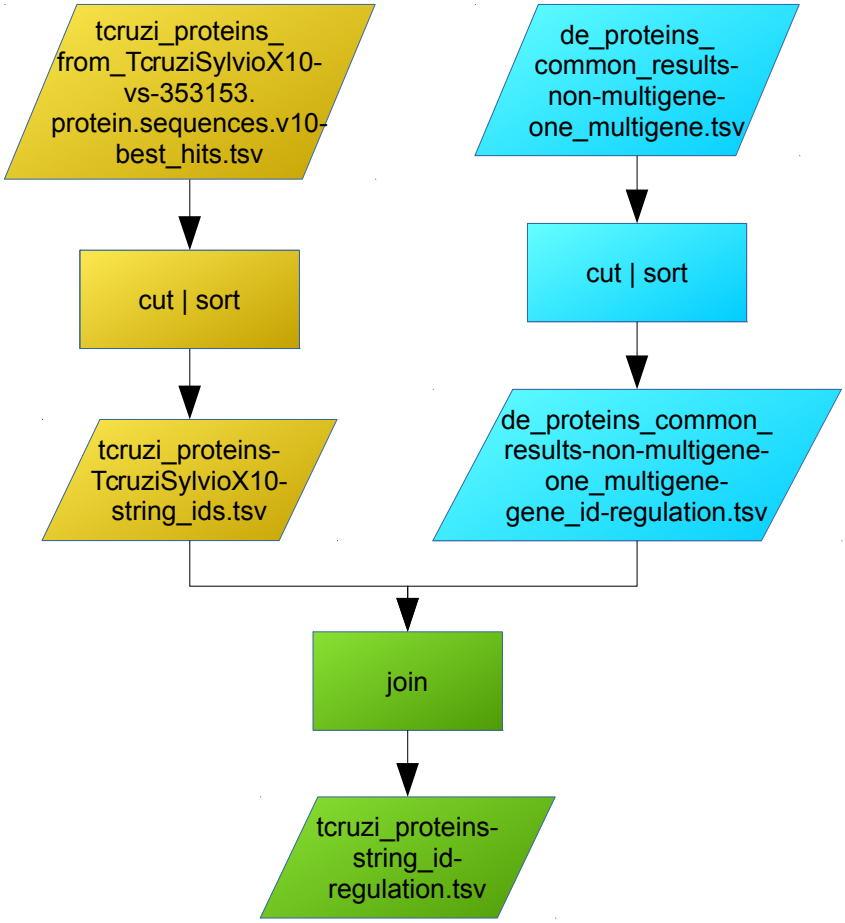


FIGURA 8: Fluxograma dos procedimentos utilizados na criação das categorias de regulação associadas às proteínas.

A partir do diretório */proteins*, geramos o arquivo *tcruzi_proteins-TcruziSylvioX10-string-ids.tsv,* com a correlação entre os *gene ids* e os *STRING ids*. Para isso, usamos a sequência de comandos abaixo.

```
cut -f1,4 blast/tcruzi_proteins_from_TcruziSylvioX10-vs-
353153.protein.sequences.v10-best_hits.tsv | sort >
work/tcruzi_proteins-TcruziSylvioX10-string_ids.tsv
```

Em seguida, a partir do diretório */proteins/work*, utilizamos a sequência de comandos abaixo para extrair as colunas *Gene_id* e *Regulation* do arquivo *de_proteins_common_results-non-multigene-one_multigene.tsv*, ordenando o resultado pela coluna *Gene_id*. Com isso, foi gerado o arquivo *de_proteins_common_results-non-multigene-one_multigene-gene_id-regulation.tsv*.

```
cut -f1,22 ../de_proteins_common_results-non-multigene-
one_multigene.tsv | sort > de_proteins_common_results-non-
multigene-one_multigene-gene_id-regulation.tsv
```

Por fim, associamos as informações dos dois arquivos gerados em um terceiro, *de_proteins_common_results-non-multigene-one_multigene-gene_id-regulation.tsv*.

```
join --header -o 1.2,2.2 -t $'\t' tcruzi_proteins-
TcruziSylvioX10-string_ids.tsv de_proteins_common_results-non-
multigene-one_multigene-gene_id-regulation.tsv | sort -k2 >
tcruzi_proteins-string_id-regulation.tsv
```

3.7.2 Informação funcional: Natureza da interação entre as proteínas

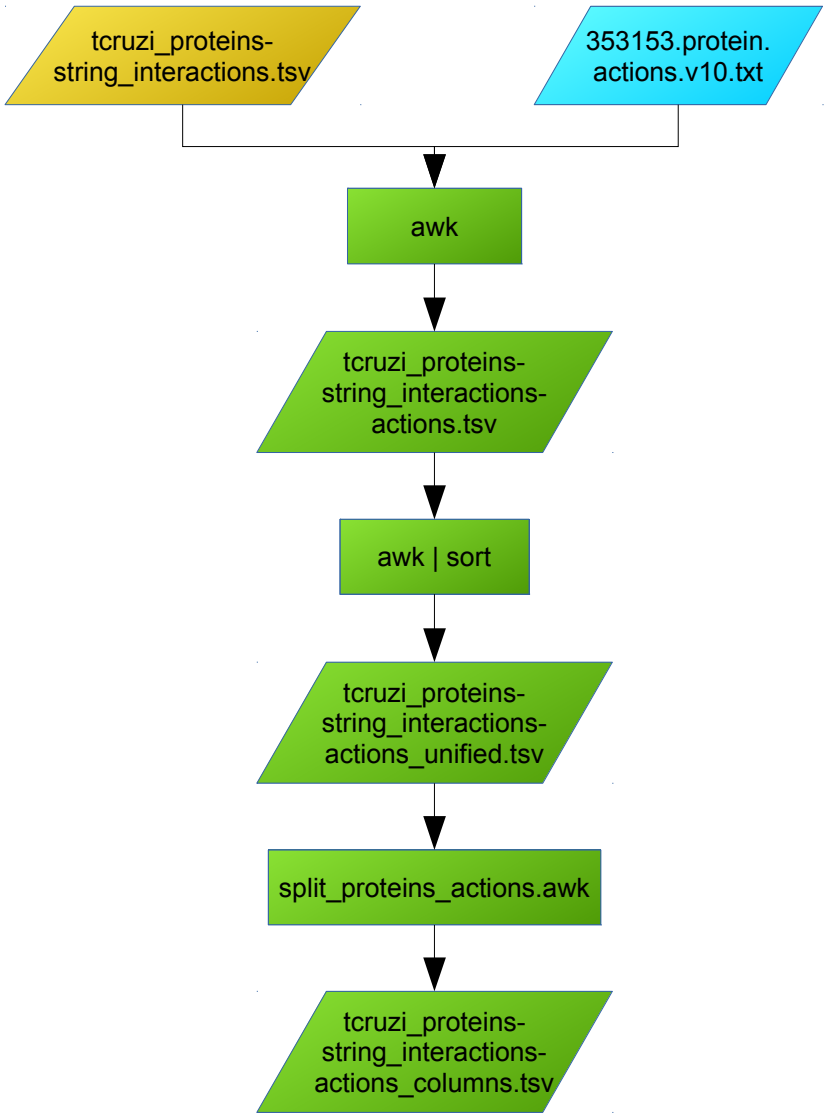


FIGURA 9: Fluxograma dos procedimentos utilizados na caracterização das interações entre as proteínas.

O arquivo com todos os tipos de interações disponíveis para *T. cruzi* pode ser obtido também no link de *Downloads* do site *STRING*. Mais uma vez, após selecionar *Trypanosoma cruzi* como organismo, fizemos a transferência do arquivo *353153.protein.actions.v10.txt.gz* para o diretório */proteins/string*. O arquivo foi descompactado após o download.

Nesse mesmo diretório, fizemos a separação dos tipos de interações disponíveis em nossa rede com o comando:

```
awk 'NR==FNR{nodes[$5,$6];next} ($1,$2) in nodes'
tcruzi_proteins-string_interactions.tsv
353153.protein.actions.v10.txt > tcruzi_proteins-
string_interactions-actions.tsv
```

Em seguida, já no diretório */proteins/work*, unificamos, na mesma linha, conexões que apresentaram mais de um tipo de interação, alterando a saída para adequar à importação no *Cytoscape*:

```
awk 'BEGIN {print "Interaction\tActions"} {FS="\t"; a[$1"
(interacts with) "$2] = a[$1" (interacts with) "$2]$3" "} END
{for(i in a){print i"\t"a[i]}}' ../string/tcruzi_proteins-
string_interactions-actions.tsv | sed 's/ $//' >
tcruzi_proteins-string_interactions-actions_unified.tsv
```

Por fim, criamos um programa em *awk* (*split_proteins_actions.awk*)para separar cada um dos tipos de interação em colunas distintas, salvamos no diretório /bin e o executamos com o seguinte comando, a partir do diretório */proteins/work*:

```
../../bin/split_proteins_actions.awk tcruzi_proteins-
string_interactions-actions_unified.tsv > tcruzi_proteins-
string_interactions-actions_columns.tsv
```

3.7.3 Informação funcional: *EC numbers*

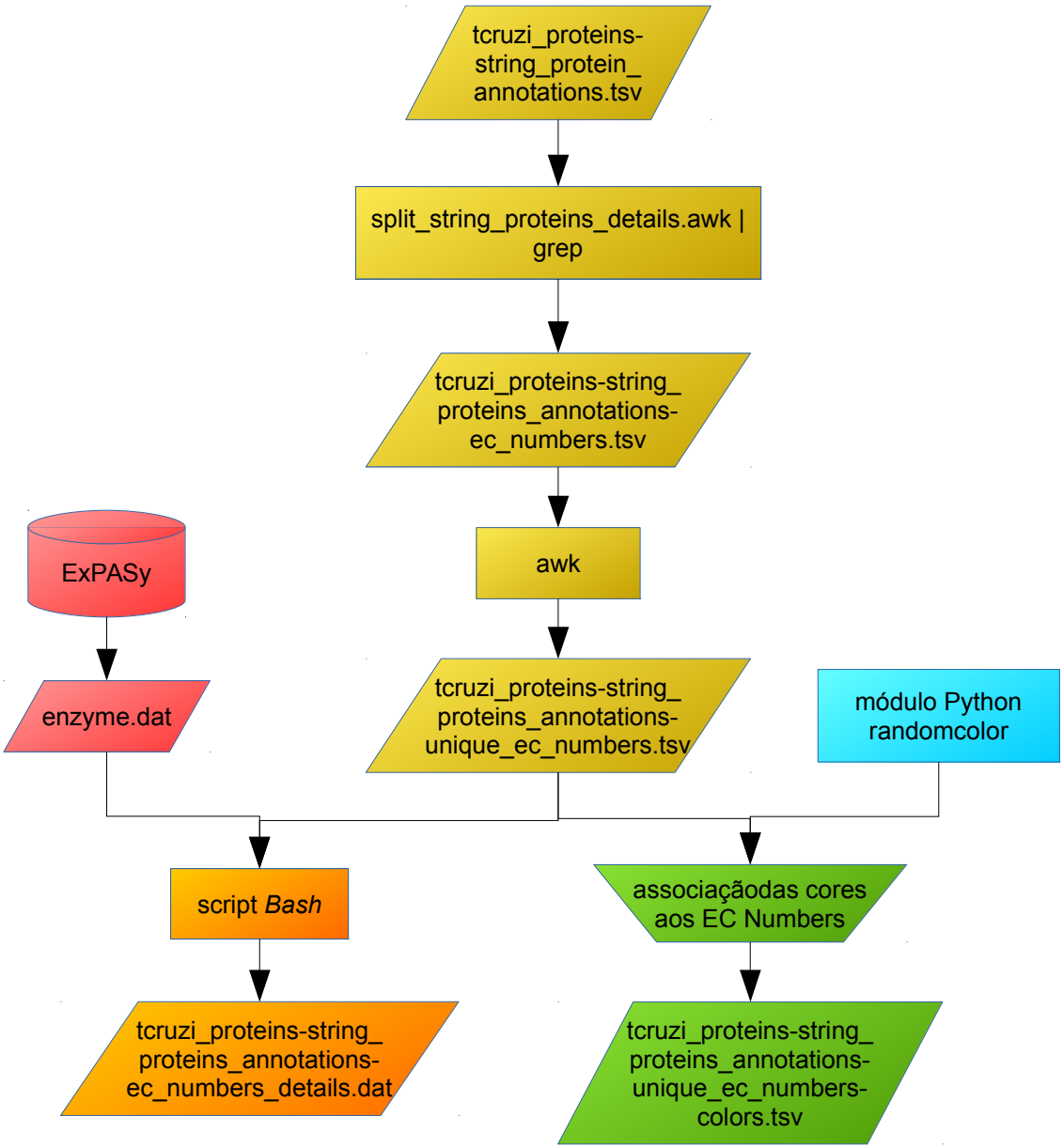


FIGURA 10: Fluxograma dos procedimentos utilizados na associação dos *EC numbers* às proteínas e às cores específicas utilizadas na construção das redes. Também descreve os procedimentos de extração das informações dos *EC numbers* encontrados a partir do banco de dados de enzimas.

Criamos um programa em *awk* (*split_string_proteins_details.awk)*, no diretório */bin*, e o utilizamos em conjunto com o comando *grep* para extrair os dados de *Enzyme Commission number* (*EC number*) disponíveis no arquivo *tcruzi_proteins-*

string_protein_annotations.tsv. A sintaxe dos comandos, executados a partir do diretório */proteins/work*, é a seguinte:

```
../../bin/split_string_proteins_details.awk
../string/tcruzi_proteins-string_protein_annotations.tsv |
grep 'string_id\|EC:' > tcruzi_proteins-
string_proteins_annotations-ec_numbers.tsv
```

Com isso, foi gerado o arquivo *tcruzi_proteins-string_proteins_annotations-ec_numbers.tsv*, com a relação de proteínas que possuem *EC number*, segundo o banco de dados *STRING*. Em seguida, ainda no mesmo diretório, isolamos cada um dos *EC numbers* únicos desse arquivo em um novo, denominado *tcruzi_proteins-string_proteins_annotations-unique_ec_numbers.tsv*:

```
awk 'BEGIN {FS="\t"; print "ec_number"} !a[$2]++{if (/EC:/)
{print $3 | "sort"}}' tcruzi_proteins-
string_proteins_annotations-ec_numbers.tsv > tcruzi_proteins-
string_proteins_annotations-unique_ec_numbers.tsv
```

A partir desse arquivo, foi gerado um novo (*tcruzi_proteins-string_proteins_annotations-unique_ec_numbers-colors.tsv*), associando cada um dos *EC numbers* encontrados a um código, que é a representação da cor em notação *RGB* hexadecimal. As cores foram geradas automaticamente, utilizando o módulo de geração aleatória de cores *randomcolor* para a linguagem *Python*, e associadas manualmente aos *EC numbers* isolados.

Buscamos os detalhes de cada um dos *EC numbers* isolados a partir do banco de dados de enzimas mantidos pelo portal *ExPASy* (http://enzyme.expasy.org/). Para isso, inicialmente transferimos, a partir do diretório */original_data,* o arquivo *enzyme.dat*, que contém informações sobre todos os *EC numbers* disponíveis no *ExPASy*:

```
wget ftp://ftp.expasy.org/databases/enzyme/enzyme.dat
```

A busca foi realizada com o *script Bash* abaixo, a partir do diretório */proteins/work*:

```
for f in $(cat tcruzi_proteins-string_proteins_annotations-
unique_ec_numbers.tsv); do ec=$(echo $f | cut -d":" -f2); sed
-n -e "/ID   $ec$/,/\/\// p" ../../original_data/enzyme.dat;
done > tcruzi_proteins-string_proteins_annotations-
ec_numbers_details.dat
```

3.7.4 Informação funcional: Anotação funcional das proteínas

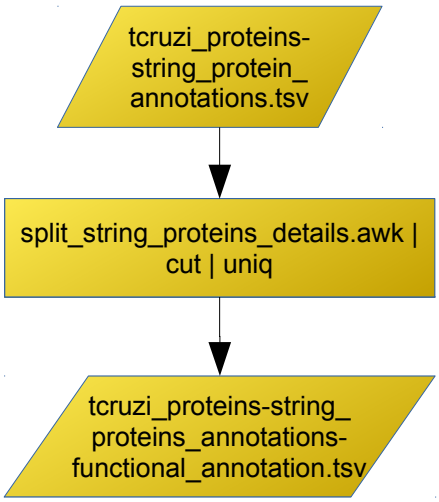


FIGURA 11: Fluxograma do procedimento utilizado na extração das informações de anotação funcional das proteínas da rede, a partir do arquivo de anotações disponibilizado pelo *STRING*.

Utilizamos novamente o programa *split_string_proteins_details.awk*, dessa vez filtrando sua saída com os comandos *cut* e *uniq*, para extrairmos as anotações funcionais de cada uma das proteínas presentes em nossa rede. Essa informação também foi obtida a partir do arquivo *tcruzi_proteins-string_protein_annotations.tsv*, utilizando a seguinte linha de comandos, a partir do diretório */proteins/work*:

```
../../bin/split_string_proteins_details.awk
../string/tcruzi_proteins-string_protein_annotations.tsv | cut
-f1,2 | uniq > tcruzi_proteins-string_proteins_annotations-
functional_annotation.tsv
```

3.8 Modelagem das redes no Cytoscape

Pela sua característica descritiva, essa seção não apresenta um fluxograma específico.

Uma vez preparados os arquivos, utilizamos o software *Cytoscape*, versão 3.4.0, para a reconstrução das redes e a incorporação dos dados.

3.8.1 Entrada e processamento dos dados

Inicialmente importamos, no *Cytoscape*, o arquivo de interações *tcruzi_proteins-string_interactions.tsv*, cuja obtenção foi descrita na seção 3.5. Ele foi aberto a partir do menu **File** → **Import** → **Network** → **File**, onde foi então localizado e confirmada sua abertura. Nesse momento, abre-se uma tela com os detalhes da importação.

Ao abrir um arquivo de importação, o *Cytoscape* reconhece automaticamente as colunas e as exibe separadamente, uma ao lado da outra, na janela de importação. É necessário definir o papel de cada uma das colunas em relação aos dados. Isso é feito clicando no título de cada coluna. Neste trabalho, nós também renomeamos as colunas. Em primeiro lugar, zeramos as configurações pré-definidas, usando o botão **Select None**. Em seguida alteramos cada uma das colunas, de acordo com a listagem abaixo:

- **#node1:** nome alterado para *source_data* e selecionado o ícone correspondente a *Source Node Attribute* (sexto ícone, folha de papel verde);
- **node2:** nome alterado para *target_data* e selecionado o ícone correspondente a *Target Node Attribute* (sétimo ícone, folha de papel laranja);
- **node1_external_id:** nome alterado para *source* e selecionado o ícone correspondente a *Source Node* (segundo ícone, bola verde);
- **node2_external_id:** nome alterado para *target* e selecionado o ícone correspondente a *Target Node* (quarto ícone, bola laranja vazada);
- **combined_score:** o nome não foi alterado e o ícone selecionado foi correspondente a *Edge Attribute* (quinto ícone, folha de papel roxa).

Após clicar em **Ok**, a rede é gerada.

Foi criada mais uma coluna na tabela de nós (*Node Table*), visível abaixo da visualização da rede. Esse passo é importante para que tenhamos, e uma única coluna, a relação de todos os identificadores de proteínas externos ao *STRING*, apresentados nas colunas *source_data* e *target_data*.

Clicando no ícone com o sinal **+**, aparecem a opções de criação da tabela. Selecionamos **New Single Column** → **String** e nomeamos a coluna como *protein_alias*. Em seguida, clicamos na primeira célula dessa coluna e entramos com a seguinte fórmula (a entrada é uma linha única):

```
=IF(${source_data:""}="",${target_data:""},${source_data:""})
```

Utilizando a opção do próprio *Cytoscape* (*Apply to entire column*, disponível ao se clicar com o botão direito na célula da fórmula), essa fórmula foi aplicada à coluna inteira.

Em seguida, renomeamos a coleção de redes (informação disponível na lateral esquerda da tela). A coleção recebeu o nome de *T. cruzi - proteins*. Também renomeamos o nome da primeira rede gerada para *STRING - website*.

Demos início, então, à importação dos dados gerados anteriormente e salvos no diretório */proteins/work*. Para isso, utilizamos a opção disponível em **File** → **Import** → **Table** → **File**. Após selecionar cada um dos arquivos, alteramos/confirmamos uma série de parâmetros na janela que se abre, entre eles o nome e atributo das colunas de dados (de forma semelhante ao realizado na importação inicial da rede). A relação dos nomes dos arquivos importados (todos presentes no diretório */proteins/work*), com as respectivas alterações, encontra-se listada abaixo:

- **tcruzi_proteins-TcruziSylvioX10-string_ids.tsv:** renomear a coluna *Query name* para *gene_id* e defini-la como *Attribute* (terceiro ícone, folha de papel); definir a coluna *Subject name* como *Key* (segundo ícone, chave); definir **Key Column for Network** como *shared name*;

- **de_proteins_common_results-non-multigene-one_multigene-gene_id-regulation.tsv:** definir a coluna *Regulation* como *Attribute*; definir a coluna Gene_id como Key; definir **Key Column for Network** como *gene_id*;
- **tcruzi_proteins-string_proteins_annotations-ec_numbers.tsv:** renomear a coluna *protein_info* para *ec_number* e defini-la como *Attribute*; definir a coluna *string_id* como Key; definir **Key Column for Network** como *shared name*;
- **tcruzi_proteins-string_proteins_annotations-unique_ec_numbers-colors.tsv:** renomear a coluna *colors* para *ec_number_color* e defini-la como *Attribute*; definir a coluna *ec_number* como *Key*; definir **Key Column for Network** como *ec_number*;
- **tcruzi_proteins-string_proteins_annotations-functional_annotation.tsv:** definir a coluna *string_id* como *Key*; definir a coluna *functional_annotation* como *Attribute*; definir **Key Column for Network** como *shared name*;
- **tcruzi_proteins-string_interactions-actions_unified.tsv:** definir a coluna *Interaction* como *Key*; definir a coluna *Actions* como *Attribute*; definir **Import data as:** *Edge Table Columns*; definir **Key Column for Network** como *shared name*;
- **tcruzi_proteins-string_interactions-actions_columns.tsv:** definir a coluna *Interaction* como *Key*; definir todas as outras colunas como *Attribute*; definir **Import data as:** *Edge Table Columns*; definir **Key Column for Network** como *shared name*.

Por fim, importamos os dados da rede com todas as interações de proteínas de *T cruzi* disponíveis no *STRING*, gerada na seção 3.6. Para isso, mais uma vez fomos ao menu **File** → **Import** → **Network** → **File** e abrimos o arquivo *353153.protein.links.detailed.v10-experiment_database-combined_400.tsv*, presente no diretório */proteins/string*. Na janela que se abre, entramos na opção **Advanced Options…** e selecionamos **SPACE** como separador de campos (item **Delimiter:**). Como no procedimento anterior, inicialmente limpamos as configurações pré-definidas de cada uma das colunas, com o botão **Select None**. Dessa vez, nenhuma

das colunas teve seu nome alterado, sendo alterados somente os papeis de cada uma, da seguinte forma:

- **protein1:** selecionado o ícone correspondente a *Source Node* (segundo ícone, bola verde);
- **protein2:** selecionado o ícone correspondente a *Target Node* (quarto ícone, bola laranja vazada);
- **combined_score:** selecionado o ícone correspondente a *Edge Attribute* (quinto ícone, folha de papel roxa).

Após confirmar as alterações, a nova rede é acrescentada ao conjunto *T. cruzi – proteins*. Ao final da estruturação da rede, o *STRING* nos deu a opção de criar sua visualização. Optamos por não fazê-lo nesse momento, por motivos de performance. Renomeamos a rede para *STRING - DB/Experiments - Combined 400*.

Ao final, eliminamos as arestas duplicadas, através do menu **Edit → Remove Duplicated Edges...** Na janela que se abre, selecionamos a rede *STRING - DB/Experiments - Combined 400* e a opção **Ignore edge direction**. Após algum tempo o *STRING* nos indicou o total de arestas eliminadas. Por fim, clicamos em **Create View** para que a visualização da rede fosse criada.

3.8.2 Análise da rede

Executamos uma análise global da rede, usando um recurso próprio do *Cytoscape,* denominado *NetworkAnalyzer*. Ele está acessível a partir do menu **Tools → NetworkAnalyzer → Network Analysis → Analyze Network**. Na janela que se abre, selecionamos a opção para rede não direcional (**Treat this network as undirected.**). Isso gerou uma série de novas colunas, usadas como entrada para outras análises.

Com o objetivo de apresentar as informações de enriquecimento funcional adicionadas posteriormente, utilizamos o recurso de “estilos” (*style*) do *Cytoscape*, que permite ajustes finos na forma e aparência da rede. Isso foi feito através da associação dos dados importados para a nossa rede em conjunto com os estilos.

4 RESULTADOS E DISCUSSÃO

Considerando-se a natureza deste trabalho, que é a construção de uma sistemática de integração de dados em RIPPs, o desenvolvimento da metodologia é o nosso resultado principal e não somente um meio para se chegar ao resultado. Assim, cada etapa metodológica foi planejada e testada exaustivamente para que fosse reprodutível nas mais diversas situações e permitisse sua implementação inclusive para outros organismos modelo.

Tendo isso em vista, os resultados de cada uma das etapas metodológicas está descrito em detalhes, seção por seção, para permitir, inclusive, que elas sejam implementadas separadamente ou adaptadas para atender a outros protocolos.

4.1 Obtenção e consenso dos dados

Conforme descrito no item 3.1 da metodologia, os dados iniciais utilizados neste trabalho, foram obtidos a partir do resultado de um projeto de pesquisa envolvendo diversos pesquisadores, que trabalharam com dados de sequenciamento massivo de RNA e proteínas de dois clones de *T. cruzi* com diferentes perfis de infectividade.

A partir desses dados, foi gerado um consenso entre genes e proteínas, ambos diferencialmente expressos. Como resultado foram selecionados 233 genes/proteínas. Conforme descrito na metodologia, eles foram agrupados de acordo com a regulação de RNAs e proteínas (regulação negativa – neg; regulação positiva – pos) e organizados em arquivos separados. Cada arquivo reuniu um dos quatro agrupamentos, a saber: RNA neg / proteína neg (26 genes), RNA neg / proteína pos (8 genes), RNA pos / proteína neg (2 genes) e RNA pos / proteína pos (197 genes). Esses dados estão representados na TABELA 2.

Uma vez que estamos trabalhando com a representação das proteínas codificadas por esses genes em um grafo, cada proteína deve corresponder a um único nó. Por isso, foi separado um representante para cada família multigênica, uma vez que genes diferentes da mesma família multigênica codificam a mesma proteína. A relação com esses representantes também foi armazenada em um arquivo à parte. Importante destacar que na representação gráfica das redes geradas cada nó

anotado como “familia multigênica”, diferentemente do resto da rede, representa a existência de, pelo menos, dez cópias do referido gene.

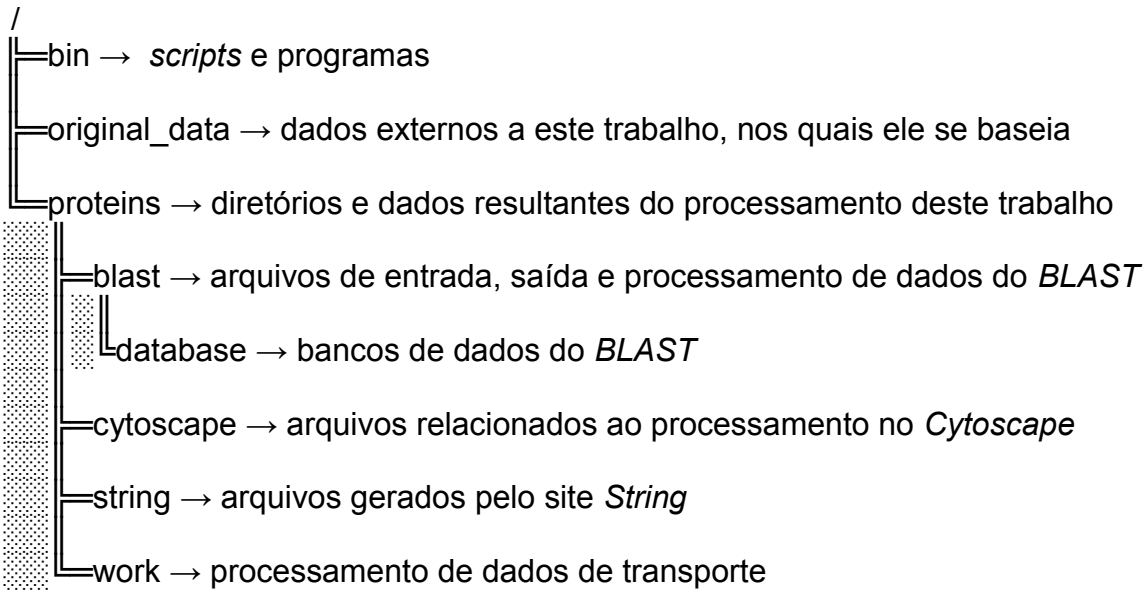
TABELA 2: Relacionamentos entre pares: Regulação positiva e negativa de RNA e proteína para os genes em estudo.

	RNA neg / proteína neg	**RNA neg / proteína pos**	**RNA pos / proteína neg**	**RNA pos / proteína pos**
Genes diferencialmente expressas após o corte de logFC ± 0.99 e p-valor ajustado menor que 0.05	26	8	2	197
Genes pertencentes a famílias multigênicas	19/26	5/8	1/2	187/197
Genes não pertencentes a famílias multigênicas	7/26	3/8	1/2	10/197
Famílias multigênicas identificadas	5	3	1	7
Genes não pertencentes a famílias multigênicas + um representante de cada família multigênica	12/26	6/8	2/2	17/197

4.2 Configuração do espaço de trabalho para o desenvolvimento do projeto

A definição da árvore de diretórios foi uma etapa importante deste trabalho, uma vez que, ao executar cada uma das linhas de comando, são referenciados arquivos de

entrada e de saída. Por isso foi criada uma estrutura pré-definida, que está descrita abaixo, com a explicação de cada conteúdo associado:



Os nomes dos arquivos também possuem uma lógica pré-definida. Eles representam a origem, as adições/alterações sofridas ao longo dos processamentos e o contexto a que se referem. Arquivos que foram gerados por processamento desse trabalho, ou que tenham sido transferidos a partir do *STRING*, possuem o prefixo *tcruzi_proteins*. Por exemplo, o arquivo *tcruzi_proteins-string_interactions.tsv* contém dados de interações de proteínas e foi obtido no *site* do *STRING*.

Ao longo das análises realizadas, utilizamos alguns *scripts* para automatizar as tarefas de processamento computacional. Esses *scripts* são sequências de comandos do *shell* do *GNU/Linux*, criados para funcionar no interpretador de comandos *Bash* (mas também passíveis de serem reconhecidos por outros interpretadores de comandos do *GNU/Linux*, do *MacOSX* e mesmo do *Windows 10*). Esses códigos de programação serão referenciados neste trabalho como *scripts shell*. A não ser quando declarado, todos os *scripts* citados nessa metodologia foram digitados em uma única linha (*script* de uma linha). Utilizamos também programas construídos com as linguagens de programação *Perl* e *Awk*. Esses programas foram armazenados no diretório */bin*.

4.3 Construção dos arquivos utilizados na origem dos dados de análise

A escolha pela extensão ".tsv" para o arquivo *de_proteins_common_results-non-multigene-one_multigene-only_gene_ids.tsv* teve por objetivo a adequação à nomenclatura mais correta. Em relação à sua estrutura, arquivos *TSV* (*Tab-Separated Values*) e *CSV* (*Comma-Separated Values*) são idênticos: ambos são arquivos em formato texto puro constituído por colunas delimitadas por um separador comum. O que os diferencia é o separador: por definição, arquivos *CSV* deveriam usar a vírgula e arquivos *TSV*, tabulações. Com os problemas associados ao uso da vírgula como separador, por exemplo, a possibilidade da existência de dados que incluam a sinal gráfico, o formato *TSV* surgiu como uma alternativa que eliminasse esse tipo de ambiguidade. Entretanto, ainda é comum o uso de arquivos que usem tabulações como separadores, mas possuam a extensão *.csv*.

4.4 Pesquisa pelos identificadores do *STRING* utilizando busca por similaridade de sequência

Existem duas formas de procurar por proteínas no *STRING*: fornecendo os identificadores das proteínas ou através da busca por similaridade de sequências de aminoácidos, realizada no próprio site do *STRING*.

A opção mais simples seria a utilização dos identificadores já associados às proteínas de estudo. Entretanto, devido à ausência de um identificador único que possa ser empregado em todos os bancos de dados, existe o risco de ambiguidades e duplicidades desses identificadores entre os diferentes bancos (PHILIPPI; KÖHLER, 2006). Além disso, as frequências de atualizações desses bancos são variáveis. Assim, o identificador de uma proteína em um banco de dados pode estar desatualizado em relação a outros. Por esse motivo, optou-se pela utilização das proteínas selecionadas como positiva ou negativamente reguladas nas buscas por similaridade de sequências para identificação da correlação com os *STRING ids.*

Apesar de o *STRING* oferecer a opção por esse tipo de busca, os resultados fornecidos pelo site apresentam somente as seguintes informações de saída: identificador da proteína, nome do organismo, anotação relacionada a essa proteína e o *bitscore* e o *e-value* do resultado encontrado. Assim sendo, como não foi possível avaliar neste resultado os parâmetros analíticos estabelecidos para o

desenvolvimento deste trabalho (taxa de identidade e de cobertura), optamos por realizar as buscas por similaridade em servidor local.

Importante destacar que no caso de famílias multigênicas, a proteína escolhida como *query* foi aquela apresentando maior cobertura e similaridade.

Para a execução da busca por similaridade de sequências com o *BLAST,* é necessário criar um banco de dados que servirá como alvo da busca. Uma vez que é possível fazer a transferência, para um computador local, de arquivos contendo as sequências de aminoácidos de todas as proteínas armazenadas para cada um dos organismos cadastrados no site do *STRING*, utilizamos esse recurso para a criação do banco de dados do *BLAST*.

Executada a busca, iniciamos o trabalho de interpretação dos dados. Para agilizar essa tarefa, criamos um programa em *Perl* que extrai uma série de informações do resultado bruto do *BLAST*. Esse programa foi denominado, *blast-extract-info.pl* e encontra-se detalhado no Apêndice 2. Para este trabalho nos interessavam duas informações que relacionavam a sequências de entrada (*query*) com o alvo no banco de dados (*subject*): as taxas de identidade (*identity*) e cobertura (calculada a partir do percentual de sobreposição da *query* sobre o *subject*). A partir do arquivo de saída gerado com o *script* anterior (*tcruzi_proteins_from_TcruziSylvioX10-vs-353153.protein.sequences-info_extraction.tsv*), selecionamos os *best hits* para cada um dos resultados do *BLAST*, considerando-se as informações de identidade e cobertura. O critério de corte estabelecido foi de que ambas as taxas fossem superiores a 90%. Dentro dos parâmetros descritos acima 21 correlações distintas foram estabelecidas entre as proteinas utilizadas como query e os *STRING ids. O* arquivo *tcruzi_proteins_from_TcruziSylvioX10-vs-353153.protein.sequences.v10-best_hits.tsv* listando essas correlações encontra-se no Apêndice 3.

4.5 Estruturação da rede no *STRING*

Após utilizarmos a coluna de *STRING ids* obtida através do estabelecimento das correlações descritas no item 4.4 como dado de entrada no *STRING,* a representação gráfica inicial da rede em estudo foi gerada (FIGURA 12).

A lista de proteínas fornecida inicialmente como dado de entrada (*STRING ids*) não possui informação de interação entre elas. Essa atribuição é realizada diretamente pelo *site* e, caso não possa ser estabelecida, o *STRING* as desenha como esferas isoladas.

Como esse resultado evidenciou somente a presença de esferas (nós) isoladas, iniciamos um estudo visando o estabelecimento de parametrizações que viabilizassem o estabelecimento das relações entre as proteínas em questão e seus vizinhos.

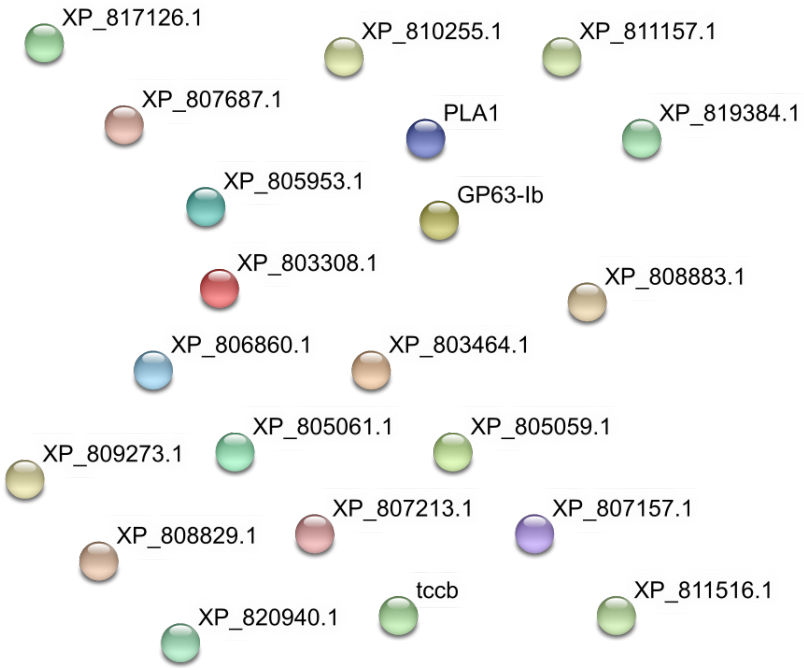


FIGURA 12: Representação gráfica inicial da rede gerada pelo STRING. As cores atribuídas não possuem nenhum significado, sendo apenas ilustrativas. Na representação em questão não existe conexão direta entre as proteínas.

Dentro desta perspectiva, avaliamos os resultados relacionados à alteração dos valores da variável "*1st shell*" que passou do valor 0 (zero) para 500. Está variável define os primeiros vizinhos de cada uma das proteínas pesquisadas, e o valor 500 define o total de interações do conjunto de todas as proteínas mapeadas na rede e não delas individualmente. Além disso, esse valor garante o estabelecimento do maior número possível de vizinhos na rede.

Adicionalmente, tendo como objetivo manter um equilíbrio entre o rigor na curadoria da qualidade dos dados e um volume de interações plausíveis de serem analisadas, utilizamos os seguintes opções, como fontes de interações (*active interaction sources*): *experiments* e *databases.* Neste contexto, o valor de escore mínimo de interação (*minimum required interaction score*) escolhido foi 400. Este valor define, dentro do contexto do banco STRING, um grau de confiança média das interações. O resultado final das configurações pode ser visto na FIGURA 13.

A rede final gerada pelo *STRING*, a partir desses parâmetros, é composta por 70 nós e 529 arestas, e pode ser visualizada na FIGURA 14. Essa imagem nos permite fazer uma análise visual da rede como um todo, contudo visando avaliar se a rede gerada apresentava características de uma rede biológica iniciamos uma investigação mais aprofundada dos elementos e métricas da rede. Considerando que o *STRING* carece de ferramentas mais adequadas para esse tipo de análise, utilizamos a opção de exportação de dados do site.

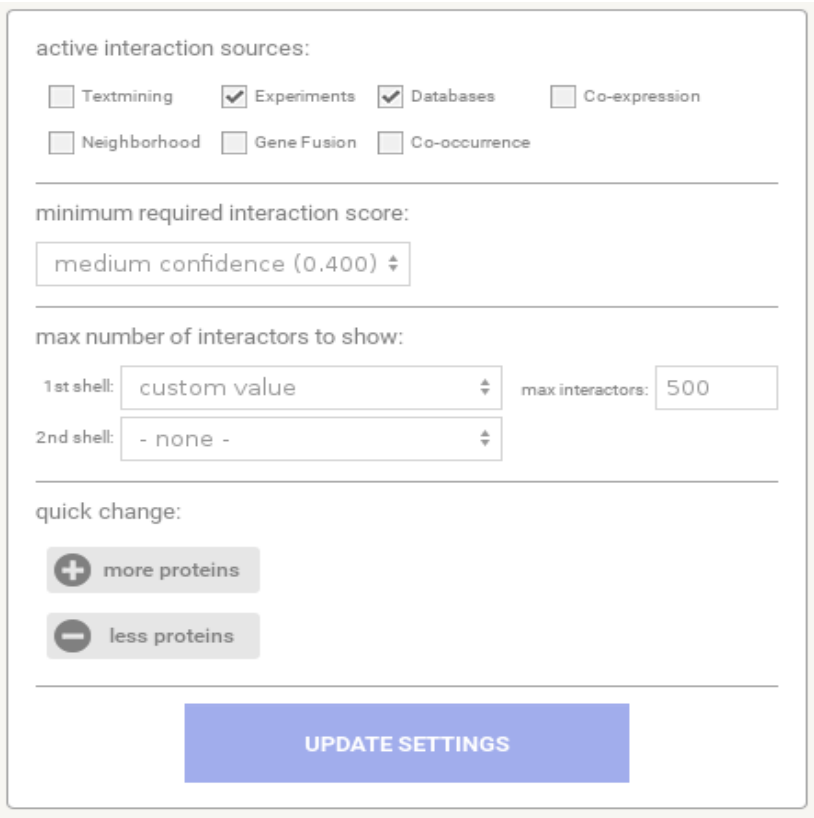


FIGURA 13: Parâmetros aplicados ao STRING para a montagem da rede adicionando informações dos primeiros vizinhos e critérios de qualidade.

Os arquivos baixados estão listados abaixo, com os nomes atribuídos entre parênteses e uma breve descrição de cada um:

- … *as simple tabular text output* (*tcruzi_proteins-string_interactions.tsv*): arquivo com as interações entre as proteínas;
- … *protein sequences* (*tcruzi_proteins-string_protein_sequences.fa*): sequências de aminoácidos de cada uma das proteínas presentes na rede;
- … *protein annotations* (*tcruzi_proteins-string_protein_annotations.tsv*): informações extras sobre cada uma das proteínas presentes na rede, tais como descrição, outros nomes e informações de anotação funcional.

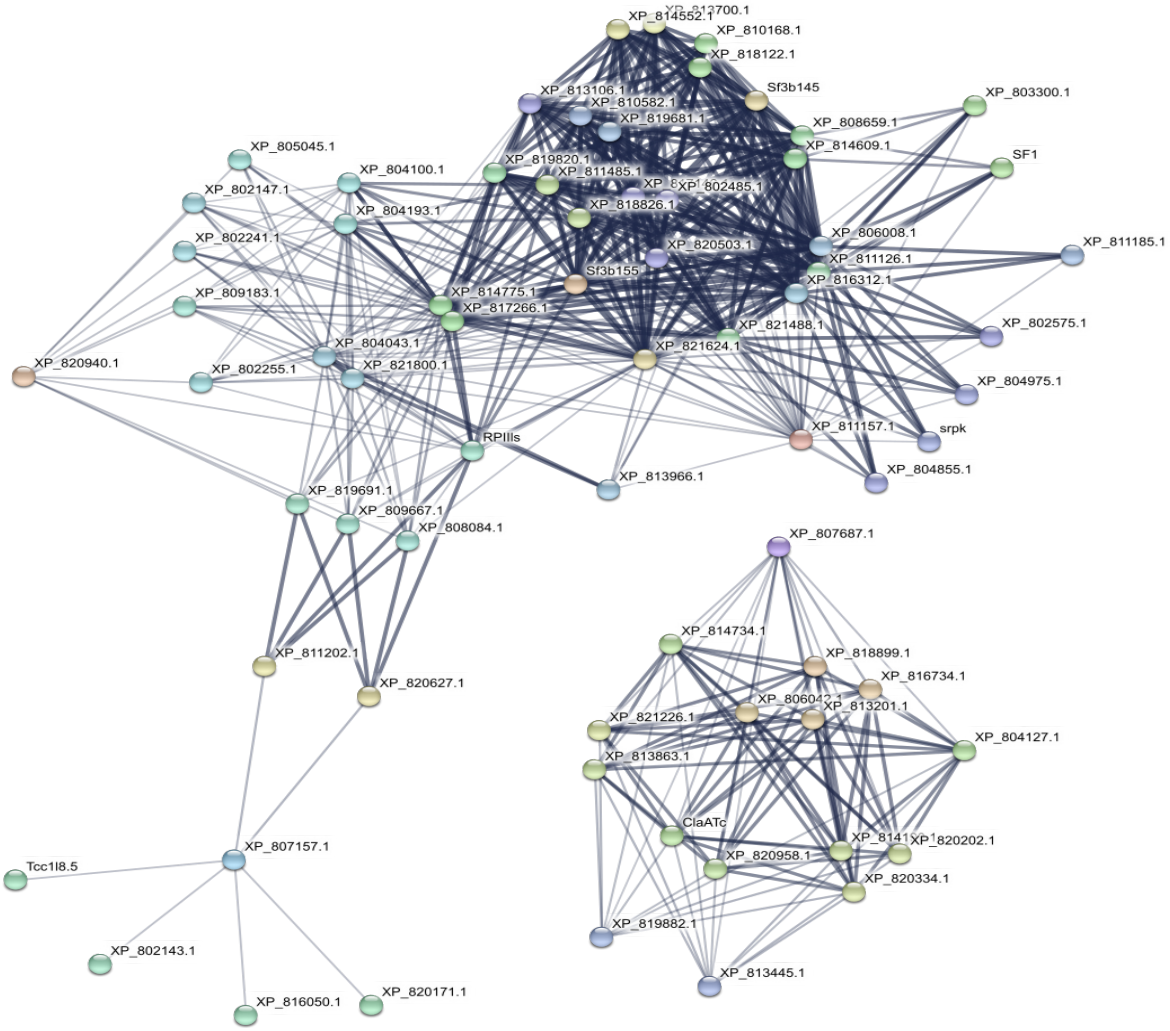


FIGURA 14: Rede produzida no *STRING*, a partir das proteínas selecionadas (regulação gênica positva e negativa) associadas aos seus primeiros vizinhos. Os vizinhos foram obtidos no banco de dados do *STRING*, a partir dos critérios de corte: *minimum required interaction score*: 0.400 e *active interaction sources*: *Experiments* e *Databases*.

4.6 Obtenção e filtragem da rede com todas as interações de proteínas de *T. cruzi* disponíveis no *STRING*

Uma vez que um dos objetivos deste trabalho e analisar métricas da rede estruturada na etapa 3.5, buscamos o arquivo contendo o total de interações de proteínas do *T. cruzi*, disponível no *STRING*. Isso se fez necessário porque, ao avaliar as métricas da rede gerada, devemos considerá-la como uma sub-rede do total de interações proteicas do organismo. Assim, os elementos presentes nessa sub-rede, ao serem considerados no contexto da rede inteira, apresentam alterações significativas nos valores de determinados parâmetros avaliados, tais como os de *centralidade* e de *grau.*

Uma vez obtido o arquivo com o total de interações, também precisamos adequá-lo aos critérios de corte utilizados também na etapa 3.5, a saber: experimentos e bancos de dados como fontes de interação, com um escore mínimo de 400. Efetuamos essa adequação através de uma linha de comando em *awk*, utilizada para calcular o escore adequado e filtrou as entradas que atendessem a esse escore.

Inicialmente nos baseamos na fórmula disponível na bibliografia (MERING et al., 2005) para fazermos a filtragem. A fórmula é a seguinte:

$$S_c = 1 - \prod_i (1 - S_i)$$

onde S_c é o escore combinado e S_i, cada um dos escores individuais disponíveis para aquela proteína.

Como controle da acuidade do cálculo, decidimos testar a fórmula utilizando os dados já disponíveis para a rede gerada. Em nosso teste inicial ocorreu uma discrepância entre o resultado esperado e o resultado calculado. Encontramos, então, na página de perguntas frequentes do *STRING*, um novo cálculo para o escore combinado, que incluía uma constante, denominada *prior* no cálculo. O valor de *prior* segundo definido na página deveria ser 0,063. A nova fórmula foi assim representada:

$$S_c = 1 - (1 - 0{,}063) \cdot \prod_i ((1 - S_i/1000)/(1 - 0{,}063))$$

Entretanto, os resultados dessa fórmula também apresentaram discrepâncias com o esperado. Entramos então em contato com a equipe do *STRING* e, após troca de algumas correspondências eletrônicas, eles descobriram que o valor de *prior* apresentado na página estava errado e nos enviaram o valor correto, que é 0,04113021. A fórmula definitiva ficou então:

$$S_c = 1 - (1 - 0.04113021) \cdot \prod_i ((1 - S_i/1000)/(1 - 0.04113021))$$

Os resultados obtidos com essa fórmula foram correspondentes aos esperados, por isso essa foi a fórmula usada em nosso comando *awk*. Ao final do comando, foi obtida uma rede com 4.897 nós (proteínas) e 129.091 arestas, denominada *353153.protein.links.detailed.v10-experiment_database-combined_400.tsv*. O conteúdo desse arquivo não se encontra em anexo devido ao seu número excessivo de linhas (258.183). Mas ele pode ser facilmente obtido aplicando-se a metodologia descrita na seção 3.6.

4.7 Preparação dos dados utilizados no enriquecimento funcional da rede

Apesar das informações compiladas e dados gerados até a presente etapa de desenvolvimento do projeto serem suficientes para a criação da estrutura da RIPP, com o objetivo de caracterizar funcionalmente essa rede e ampliar as possibilidades de cálculos de suas métricas, adicionamos algumas outras informações aos seus nós e arestas. Para tanto, geramos uma série de arquivos com resultados de análises e informações funcionais que foram sucessivamente incorporados à rede com atribuições de característica iconográfica ou de cor para melhor visualização. Todos esses arquivos foram salvos no diretório */proteins/work*.

4.7.1 Informação funcional: Categorias de regulação RNA/proteína

Como estratégia de integração da informação funcional relacionada ao controle positivo e negativo de expressão gênica (vide TABELA 2), separamos as 20 proteínas originalmente fornecidas ao *STRING* para a construção da rede, além de associá-las ao seu grupo de regulação de expressão. A listagem dessas proteínas e respectiva informação funcional encontram-se no arquivo *tcruzi_proteins-string_id-regulation.tsv* (Apêndice 4).

4.7.2 Informação funcional: Natureza da interação entre as proteínas

Com o propósito de contextualizar a natureza das interações entre as proteínas, extraímos também as informações relacionadas ao tipo de interação entre elas. O *STRING* divide as interações em seis classes: *binding*, *reaction*, *expression*, *activation*, *catalysis* e *ptmod* (*post-translational modifications*). *Binding* são interações físicas (diretas). Todas as outras são interações funcionais (indiretas) (SZKLARCZYK et al., 2015, p. 10). Importante destacar que as interações não são necessariamente excludentes. Por exemplo, em nosso trabalho, todas as interações indiretas também são diretas, ou seja, todas as interações funcionais também são classificadas como físicas.

Operacionalmente, extraímos tal informação do arquivo original do STRING do *STRING* (*353153.protein.actions.v10.txt*) que apresenta as proteínas que interagem entre si e o tipo de interação que existe entre elas. Devido à importância da incorporação desse dado em nossa rede, reunimos todos os tipos de interação (denominadas pelo *STRING* como *action*) de cada ligação (vide item 3.7.2). Vale destacar que cada par de proteínas pode apresentar mais de um tipo de interação física e funcional e assim sendo para o mesmo par protéico podem existir várias entradas no arquivo original que são representadas por linhas distintas. Após isso, construímos um programa em *awk* para separar os tipos de interação em colunas distintas. O código desse programa está disponível no Apêndice 5.

Ao final desses procedimentos obtivemos dois arquivos: um contendo todos os tipos de interação na mesma coluna (*tcruzi_proteins-string_interactions-actions_unified.tsv*) e outro separando os tipos de interação em colunas diferentes (*tcruzi_proteins-string_interactions-actions_unified.tsv*). Disponibilizamos o conteúdo do arquivo *tcruzi_proteins-string_interactions-actions_unified.tsv* no Apêndice 6.

4.7.3 Informação funcional: *EC numbers*

O *EC number* é uma classificação numérica utilizada para enzimas, baseada na reação química que elas catalisam. A nomenclatura foi originalmente publicada em 1992, pelo *Nomenclature Committee* da *International Union of Biochemistry and Molecular Biology* (*IUBMB*), com sucessivas atualizações nos anos subsequentes (“Enzyme Nomenclature”, 2016).

Um *EC number* é composto pelas letras “EC” seguidas por um código numérico no formato 1.2.3.4, que indicam diversos níveis de classificação da reação enzimática. O primeiro campo, indicando as classificações mais genéricas, vai de 1 a 6, representando cada uma das reações enzimáticas principais, a saber, segundo LEHNINGER; NELSON; COX (2011):

- **EC 1:** Oxidorredutases – transferência de elétrons (íons hidreto ou átomos de hidrogênio);
- **EC 2:** Transferases – reações de transferência de grupos;
- **EC 3:** Hidrolases – reações de hidrólise (transferência de grupos funcionais para a água);
- **EC 4:** Liases – adição de grupos de ligações duplas ou formação de ligações duplas por remoção de grupos;
- **EC 5:** Isomerases – transferência de grupos dentro de uma mesma molécula, produzindo formas isoméricas;
- **EC 6:** Ligases – formação de ligações C—C, C—S, C—O e C—N por reações de condensação acopladas à hidrólise de ATP ou cofatores similares.

Os campos subsequentes indicam detalhamentos na classificação da reação. Por exemplo, o *EC number* **1.1.3.4** (glicose oxidase) indica que as enzimas associadas a ele são oxidorredutases (**1.**), que atuam no grupo CH—OH dos doadores (**1.1.**), utilizando oxigênio como aceptor (**1.1.3.**) oxidando a glicose (**1.1.3.4**).

É importante destacar que os *EC numbers* não podem ser usados para identificar enzimas específicas, mas sim a reação catalizada pela enzima. Dessa forma, devemos olhar para os *EC numbers* como agrupamentos de enzimas com a mesma função catalítica.

Em nosso trabalho, utilizamos a associação dos *EC numbers* às proteínas como uma forma de caracterizar o seu papel no metabolismo do *T. cruzi*. Para isso, extraímos essa informação do arquivo de anotações de proteínas obtido a partir do *STRING* (*tcruzi_proteins-string_protein_annotations.tsv*). Esse arquivo contém uma

série de dados associados a cada uma das proteínas mapeadas na rede, tais como nomes alternativos, representação gráfica do domínio proteico e informações funcionais, como a via metabólica e o *EC number* eassociados às proteínas que porventura possuam tais informações. Construímos um programa na linguagem *awk* para isolar cada uma dessas informações e associá-las às proteínas correspondentes. Esse programa recebeu o nome de *split_string_proteins_details.awk* e encontra-se no Apêndice 7.

Como esse programa separa todas as informações referentes a cada uma das proteínas, sua execução gera uma quantidade massiva de dados. Por isso, a sua saída foi filtrada com o comando *grep*, uma vez que nos interessam somente os dados de *EC number*. O arquivo resultante (*tcruzi_proteins-string_proteins_annotations-ec_numbers.tsv*) encontra-se no Apêndice 8.

Dando continuidade ao processo de integração da informação funcional relaciona aos *EC Numbers*, precisávamos de uma forma de identificar cada um deles no momento em que fossem incorporados à rede. Para isso, decidimos usar cores: cada *EC Number* teria uma cor diferente, utilizando matizes para diferenciar os grandes grupos. Operacionalmente, essa tarefa foi concluída através da associação de cada um dos *EC Numbers* encontrados a um código, que é a representação da cor em notação *RGB* hexadecimal. Essa notação é reconhecida pelo *Cytoscape*, que a converte para a cor equivalente. As cores foram geradas automaticamente, utilizando o módulo de geração aleatória de cores *randomcolor* para a linguagem *Python*. Uma vez que foram encontrados *EC numbers* pertencentes a três dos seis grupos superiores (EC2, EC3 e EC6) decidimos utilizar um recurso desse módulo que é a possibilidade de definir matizes diferentes para os grupos (EC2: vermelho, EC3: verde, EC6: azul). As cores associadas a cada *EC number* encontram-se na TABELA 3.

TABELA 3: Associação entre os códigos de cores a cada um dos EC *numbers* válidos encontrados.

EC number	**Código da cor**
EC:2.7.11.1	#ff8e2b
EC:2.7.1.40	#efa5ba
EC:2.7.7.6	#cf092a
EC:3.1.2.6	#3fdb42
EC:3.1.3.48	#437e18
EC:6.3.2.19	#7aced6

Ainda em relação aos *EC numbers* encontrados, com o objetivo de melhor caracterizá-los, optou-se por buscar os detalhes relativos a cada um no banco de dados de enzimas (*ENZYME*) mantido pelo portal *ExPASy* (http://enzyme.expasy.org/). O *ENZYME* é um repositório que reúne tanto informações de nomenclatura quanto descrições sobre cada uma das enzimas relacionadas nas recomendações da *IUBMB* (BAIROCH, 2000). A busca pelos *EC numbers* poderia ter sido feita diretamente no site, entretanto, o *ENZYME* disponibiliza um arquivo que é basicamente um espelho do conteúdo encontrado no site, contendo as seguintes informações:

- *EC number*;
- nome recomendado;
- nome alternativo (caso exista);
- atividade catalítica;
- cofatores (caso existam);
- indicadores de entradas do *Swiss-Prot* que correspondam à enzima (caso existam).

Optamos pela transferência desse arquivo (*enzyme.dat*), já que isso nos oferecia uma maior flexibilidade nas buscas e no processamento da informação. Esse arquivo possui uma estrutura fixa pré-definida, em que os dois primeiros caracteres (termos descritores) de cada linha correspondem a um código que indica o tipo de informação ali contida, seguidos de três caracteres em branco e, por fim, a

informação propriamente dita, ou seja, os termos anotadores definidos anteriormente. Maiores detalhes sobre o formato desse arquivo, bem como a codificação utilizada, podem ser obtidos a partir do próprio repositório *FTP* do *ExPASy*, no endereço *ftp://ftp.expasy.org/databases/enzyme/enzuser.txt*.

Tendo como referência o arquivo com a relação dos *EC numbers* encontrados na rede, criado previamente (*tcruzi_proteins-string_proteins_annotations-unique_ec_numbers.tsv*), buscamos os detalhes de cada enzima no banco enzyme.dat, através de um *script Bash*, o que gerou um novo arquivo com essas informações (*tcruzi_proteins-string_proteins_annotations-ec_numbers_details.dat,* Apêndice 9).

Todos esses arquivos foram incorporados individualmente à rede montada pelo *Cytoscape*. A ordem de incorporação é importante e será descrita mais adiante neste trabalho.

4.7.4 Informação funcional: Anotação funcional das proteínas

Para finalizar a etapa de organização da informação funcional, extraímos a anotação funcional de cada uma das proteínas que fazem parte da rede. Mais uma vez utilizamos o programa *split_string_proteins_details.awk*, desta vez filtrando seu resultado com o comando *cut* e unificando as repetições da saída com o comando *uniq*. O conteúdo do arquivo resultante deste processamento contendo colunas que associando cada *STRING id* à sua respectiva anotação (*tcruzi_proteins-string_proteins_annotations-functional_annotation.tsv*) encontra-se no Apêndice 10.

Todos os procedimentos de obtenção e tratamento dos dados descritos até essa etapa foram resumidos em um fluxograma unificado, que está disponível no Apêndice 11.

4.8 Modelagem das redes no *Cytoscape*

Considerando que este trabalho foi direcionado para a construção de um protocolo de tratamento de dados para criação de RIPPs, exploraremos as análises que podem ser feitas com os recursos oferecidos pelo *Cytoscape*, com o objetivo de instrumentalizar os pesquisadores nos potenciais usos dessa ferramenta.

Entendemos ser possível aprofundar ainda mais a análise dos dados apresentados em termos da exploração das interrelações biológicas potenciais entre as proteínas integrantes da rede, mas isso ampliaria em muito a discussão desses resultados e fugiria do escopo metodológico deste trabalho.

4.8.1 A ferramenta *Cytoscape*

O *Cytoscape* é um programa criado com o objetivo de integrar, em um ambiente gráfico único, redes de interação de moléculas biológicas e dados de expressão de larga escala (SHANNON et al., 2003). Ele é desenvolvido na linguagem de programação *Java* e possui uma comunidade voltada ao desenvolvimento do mesmo bastante ativa, com novas versões lançadas regularmente. Além disso, está disponível para as plataformas *GNU/Linux*, *MacOS X* e *Microsoft Windows* 64-bit.

A tela padrão do *Cytoscape*, apresentada na FIGURA 15, é dividida em três grandes áreas. À esquerda (FIGURA 15-1) temos a região onde são listadas as redes, referenciada neste trabalho como “lista de redes”. O *Cytoscape* trabalha com uma lógica de “coleções de redes”, em que várias redes distintas podem se agrupadas e compartilhar dados e/ou estilos. À direita (FIGURA 15-2) temos a área de visualização da rede que está selecionada na coleção, referenciada neste trabalho como “área de visualização da rede”. Cada um dos nós da rede lá exibida pode ser livremente movimentado e pode-se também controlar o zoom da imagem. Existe um campo com a miniatura da rede no canto inferior direito, que indica a região que está sendo visualizada e qual o grau do zoom aplicado nessa visualização. Por fim, na parte inferior da tela (FIGURA 15-3) encontramos a tabela com todos os dados associados à rede selecionada, referenciada neste trabalho como “tabela de dados da rede”. Esses dados são compartilhados por todas as redes pertencentes à mesma coleção e podem ser acrescentados manualmente, inclusive com a opção de utilização de fórmulas, ou através da importação de tabelas de dados. Inclusive uma característica importante do *Cytoscape*, e que é um dos elementos tratados neste trabalho, é a possibilidade de se definir estilos de formatação dos elementos da rede com base nos dados disponíveis nessa tabela. Estes estilos podem configurar tanto características gerais, como uma cor ou um tamanho único para todos os elementos

da rede, quanto específicas, como variar a cor ou o tamanho de nós específicos, de acordo com valores presentes na tabela de dados da rede.

Além do recurso de estilos, o *Cytoscape* também apresenta uma série de outras funcionalidades de construção, edição e manipulação de redes biológicas. Por possuir uma arquitetura aberta, é possível estender os seus recursos com a utilização de *plugins* (denominados, no contexto do *Cytoscape*, como *apps*), que estão reunidos em uma página própria, dentro do próprio projeto (http://apps.cytoscape.org) e podem ser instalados a partir do próprio software, na opção **App Manager**, disponível a partir do menu **Apps**.

Como o cerne do presente trabalho é estruturar uma metodologia que seja robusta em relação ao tratamento dos dados de predições de redes biológicas, mas, ao mesmo tempo, simples de ser implementada, optamos por não utilizar nenhum *app*, uma vez que existe o risco de incompatibilidade entre versões dos *apps* e do *Cytoscape*, além da interrupção de seu desenvolvimento.

Dessa forma, para a tarefa de estruturação de nossas redes, utilizamos somente a implementação dos estilos e o *NetworkAnalyzer* que é uma ferramenta disponível no *Cytoscape.* Inicialmente distribuída como um *plugin*, atualmente esse recurso se encontra incorporado ao *Cytoscape* (ASSENOV et al., 2008). Essa ferramenta permite a análise de uma série de métricas de redes tanto direcionais quanto não-direcionais, adicionando os resultados à tabela de dados da própria rede. Uma vez efetuada a análise, é possível associar seus resultados aos estilos e alterar a representação visual da rede.

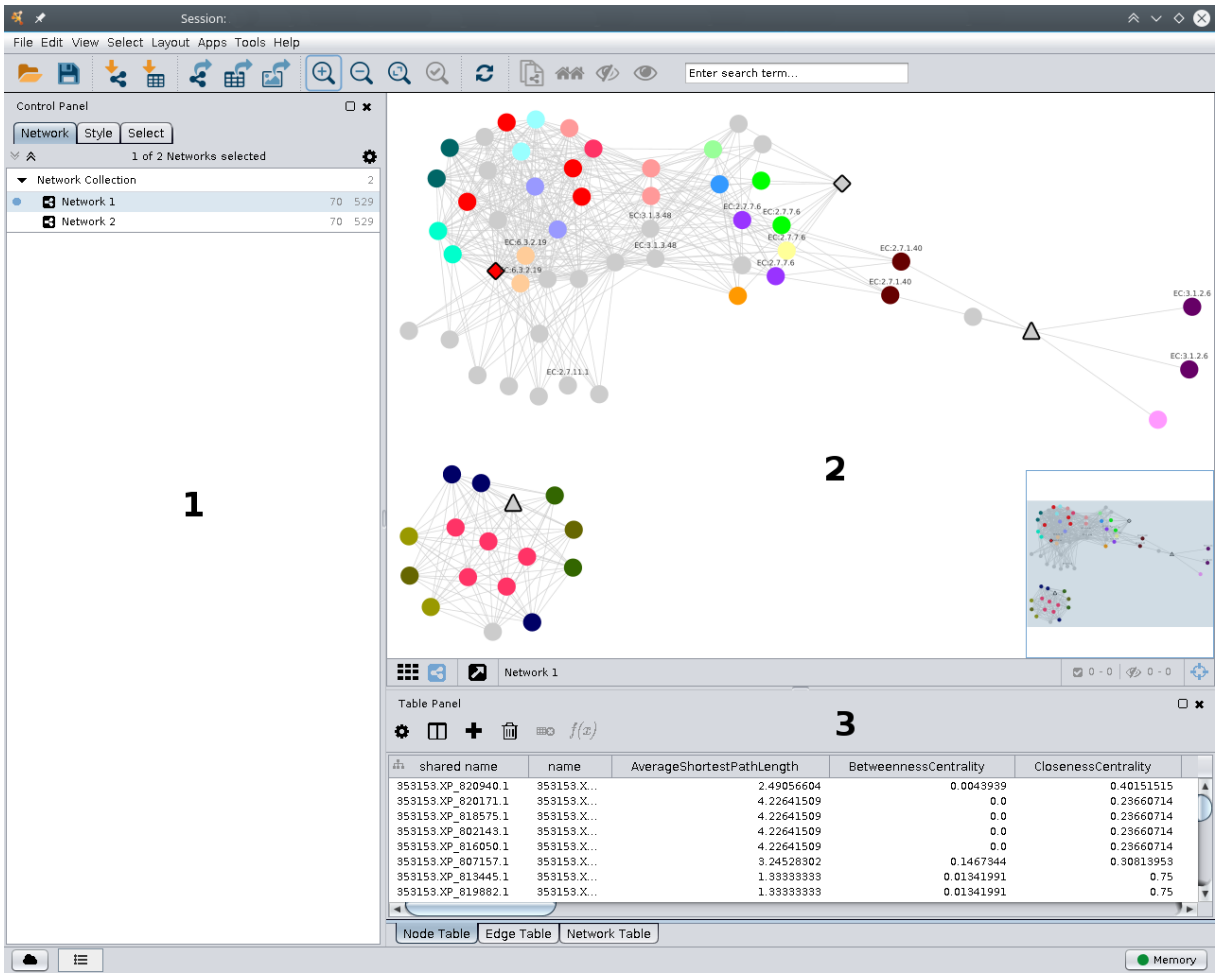


FIGURA 15: Tela padrão do *Cytoscape*, apresentando as três divisões principais: **1**-lista das coleções de redes, com as redes associadas a cada uma delas; **2**-área de visualização do desenho da rede selecionada na coleção; **3**-tabela com todos os dados associados à rede selecionada na coleção.

4.8.2 Entrada e processamento dos dados

Nesta etapa analítica do trabalho consideramos como dados de entrada os resultados das predições e análises funcionais obtidos através de diferentes metodologias e formatados de maneira tabular em arquivos com nomenclaturas específicas que refletem a sua origem e o seu processamento.

O estabelecimento das correspondências destas análises com os elementos da rede representa o nosso objetivo neste momento. A definição das colunas que servirão como consulta nas tabelas de dados de entrada (atributo *Key* na tela de importação de tabelas) e de indexadora na rede já existente (parâmetro **Key Column for Network**, na tela de importação de tabelas) representa um passo crítico no

estabelecimento das associações. Isso porque o *Cytoscape* faz, automaticamente, a correspondência entre os dados de entrada e os dados da coluna utilizada como indexador. Caso seja definida uma coluna incorreta nos dados já existentes ou na tabela a ser incluída, a correspondência não ocorrerá.

Além disso, considerando que colunas de uma tabela importada podem servir como indexadoras para tabelas subsequentes, a ordem de entrada das tabelas de dados também é importante. Como exemplo específico podemos mencionar a associação estabelecida entre os resultados obtidos das análises de vias metabólicas através do *EC number* e da sua respectiva atribuição de cor (arquivos *tcruzi_proteins-string_proteins_annotations-ec_numbers.tsv* e *tcruzi_proteins-string_proteins_annotations-unique_ec_numbers-colors.tsv)* onde a coluna *ec_number* do primeiro conjunto de dados serviu como indexador para a coluna *ec_number_color* do segundo conjunto.

Ao fazer a correspondência entre as colunas de consulta e de índice o *Cytoscape* também elimina automaticamente toda não correspondência. Devido a isso, é normal encontrarmos um número menor de linhas em algumas colunas, ao final da importação.

4.8.3 Análise da rede gerada

Com o objetivo de identificar características gerais da rede e ampliar a disponibilidade de dados utilizáveis na melhor caracterização dos elementos individuais das nossas redes, utilizamos o *NetworkAnalyzer* para efetuar uma análise da rede *STRING - DB/Experiments - Combined 400*, que contém todas as interações de proteínas de *T. cruzi* disponíveis no *STRING*. Na execução do *NetworkAnalyzer* a rede foi tratada como não-direcional, uma vez que redes direcionais só podem ser analisadas por essa ferramenta quando todas as suas arestas possuem informação de direcionamento. Após executarmos o *NetworkAnalyzer* ele nos apresenta uma tela com as características gerais da análise da rede (FIGURA 16).

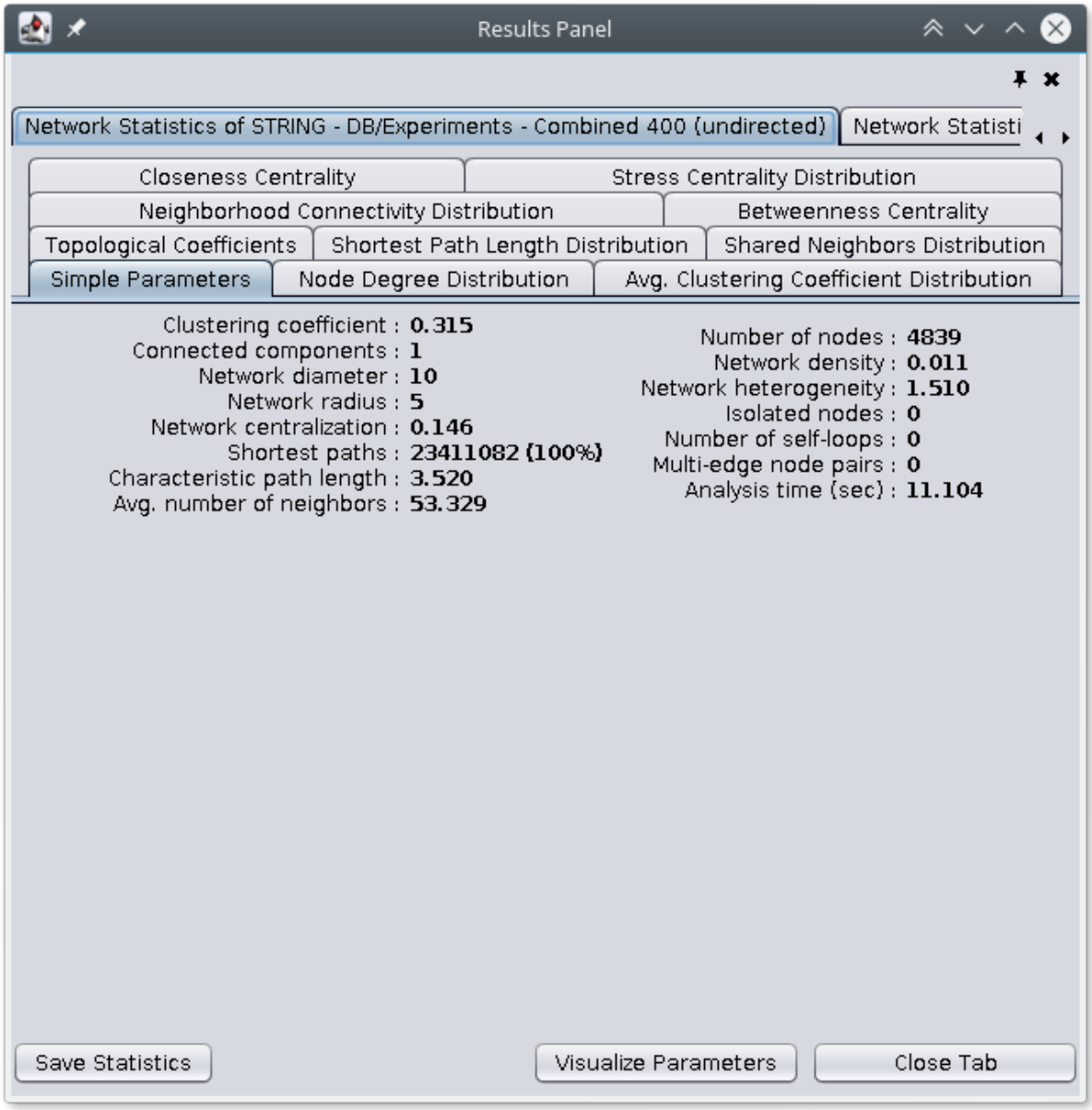


FIGURA 16: Tela de resultados da ferramenta *NetworkAnalyzer*, do *Cytoscape*, após analisar a rede completa de *T. cruzi*.

Nos interessa, no escopo deste trabalho, a análise encontrada na aba *Node Degree Distribution*, por ser um indicativo do tipo de rede. Conforme descrito na seção 1.2, redes biológicas tendem a ser do tipo livres de escala, o que pode ser avaliado quando se analisa o gráfico de distribuição dos nós da rede. Uma vez que redes livres de escala seguem uma distribuição de lei de potência, na forma $y = ax^b$, aplicamos um *fitting* a esse gráfico e obtivemos o resultado apresentado na FIGURA

17. A correlação encontrada foi de 0,619, que não é expressiva, mas, ainda assim, suficiente para corroborar que a distribuição dos nós segue uma lei de potência.

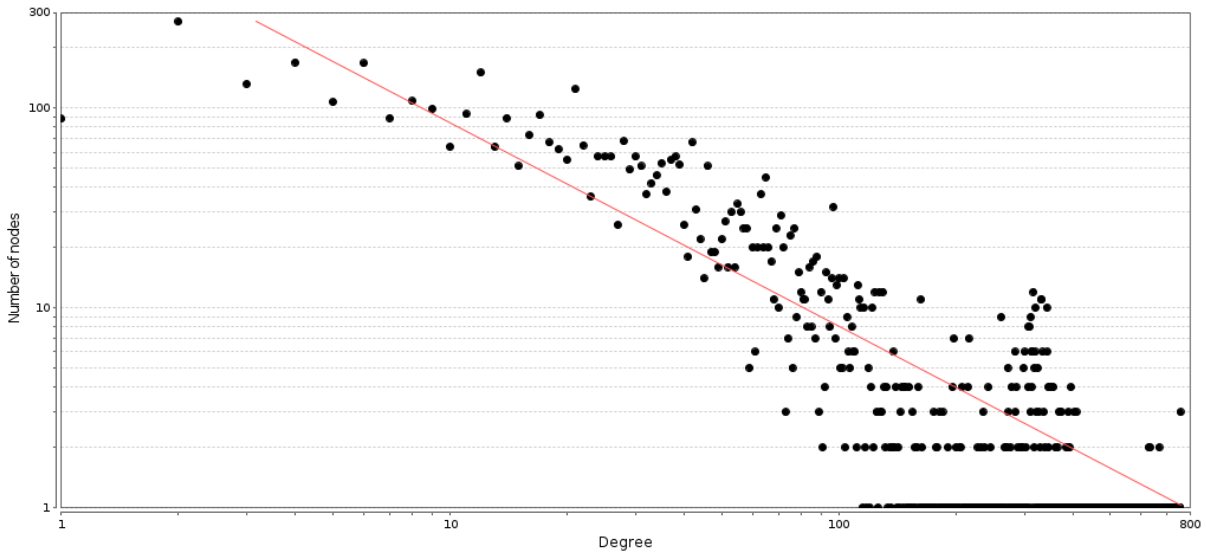


FIGURA 17: Gráfico da distribuição de graus dos nós (proteínas) da rede completa de *T. cruzi*, calculado pelo *NetworkAnalyzer*. A linha representa o resultado do *fitting* desta distribuição através da lei de potência, de forma $y = ax^b$.

4.8.4 Exploração dos estilos de formatação da rede

Utilizamos os dados que foram incorporados à nossa rede para apresentar o potencial de uso dos estilos como forma de destacar características biológicas significativas.

Através dos estilos é possível controlar diversos detalhes na exibição dos elementos da rede, tais como cor, transparência, tamanho, visibilidade e estilo tanto dos ícones representando os nós quando das linhas representando as arestas, além da posição, cor, transparência e visibilidade de todas as etiquetas vinculadas aos nós e arestas. Conforme descrito anteriormente, esses parâmetros podem ser definidos de maneira única para os elementos ou podem estar atrelados a alguma coluna da *tabela de dados da rede*. Neste caso, os parâmetros a serem utilizados podem vir diretamente da tabela, como as cores associadas aos *EC numbers*, podem ser calculados em uma escala contínua (caso os dados sejam numéricos) ou podem ser inseridos manualmente, associando-se um valor a cada elemento único da coluna

da tabela. Em nosso trabalho, utilizamos os três modelos, conforme descrito mais adiante.

Como critério de comparação do quanto os estilos podem interferir na visualização da rede, apresentamos, na FIGURA 18, a rede *STRING* - *website*, sem nenhum tipo de formatação.

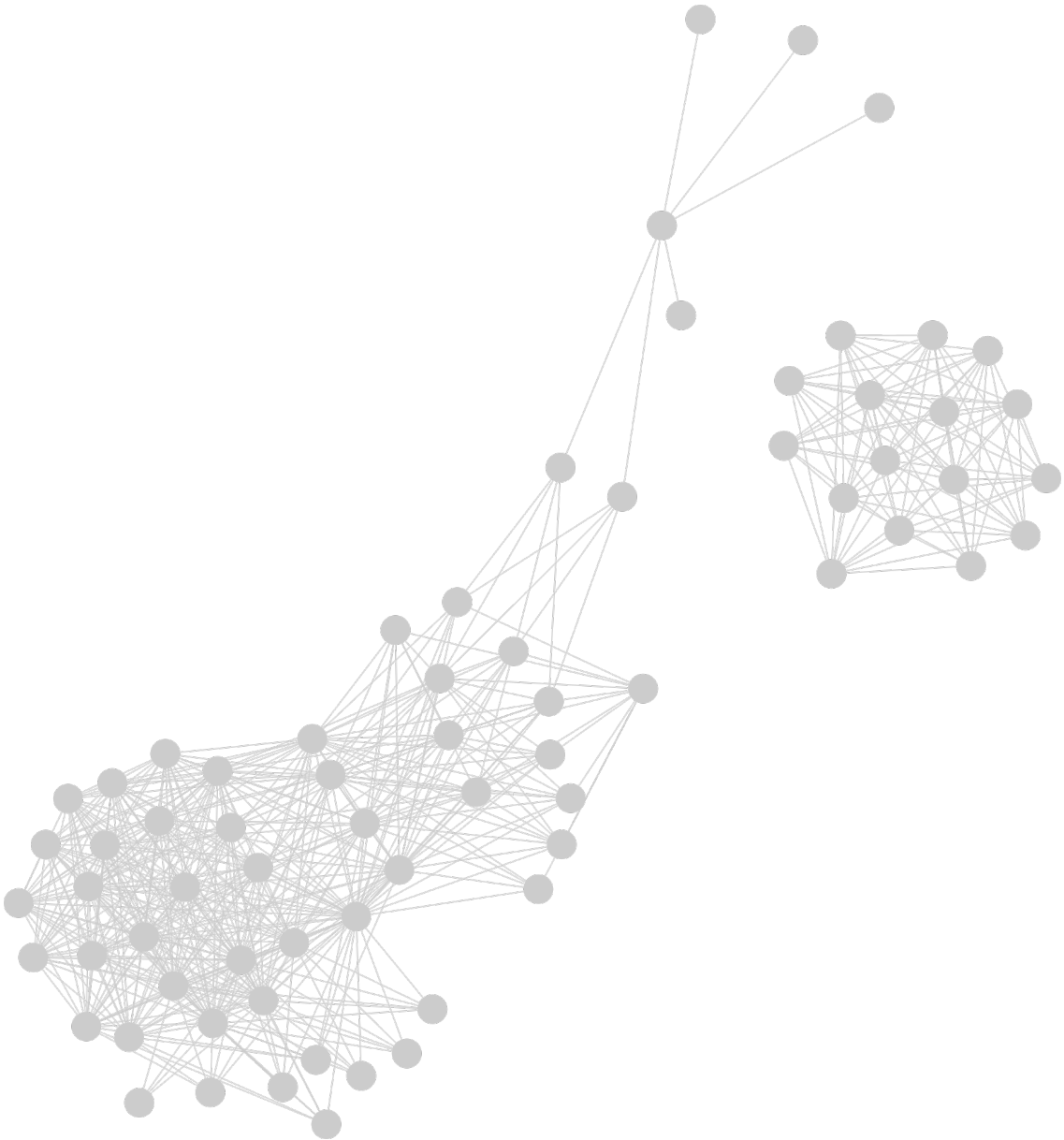


FIGURA 18: Rede de interações de proteínas construída no *Cytoscape* a partir dos dados gerados pelo site do *STRING*, sem nenhum tipo de formatação adicional.

Criamos então um estilo para esta rede, nomeando-o como *t.cruzi-ec-expression-interactions* e definindo os seguintes parâmetros:

- alteração dos ícones dos nós que representavam proteínas com informação de regulação gênica, para triângulo (▲), caso a regulação seja *rna pos/ptn pos*, ou losango (◆), caso seja *rna pos/ptn neg* (que foram os dois únicos tipos de regulação que surgiram em nossa rede após a correspondência de dados, conforme descrito na seção 4.8.2);
- alteração do valor da largura das bordas (para *5*) de todos os ícones desses mesmos nós, representantes de proteínas que apresentavam informação de regulação gênica;
- alteração das cores dos nós que representavam as proteínas que possuíam informação de *EC number*, utilizando os valores da coluna *ec_number_color* como referência;
- alteração da exibição de rótulos somente dos *EC numbers*, utilizando os valores da coluna *ec_numbers* como referência;
- alteração das cores das arestas de acordo com o tipo de ligação que elas representam. Esses valores foram fornecidos manualmente, para que fossem escolhidas cores específicas para cada um dos valores, utilizando a coluna *action* como referência.

Alterados esses parâmetros, a visualização da rede assumiu um aspecto muito mais rico em termos de informações biologicamente relevantes, conforme pode ser visto na FIGURA 19. O *STRING* oferece a opção de exportar as configurações do estilo criado para um arquivo XML, que pode ser posteriormente recuperado e aplicado em uma outra rede. Exportamos então esse estilo para o arquivo *t.cruzi-ec_number-actions.xml*, cujo conteúdo está disponível no Apêndice 12.

A adição de elementos de formatação já destaca visualmente algumas propriedades em nossa rede facilitando a curadoria e interpretação biológica dos resultados. Inicialmente percebemos que três das quatro proteínas do conjunto original de dados (selecionadas e organizadas conforme descrito na seção 3.3), fazem parte de um único componente conectado e estão separadas umas das outras por, no máximo, dois vizinhos. E apesar dessas três proteínas não apresentarem associação com nenhum *EC number*, elas estão ligadas diretamente a 13 outras proteínas que apresentam tal associação, conforme detalhado na TABELA 4.

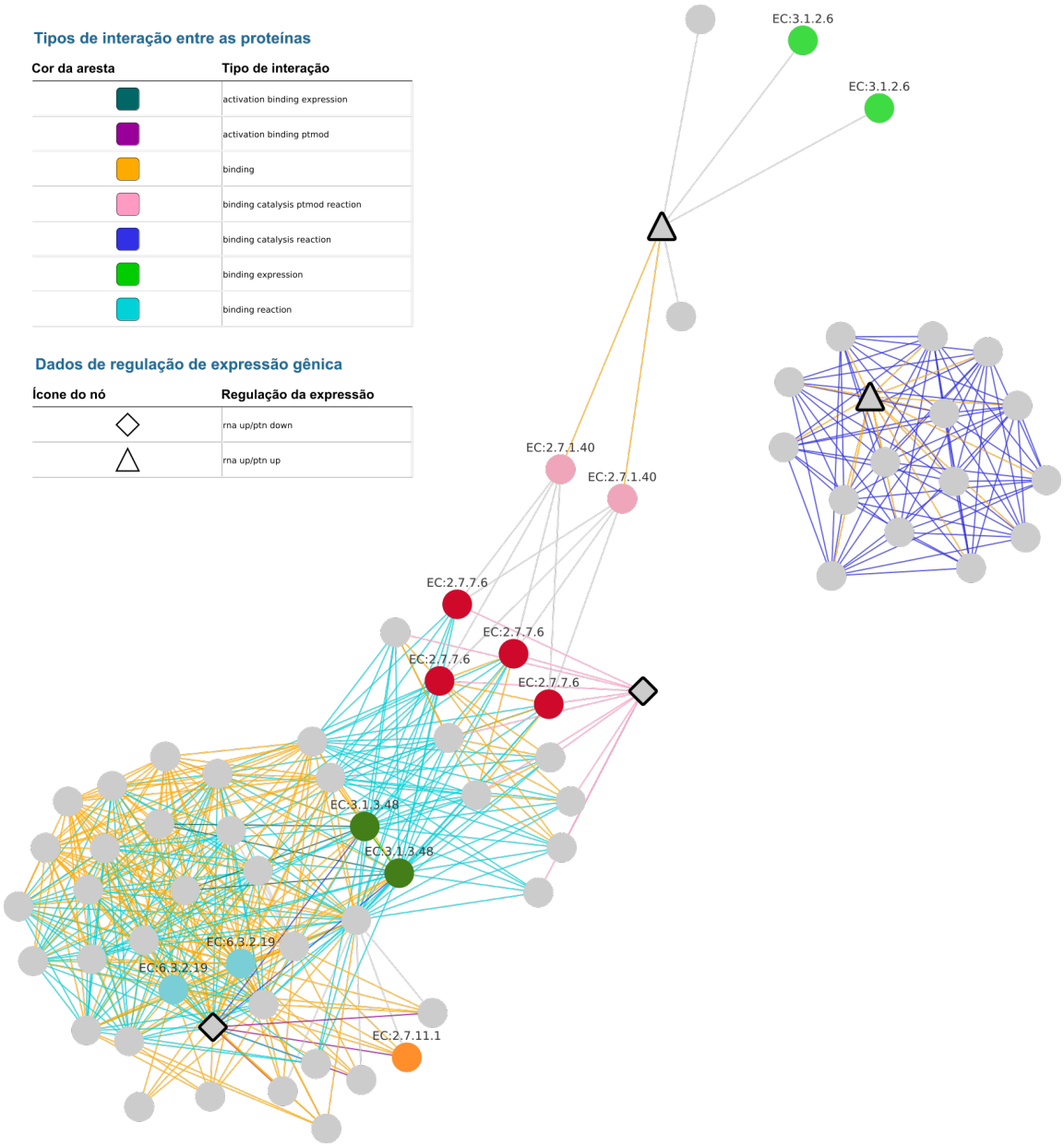


FIGURA 19: Rede de interações de proteínas construída no *Cytoscape* a partir dos dados gerados pelo site do *STRING*, visualmente reformatada segundo informações dos dados de enriquecimento funcional alimentados na rede. As cores dos nós correspondem aos *EC numbers*. A legenda foi gerada através do recurso de criação de legendas do próprio *Cytoscape*, editada digitalmente e adicionada posteriormente à imagem.

Cabe destacar que essas proteínas associadas a *EC numbers* não faziam parte do conjunto original de dados e só foram descobertas por apresentarem algum tipo de interação com as proteínas originais.

TABELA 4: Ligações entre as proteínas utilizadas inicialmente para a geração da rede no *STRING* e seus vizinhos que apresentam associação com algum *EC number*. O *EC number* 6.3.2.19 teve sua entrada transferida para três outros *EC numbers*: 2.3.2.23, 2.3.2.27 e 6.2.1.45. Estamos trabalhando com a entrada 6.2.1.45 neste exemplo.

***Gene id*/anotação funcional/regulação**	**Primeiros vizinhos**		
	STRING id	***EC number***	**Anotação funcional**
TCSYLVIO_000129 actin interacting protein-like rna pos/ptn pos	353153.XP_811202.1	2.7.1.40	pyruvate kinase 2
	353153.XP_820627.1	2.7.1.40	pyruvate kinase 2
	353153.XP_818575.1	3.1.2.6	hydroxyacylglutathione hydrolase
	353153.XP_820171.1	3.1.2.6	hydroxyacylglutathione hydrolase
TCSYLVIO_006124 hypothetical protein rna pos/ptn neg	353153.XP_808084.1	2.7.7.6	DNA-directed RNA polymerase I largest subunit
	353153.XP_809667.1	2.7.7.6	DNA-directed RNA polymerase I largest subunit
	353153.XP_812569.1	2.7.7.6	RNA polymerase IIA largest subunit
	353153.XP_819691.1	2.7.7.6	DNA-directed RNA polymerase III largest subunit
TCSYLVIO_001850 RNA-binding protein rna pos/ptn neg	353153.XP_818985.1	2.7.11.1	serine/arginine-rich protein specific kinase SRPK
	353153.XP_804043.1	3.1.3.48	dual specificity protein phosphatase
	353153.XP_821800.1	3.1.3.48	dual specificity protein phosphatase
	353153.XP_806008.1	6.2.1.45 (6.3.2.19)*	hypothetical protein
	353153.XP_816312.1	6.2.1.45 (6.3.2.19)*	hypothetical protein

A seguir detalhamos as características funcionais das proteínas ligadas aos *EC numbers* encontrados e apresentados na FIGURA 19 (elementos coloridos) e TABELA 4:

- **2.7.1.40 – Piruvato quinase:** A piruvato quinase é a última enzima da via glicolítica. É regulada alostericamente pela frutose 2,6-bisfosfato e catalisa a transferência do fosfoenolpiruvato para o difosfato de adenosina (ADT) para gerar ATP e piruvato. Ocorre tanto no inseto vetor da doença de Chagas quanto no hospedeiro mamífero (GRÜNING et al., 2011). Em nosso trabalho,

as duas proteínas associadas a esse *EC number* estão anotadas como *pyruvate kinase 2*.

- **2.7.7.6 – RNA polimerase:** Nos tripanosomatídeos, a transcrição é realizada por, pelo menos, três RNAs polimerases, onde as RNA polimerase I e III são responsáveis, respectivamente, pela transcrição de RNAs ribossomais (rRNA) e RNAs transportadores (tRNA). Por sua vez, a RNA polimerase II é responsável pela transcrição de genes codificantes para proteínas (KISSINGER, 2006; PAPADOPOULOU et al., 2003; STUART et al., 2008). Neste trabalho, associadas a esse *EC number*, temos duas proteínas anotadas como *DNA-directed RNA polymerase I largest subunit*, uma *DNA-directed RNA polymerase III largest subunit* e uma *RNA polymerase IIA largest subunit*.

- **2.7.11.1 – Serina/treonina quinase:** Essas enzimas catalisam a fosforilação de grupos OH dos aminoácidos serina ou treonina. Elas também atuam na homeostase celular e em vias de sinalização com a habilidade de fosforilar fatores de transcrição e reguladores do ciclo celular (CAPRA et al., 2006; PARSONS et al., 2005). A única proteína associada a esse *EC number*, está anotada como *serine/arginine-rich protein specific kinase SRPK*.

- **3.1.2.6 – Hidroxiacilglutationa hidrolase:** Um dos produtos dessa reação de hidrólise é a glutationa, que está relacionada à proteção nos processos de defesa antioxidante do parasito (ERCOLANI et al., 2016). Associadas a esse *EC number*, encontramos duas proteínas anotadas como *hydroxyacylglutathione hydrolase*.

- **3.1.3.48 – Tirosina fosfatase:** Está diretamente associada a cascatas de sinalização transmembrana. Elas atuam nos processos de desfosforilação, removendo grupos fosfato de vários aminoácidos como, por exemplo, tirosina, serina e treonina. Além de regularem a sinalização intracelular ativada pelas proteínas quinases ativadas por mitógenos (MAPK, do inglês, *Mitogen Activated Protein Kinases***)** e por fatores de estresse também estão relacionadas a modulação da progressão de algumas etapas do ciclo celular

atuando sobre as ciclinas dependentes de quinases (BRENCHLEY et al., 2007; GARG; GOYAL, 2015; WHITMARSH; DAVIS, 2000). Duas proteínas estão associadas a esse *EC number*, neste trabalho, ambas anotadas como *dual specificity protein phosphatase*.

- **6.2.1.45 (6.3.2.19) – Enzima ativadora da ubiquitina:** Essa relacionada ao processo de degradação de proteínas desnaturadas pelo complexo proteassoma além de atuarem no reparo de DNA, progressão do ciclo celular, endocitose e apoptose (DE DIEGO et al., 2001; KIRCHHOFF et al., 1988).

Um dos recursos do *Cytoscape* é a possibilidade de selecionarmos nós específicos da rede e gerarmos um sub-grafo a partir dessa seleção. Isso é particularmente útil quanto queremos analisar com mais detalhes regiões específicas de interesse biológico da rede, eliminando outras interações que poderiam confundir a visualização dos elementos de interesse. Selecionamos então três conjuntos de proteínas associadas aos *EC numbers*: 2.7.1.40, 2.7.7.6 e 3.1.3.48 e montamos uma nova rede, que pode ser vista na FIGURA 20.

Esses *EC numbers* foram escolhidos por estarem ligados aos mecanismos de variabilidade genética do *T. cruzi*: RNA polimerase, quinase e fosfatase. Destaca-se nessa sub-rede o fato de que tanto as proteínas quinases quanto as fosfatases estarem ligadas a todas as RNA polimerases. Inclusive as fosfatases estabelecem, além de interação física, interação de reação. A informação do tipo de interação que as quinases estabelecem com as RNA polimerases não está disponível no *STRING*.

Aqui cabe ressaltar que a variabilidade genética dos tripanossomatídeos está relacionada à adaptação do parasita ao seu hospedeiro vertebrado e invertebrado e pode contribuir para a diversidade genética do parasita e sua taxa de infectividade.

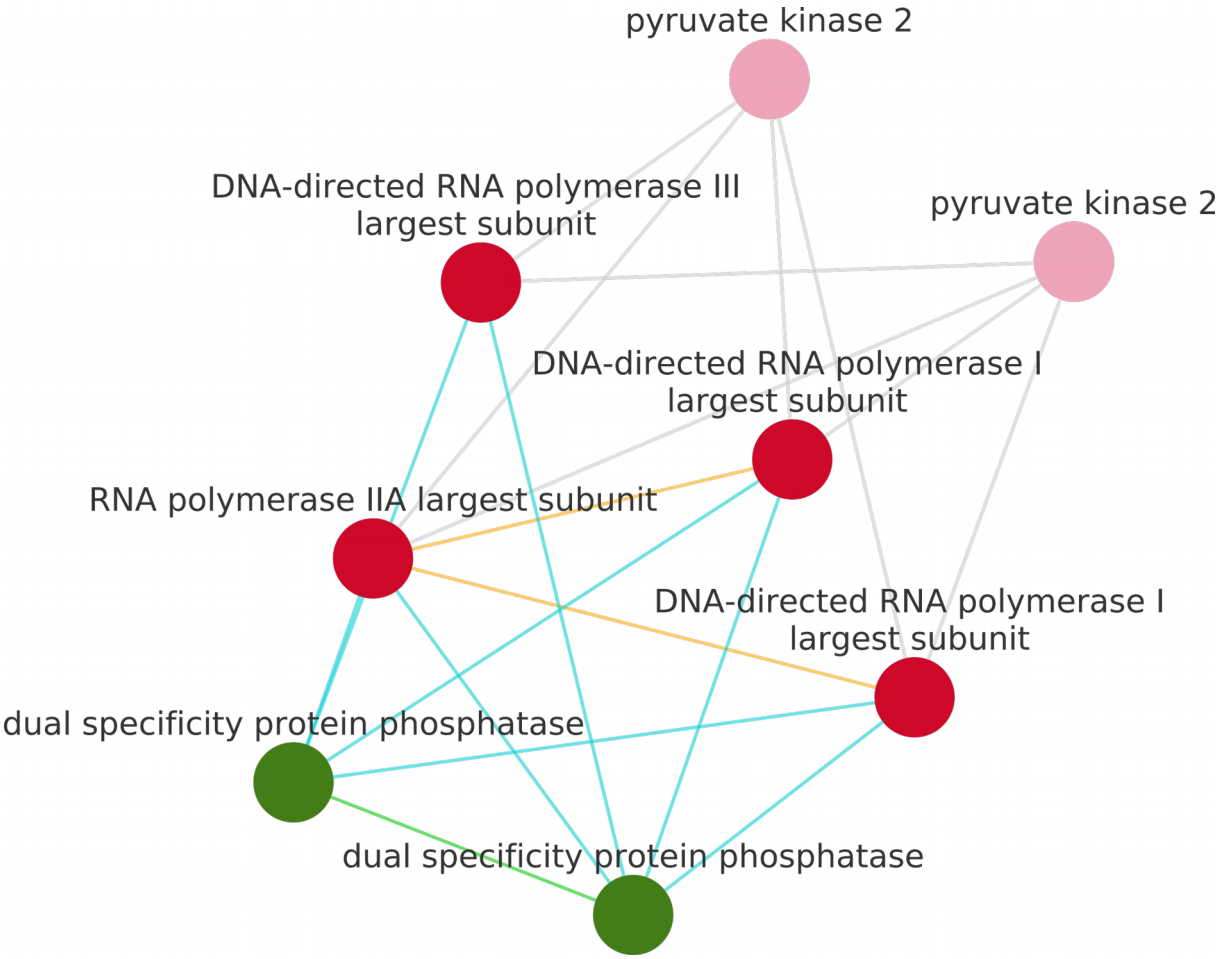


FIGURA 20: Sub-rede gerada a partir da seleção das proteínas associadas aos *EC numbers* 2.7.1.40, 2.7.7.6 e 3.1.3.48. As cores dos elementos são as mesmas empregadas na FIGURA 19.

Mais dois exemplos de destaques de sub-redes podem ser vistos nas FIGURAS 21 e 22. Em ambas selecionamos os primeiros vizinhos de duas das proteínas utilizadas na construção original da rede. Em relação à primeira, cujo *Gene id* é *TCSYLVIO_006124*, e está anotada como proteína hipotética, podemos perceber que todas as suas interações são com proteínas anotadas como *RNA polimerases*. Vale ressaltar nesta representação que as proteínas anotadas como RNA polimerase I e II apresentam diversas interconexões, enquanto as descritas como RNA polimerase III interagem diretamente apenas com a proteína hipotética. Além disso, essa proteína possui regulação da expressão positiva para RNA e negativa para proteína e quando analisamos as suas interações conhecidas, todas são do tipo físico e modificação pós-traducional. Interações ocorridas após a tradução podem

indicar uma ativação pós-traducional, o que explicaria a diferença de regulação entre o RNA e a proteína.

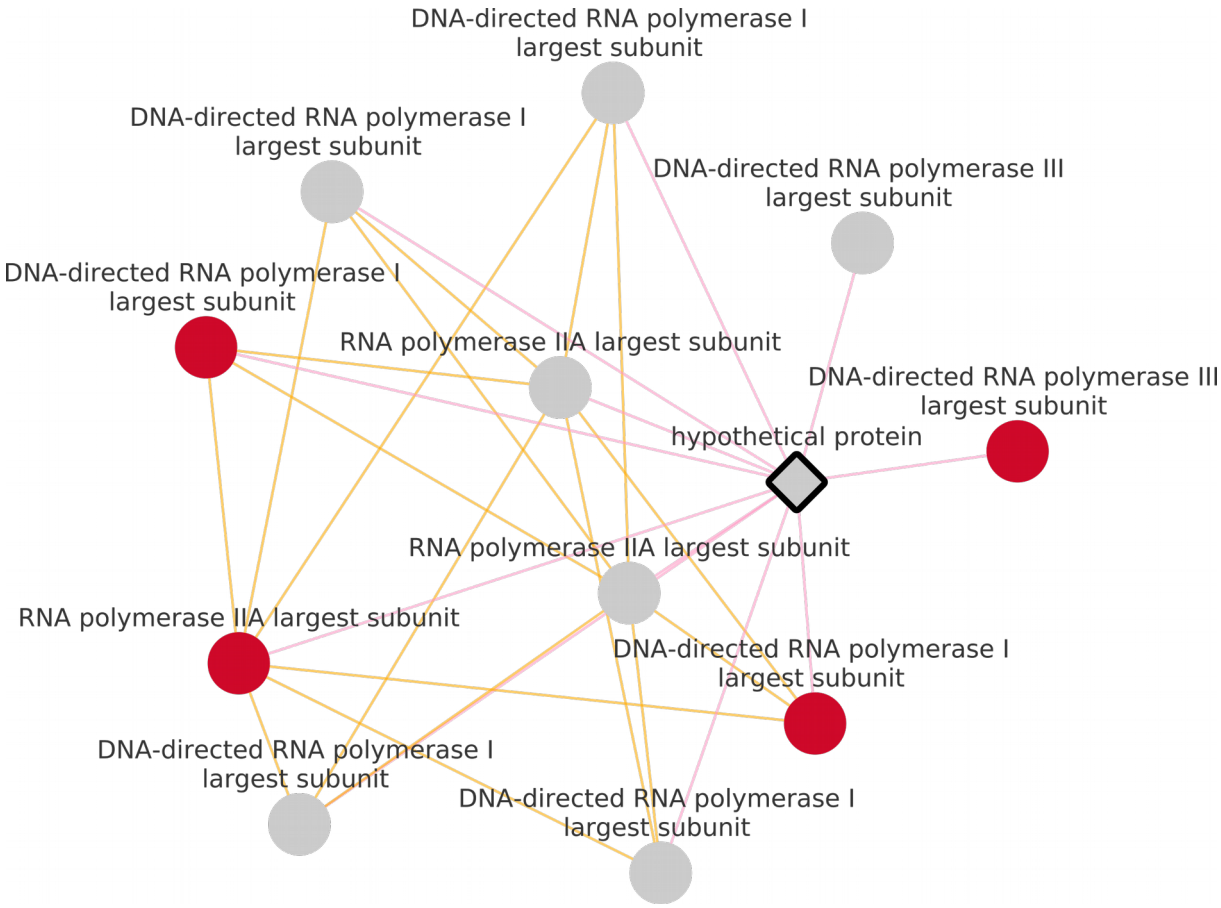


FIGURA 21: Interação estabelecidas pela proteína de *gene id TCSYLVIO_006124*, uma das utilizadas na geração da rede original. Todas as suas interações são com proteínas anotadas como RNA polimerase e ela serve como mediadora nas interações entre aquelas identificadas como RNA polimerase III e as identificadas como RNA polimerase I e II.

Já a FIGURA 22 foi obtida isolando-se a proteína que não aparece no componente principal da rede, cujo *gene id* é *TCSYLVIO_006143* e também anotada como proteína hipotética. Quando analisamos a sua vizinhança, percebemos que todas as proteínas estão associadas ou à montagem de clatrinas ou são adaptinas, que, junto com as clatrinas atuam na formação de vesículas de transporte. As clatrinas estão bem caracterizadas no *T. cruzi* (KALB et al., 2014, 2016) e existem indícios de que as vesículas mediadas por elas possam estar associadas à infectividade desse organismo (GARCIA-SILVA et al., 2014). Tal informação é consistente com o fato de

que a proteína hipotética ligada a essa vizinhança apresenta regulação positiva tanto para RNA quando para proteína.

Avançando na utilização dos estilos, partimos para a integração dos dados biológicos com os resultantes da análise realizada pelo *NetworkAnalyzer*. Consideramos aqui apenas três métricas, dentre as disponíveis: o grau (degree), a centralidade de intermediação (*betweenness centrality*) e o coeficiente de agrupamento (*clustering coefficient*).

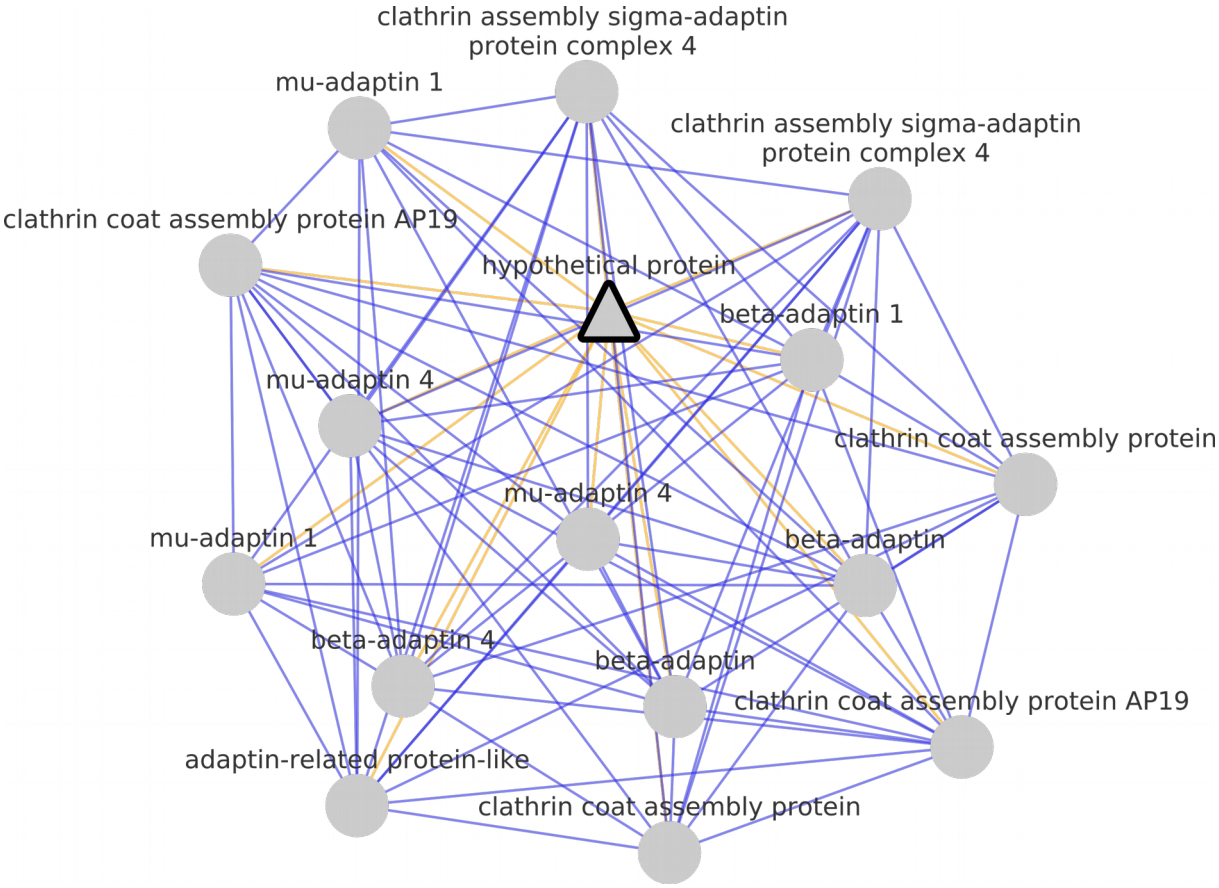


FIGURA 22: Interação estabelecidas pela proteína de *gene id TCSYLVIO_006143*, uma das utilizadas na geração da rede original. Todas as suas interações são com proteínas adaptinas ou então associadas à montagem de clatrinas, ambos elementos atuantes na formação de vesículas de transporte.

Os grafos presentes na FIGURAS 23, 24 e 25 são reformatações daquele apresentado na FIGURA 19, com a diferença que agora estamos destacando as informações de grau, centralidade de intermediação e coeficiente de agrupamento

para cada um dos nós. Quanto maior o ícone do nó, maior o valor correspondente a cada uma dessas métricas.

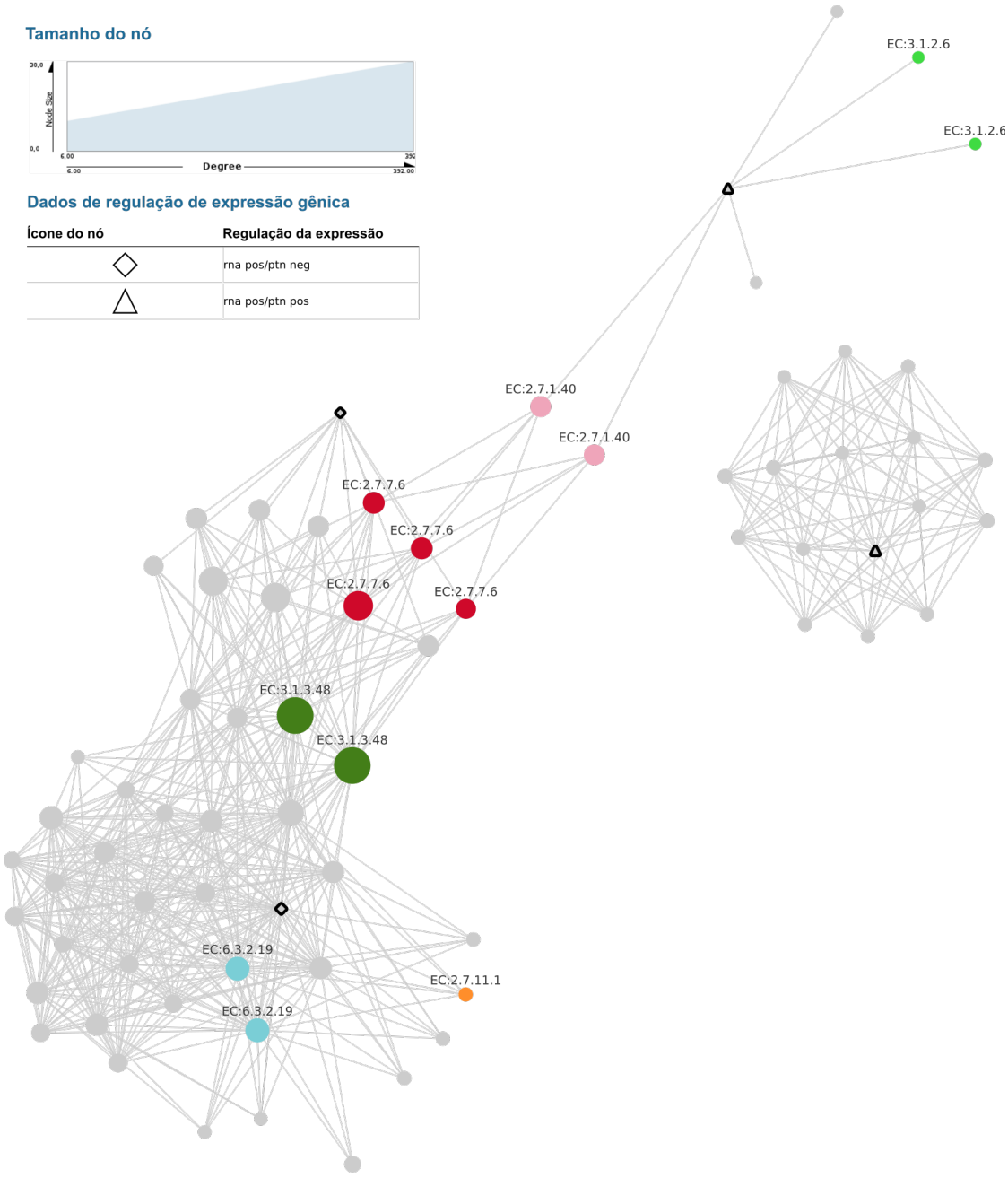


FIGURA 23: Rede de interações de proteínas construída no *Cytoscape* a partir dos dados gerados pelo site do *STRING*, visualmente reformatada segundo informações dos dados de enriquecimento funcional alimentados na rede, com destaque aqui para associação entre o *grau* do nó e o tamanho do seu ícone. As cores dos nós correspondem aos *EC numbers*. A legenda foi gerada através do recurso de criação de legendas do próprio *Cytoscape*, editada digitalmente e adicionada posteriormente à imagem.

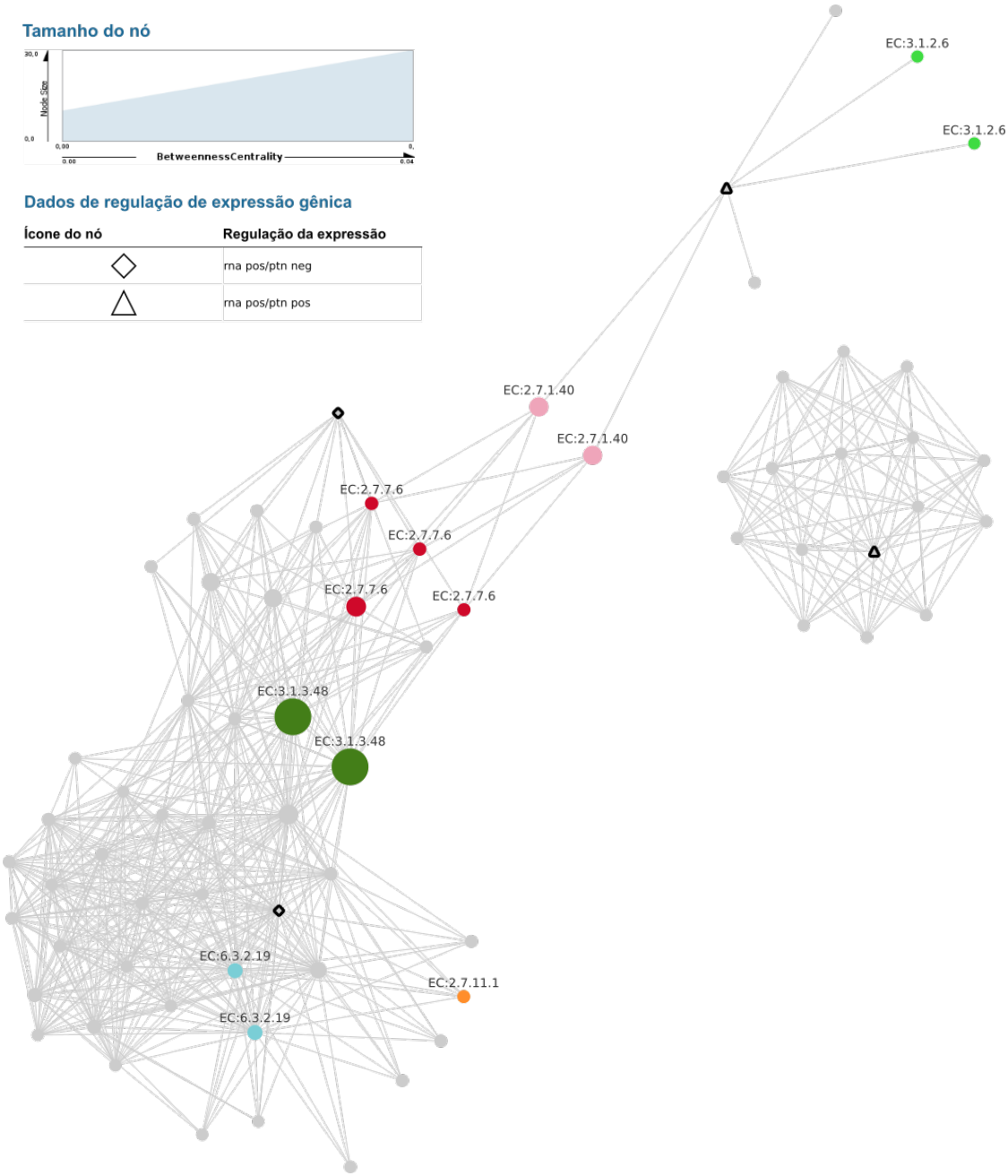


FIGURA 24: Rede de interações de proteínas construída no *Cytoscape* a partir dos dados gerados pelo site do *STRING*, visualmente reformatada segundo informações dos dados de enriquecimento funcional alimentados na rede, com destaque aqui para associação entre a *centralidade de intermediação* do nó e o tamanho do seu ícone. As cores dos nós correspondem aos *EC numbers*. A legenda foi gerada através do recurso de criação de legendas do próprio *Cytoscape*, editada digitalmente e adicionada posteriormente à imagem.

As métricas de grau e intermediação estão entre os chamados “indicadores de centralidade”, que estão entre os conceitos mais estudados nas análises de redes

(BORGATTI, 2005). Em relação essas métricas existem evidências de que, apesar de estarem conceitualmente relacionadas, suas medidas podem ser distintas (VALENTE et al., 2008). Isso fica claro ao compararmos as FIGURAS 23 e 24, em que as proteínas associadas aos *EC numbers* 2.7.7.6 e 6.3.2.19 apresentam diferenças visíveis, o mesmo não ocorrendo com as associadas ao 3.1.3.48.

Existem também evidências de que, do ponto de vista biológico, inclusive por representar uma métrica que considera a rede como um todo, a medida da intermediação tenha um papel mais relevante que a de grau, podendo inclusive representar indícios de como a rede de interações evoluiu com o tempo (JOY et al., 2005).

Um indício do papel da medida da intermediação no entendimento da importância da proteína para o organismo pode ser verificado ao se analisar esse valor para as duas proteínas associadas ao *EC number* 3.1.3.48, anotadas como *dual specificity protein phosphatase*. Além do destaque visual nas figuras, ao consultarmos o valor dessa métrica na tabela de dados da rede, verificamos que é terceiro maior valor entre todas as medidas de intermediação da rede inteira (que possui 4.839 proteínas). Também é o oitavo maior valor de grau. Ou seja, essa proteína é a oitava com maior número de conexões na rede e também a terceira em número de intermediações de menores caminhos na rede, resumindo, ela ocupa uma posição privilegiada na topologia da rede e pode, inclusive, ser utilizada na conexão de processos biológicos em agrupamentos funcionais distintos na rede, fato que condiz com a importância do papel das fosfatases na biologia do *T. cruzi*.

Por fim, utilizamos a FIGURA 25 para destacar a medida do coeficiente de agrupamento. Apesar de, isoladamente, essa métrica não ser determinante quanto à caracterização de agrupamentos na rede, ela nos dá pistas de potenciais grupos. No caso específico de nossa rede, ela reforça a análise efetuada anteriormente, a partir da FIGURA 22, indicando grupos de proteínas associadas à produção de vesículas de transporte.

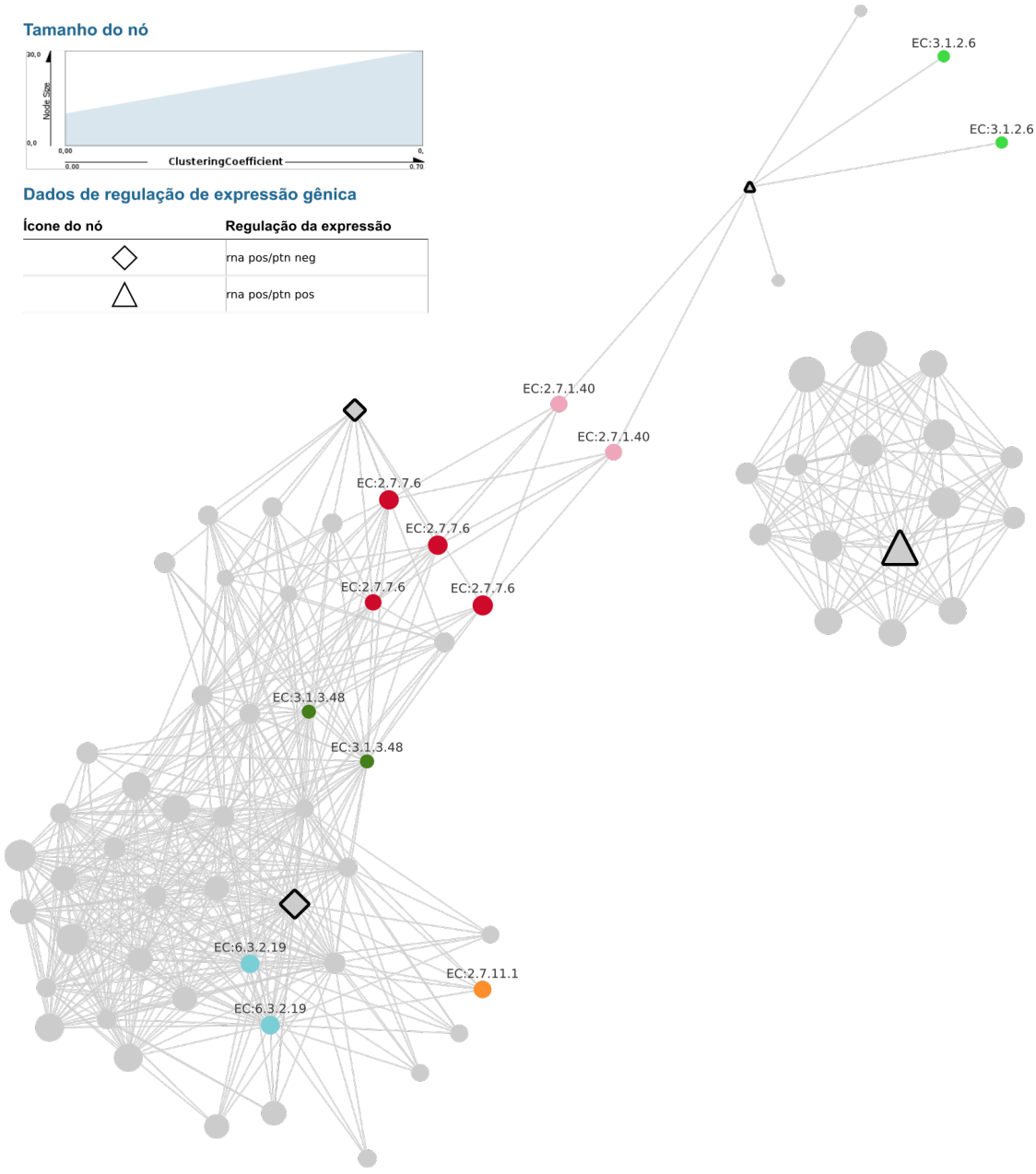


FIGURA 25: Rede de interações de proteínas construída no *Cytoscape* a partir dos dados gerados pelo site do *STRING*, visualmente reformatada segundo informações dos dados de enriquecimento funcional alimentados na rede, com destaque aqui para associação entre o *coeficiente de agrupamento* do nó e o tamanho do seu ícone. As cores dos nós correspondem aos *EC numbers*. A legenda foi gerada através do recurso de criação de legendas do próprio *Cytoscape*, editada digitalmente e adicionada posteriormente à imagem.

5 CONCLUSÃO

Com o propósito de desenvolver e implementar uma metodologia de integração de dados ômicos em uma rede de interação proteína-proteína, ao longo deste trabalho desenvolvemos uma metodologia computacional que associou a utilização de dados do banco *STRING* v.10, a um outro conjunto de dados, gerados pelo grupo Informática de Biossistemas e Genômica, oriundos de projetos que envolveram a análise e aquisição de informações provenientes da análise massiva de RNA e proteínas.

Ao final do processo, obtivemos um conjunto de procedimentos e ferramentas utilizáveis no tratamento e integração dos dados disponíveis em redes biológicas. Nos atemos ao desenvolvimento e implementação de uma metodologia que fosse plenamente reprodutível, com detalhamentos relacionados a sua execução e inclusão de todos os códigos utilizados. Ainda pensando em sua reprodutibilidade, todas as ferramentas utilizadas nesse trabalho são distribuídas sob licença de software livre, o que permite a sua utilização sem restrições. Foram desenvolvidos 11 scripts e 3 programas, cujos códigos estão disponíveis nessa dissertação. Por fim, a metodologia foi pensada em formato modular, com passos estanques, característica que permite que trechos específicos da metodologia sejam aplicadas a outras situações, organismos e dados, com alterações mínimas. Pelo seu caráter modular, trechos específicos dessa metodologia podem ser adaptadas para atender a outras demandas similares de tratamento de dados.

Foram geradas duas redes principais, uma com dados oriundos da consulta à página *web* do *STRING*, contendo 70 nós e 529 arestas, e a outra construída a partir dos dados de todas as interações de proteínas disponíveis no banco de dados do *STRING*, contendo 4.839 nós e 129.030 arestas. A análise da distribuição de graus dos nós da rede nos apontou que elas podem ser classificadas na categoria *livre de escala*, o que é um dos indícios que caracteriza uma rede biológica. Integramos à RIPP dados de regulação funcional, *EC number* e tipo de interação entre as proteínas, que associados à estrutura organizacional e a métricas específicas (grau, intermediação e coeficiente de correlação) nos forneceram um cenário propício ao

destaque, mediante estilos de formatação visual, de características subjacentes ao conjunto de dados de entrada.

Os resultados obtidos demonstraram que a metodologia utilizada cumpriu a contento o seu papel, permitindo não só a organização dos dados como a emersão de novas informações que não se encontravam disponíveis inicialmente, tais como agrupamentos funcionais e métricas demonstrativas da importância de proteínas, tudo isso graças ao caráter agregador e integrador das redes.

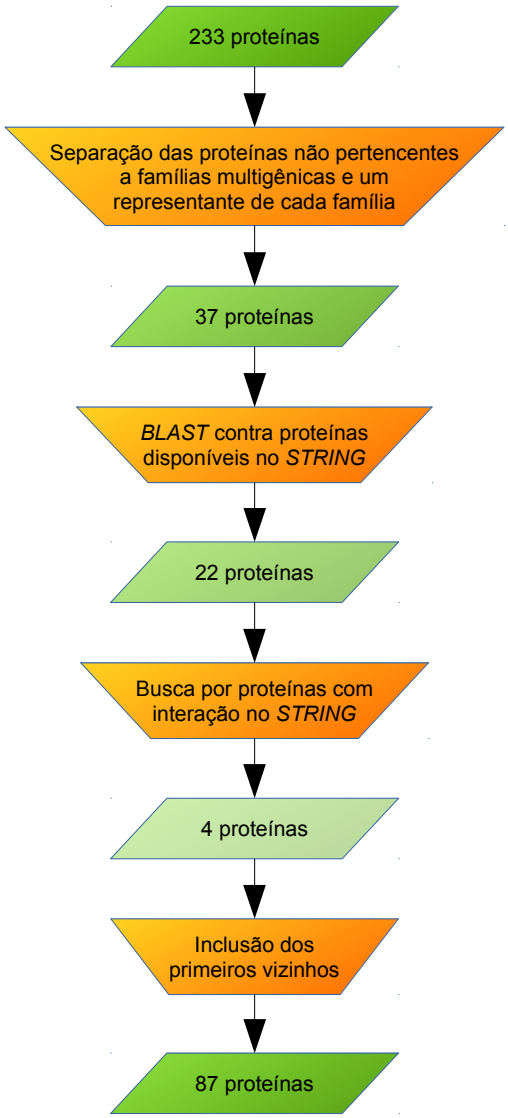


FIGURA 26: Esquema apresentando a variação do volume de dados ao longo do processamento. As sucessivas filtragens reduzem gradualmente os dados disponíveis, até o momento da construção da rede com a adição dos primeiros vizinhos, quando o valor aumenta novamente.

Tal emersão de dados apresenta-se evidente quando analisamos a variação do volume de dados ao longo dos processamentos efetuados, conforme ilustrado na FIGURA 26. Nela percebe-se que os sucessivos procedimentos reduzem gradativamente a quantidade de proteínas disponíveis para análise, até o momento em que elas são pesquisadas no banco de dados de redes. Lá é possível acrescentar os seus primeiros vizinhos, que não eram parte dos dados originais, mas, conforme discutido neste estudo, possuem importância biológica e servem para melhor caracterizar as funções biológicas do organismo.

A partir daí, concluímos então que a mineração de novas informações biológicas funcionais a partir dos dados já existentes é uma característica de destaque na utilização de redes de interação como instrumento de análise de dados de diferentes fontes. Isso ganha especial importância quando consideramos que a estratégia do uso de redes viabiliza um novo olhar estratégico sobre a pesquisa científica da atualidade que via de regra está centrada no estudo de componentes individuais e não no aspecto sistêmico das interações de um organismo. Tendo este prisma direcionando nossos estudos, fica evidente que o potencial de integração de dados biológicos de diferentes fontes (genoma, transcriptoma e proteoma) tem o potencial de trazer a luz e/ou gerar novas hipóteses sobre aspectos ainda obscuros da biologia de tripanossomatídeos que representam os principais agentes de doenças negligenciadas, abrindo perspectivas para novas pesquisas utilizando dados já disponíveis em bancos de dados públicos.

REFERÊNCIAS

ASSENOV, Y. et al. Computing topological parameters of biological networks. **Bioinformatics**, v. 24, n. 2, p. 282–284, 15 jan. 2008.

BAIROCH, A. The ENZYME database in 2000. **Nucleic Acids Research**, v. 28, n. 1, p. 304–305, 1 jan. 2000.

BARABÁSI, A.-L.; OLTVAI, Z. N. Network biology: understanding the cell's functional organization. **Nature Reviews Genetics**, v. 5, n. 2, p. 101–113, Fevereiro 2004.

BORGATTI, S. P. Centrality and network flow. **Social Networks**, v. 27, n. 1, p. 55–71, jan. 2005.

BRENCHLEY, R. et al. The TriTryp Phosphatome: analysis of the protein phosphatase catalytic domains. **BMC Genomics**, v. 8, p. 434, 2007.

CAPRA, M. et al. Frequent Alterations in the Expression of Serine/Threonine Kinases in Human Cancers. **Cancer Research**, v. 66, n. 16, p. 8147–8154, 15 ago. 2006.

CONNERS, E. E. et al. A global systematic review of Chagas disease prevalence among migrants. **Acta Tropica**, v. 156, p. 68–78, Abril 2016.

DAVIDSON, S. B.; OVERTON, C.; BUNEMAN, P. Challenges in Integrating Biological Data Sources. **Journal of Computational Biology**, v. 2, n. 4, p. 557–572, 1 jan. 1995.

DE DIEGO, J. L. et al. The Ubiquitin−Proteasome Pathway Plays an Essential Role in Proteolysis during *Trypanosoma cruzi* Remodeling. **Biochemistry**, v. 40, n. 4, p. 1053–1062, 1 jan. 2001.

Enzyme Nomenclature. Disponível em: <http://www.chem.qmul.ac.uk/iubmb/enzyme/>. Acesso em: 2 fev. 2017.

ERCOLANI, L. et al. A possible S-glutathionylation of specific proteins by glyoxalase II: An in vitro and in silico study. **Cell Biochemistry and Function**, v. 34, n. 8, p. 620–627, Dezembro 2016.

FEASEY, N. et al. Neglected tropical diseases. **British Medical Bulletin**, v. 93, n. 1, p. 179–200, 1 mar. 2010.

FERNÁNDEZ-MOYA, S. M.; ESTÉVEZ, A. M. Posttranscriptional control and the role of RNA-binding proteins in gene regulation in trypanosomatid protozoan parasites. **Wiley Interdisciplinary Reviews - RNA**, v. 1, n. 1, p. 34–46, 1 jul. 2010.

GARCIA-SILVA, M. R. et al. Extracellular vesicles shed by *Trypanosoma cruzi* are linked to small RNA pathways, life cycle regulation, and susceptibility to infection of mammalian cells. **Parasitology Research**, v. 113, n. 1, p. 285–304, 1 jan. 2014.

GARG, M.; GOYAL, N. MAPK1 of Leishmania donovani Modulates Antimony Susceptibility by Downregulating P-Glycoprotein Efflux Pumps. **Antimicrobial Agents and Chemotherapy**, v. 59, n. 7, p. 3853–3863, 1 jul. 2015.

GASCON, J.; BERN, C.; PINAZO, M.-J. Chagas disease in Spain, the United States and other non-endemic countries. **Acta Tropica**, v. 115, n. 1–2, p. 22–27, ago. 2010.

GAUDENZI, J. G. D. et al. Gene expression regulation in trypanosomatids. **Essays In Biochemistry**, v. 51, p. 31–46, 24 out. 2011.

GOBLE, C.; STEVENS, R. State of the nation in data integration for bioinformatics. **Journal of Biomedical Informatics**, Semantic Mashup of Biomedical Data. v. 41, n. 5, p. 687–693, Outubro 2008.

GOMEZ-CABRERO, D. et al. Data integration in the era of omics: current and future challenges. **BMC Systems Biology**, v. 8, n. 2, p. I1, 2014.

GRÜNING, N.-M. et al. Pyruvate Kinase Triggers a Metabolic Feedback Loop that Controls Redox Metabolism in Respiring Cells. **Cell Metabolism**, v. 14, n. 3, p. 415–427, Setembro 2011.

HAILE, S.; PAPADOPOULOU, B. Developmental regulation of gene expression in trypanosomatid parasitic protozoa. **Current Opinion in Microbiology**, Growth and Development. v. 10, n. 6, p. 569–577, Dezembro 2007.

JOY, M. P. et al. High-Betweenness Proteins in the Yeast Protein Interaction Network. **BioMed Research International**, v. 2005, n. 2, p. 96–103, 2005.

KALB, L. C. et al. Clathrin expression in *Trypanosoma cruzi*. **BMC Cell Biology**, v. 15, p. 23, 2014.

KALB, L. C. et al. Conservation and divergence within the clathrin interactome of *Trypanosoma cruzi*. **Scientific Reports**, v. 6, p. 31212, 9 ago. 2016.

KIRCHHOFF, L. V. et al. Ubiquitin genes in trypanosomatidae. **Journal of Biological Chemistry**, v. 263, n. 25, p. 12698–12704, 5 set. 1988.

KISSINGER, J. C. A tale of three genomes: the kinetoplastids have arrived. **Trends in Parasitology**, v. 22, n. 6, p. 240–243, jun. 2006.

LEE, B. Y. et al. Global economic burden of Chagas disease: a computational simulation model. **The Lancet. Infectious Diseases**, v. 13, n. 4, p. 342–348, abr. 2013.

LEHNINGER, A. L.; NELSON, D. L.; COX, M. M. **Lehninger princípios de bioquímica**. [s.l.] ARTMED, 2011.

MARTÍNEZ-CALVILLO, S. et al. Gene Expression in Trypanosomatid Parasites. **BioMed Research International**, v. 2010, p. e525241, 11 fev. 2010.

MERING, C. VON et al. STRING: known and predicted protein–protein associations, integrated and transferred across organisms. **Nucleic Acids Research**, v. 33, n. suppl 1, p. D433–D437, 1 jan. 2005.

NCBI Reference Sequence (RefSeq) Database - Release 80. Disponível em: <ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/RefSeq-release80.txt>. Acesso em: 7 fev. 2017.

PAPADOPOULOU, B. et al. Stage-Specific Regulation of Gene Expression in Leishmania. **ASM News**, v. 69, p. 282–88, 2003.

PARSONS, M. et al. Comparative analysis of the kinomes of three pathogenic trypanosomatids: *Leishmania major*, *Trypanosoma brucei* and *Trypanosoma cruzi*. **BMC Genomics**, v. 6, p. 127, 2005.

PHILIPPI, S. Data and knowledge integration in the life sciences. **Briefings in Bioinformatics**, v. 9, n. 6, p. 451–451, 1 nov. 2008.

PHILIPPI, S.; KÖHLER, J. Addressing the problems with life-science databases for traditional uses and systems biology. **Nature Reviews Genetics**, v. 7, n. 6, p. 482–488, 1 jun. 2006.

Por que negligenciadas? Disponível em: <http://www.dndial.org/pt/doencas-negligenciadas/contexto.html>. Acesso em: 23 nov. 2016.

SHANNON, P. et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. **Genome Research**, v. 13, n. 11, p. 2498–2504, 1 nov. 2003.

SOUZA, W. (ED.). **Doenças negligenciadas**. [s.l.] Academia Brasileira de Ciências, 2010.

STEIN, L. Creating a bioinformatics nation. **Nature**, v. 417, n. 6885, p. 119–120, Maio 2002.

STEIN, L. D. Integrating biological databases. **Nature Reviews Genetics**, v. 4, n. 5, p. 337–345, 1 maio 2003.

STUART, K. et al. Kinetoplastids: related protozoan pathogens, different diseases. **The Journal of Clinical Investigation**, v. 118, n. 4, p. 1301–1310, 1 abr. 2008.

SZKLARCZYK, D. et al. STRING v10: protein–protein interaction networks, integrated over the tree of life. **Nucleic Acids Research**, v. 43, n. D1, p. D447–D452, 28 jan. 2015.

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, jan. 2012.

VALENTE, T. W. et al. How Correlated Are Network Centrality Measures? **Connections (Toronto, Ont.)**, v. 28, n. 1, p. 16–26, 1 jan. 2008.

WAINWRIGHT, M. Dyes, flies, and sunny skies: photodynamic therapy and neglected tropical diseases. **Coloration Technology**, v. 133, n. 1, p. 3–14, Fevereiro 2017.

WHITMARSH, A. J.; DAVIS, R. J. Regulation of transcription factor function by phosphorylation. **Cellular and molecular life sciences: CMLS**, v. 57, n. 8–9, p. 1172–1183, ago. 2000.

WHO, W. H. O. **Chagas disease**. Disponível em: <http://www.who.int/topics/chagas_disease/en/>. Acesso em: 24 nov. 2016.

YOON, J.; BLUMER, A.; LEE, K. An algorithm for modularity analysis of directed and weighted biological networks based on edge-betweenness centrality. **Bioinformatics**, v. 22, n. 24, p. 3106–3108, 15 dez. 2006.

ZINGALES, B. et al. A new consensus for *Trypanosoma cruzi* intraspecific nomenclature: second revision meeting recommends TcI to TcVI. **Memorias Do Instituto Oswaldo Cruz**, v. 104, n. 7, p. 1051–1054, nov. 2009.

ZINGALES, B. *Trypanosoma cruzi* – um parasita, dois parasitas ou vários parasitas da doença de chagas? **Revista da Biologia**, v. 44, n. 48, 2011.

APÊNDICES

APÊNDICE 1 – Conteúdo do arquivo *de_proteins_common_results-non-multigene-one_multigene-only_gene_ids.tsv* que contém a relação dos *gene ids* utilizados no *BLAST* de consulta ao *STRING*.

Gene_id
TCSYLVIO_000604
TCSYLVIO_001905
TCSYLVIO_002420
TCSYLVIO_002437
TCSYLVIO_008205
TCSYLVIO_009320
TCSYLVIO_009850
TCSYLVIO_009378
TCSYLVIO_010762
TCSYLVIO_000562
TCSYLVIO_009900
TCSYLVIO_008058
TCSYLVIO_004013
TCSYLVIO_004114
TCSYLVIO_010859
TCSYLVIO_006635
TCSYLVIO_001125
TCSYLVIO_008689
TCSYLVIO_006124
TCSYLVIO_001850
TCSYLVIO_000129
TCSYLVIO_000362
TCSYLVIO_000575
TCSYLVIO_002521
TCSYLVIO_002797
TCSYLVIO_004793
TCSYLVIO_006143
TCSYLVIO_006144
TCSYLVIO_006746
TCSYLVIO_011104
TCSYLVIO_006674
TCSYLVIO_006975

TCSYLVIO_007621
TCSYLVIO_007701
TCSYLVIO_007917
TCSYLVIO_008159
TCSYLVIO_011072

APÊNDICE 2 – Programa *blast-extract-info.pl*, utilizado para extrair diversas informações do resultado do *BLAST*.

```
#!/usr/bin/perl

use Bio::SearchIO;
use Getopt::Long;

my ($inblast,$outfile);
my $usage = "\n$0 -i blast_file -o output_file\n\n";

GetOptions ('i=s' => \$inblast, 'o=s' => \$outfile);

if(!$inblast or !$outfile) { die $usage;}

if (! -e $inblast) {die "\nCannot find file: $seq\n\n";}

open(OUT ,">$outfile");

my $in = new Bio::SearchIO(-format => 'blast',-file =>
"$inblast");

print OUT "Query name\tQ.length(pb)\tQ.range\tSubject
name\tS.length(pb)\tS.range\tCoverage
length\tCoverage\tExtension\tIdentity\tSimilarity\n";

while(my $result = $in->next_result) {
  if ($result->num_hits == 0){
    print OUT $result->query_name, "\t", $result-
>query_length, "\t", "No hits found\n----------\n";
  }

  else {
    my $count = 0;
    while (my $hit = $result->next_hit){
      $count++;
      $count2=0;
      my $hspcount=0;

      while (my $hsp = $hit->next_hsp){
        $count2++;
        $query_length = $result->query_length;
        @query_coords = $hsp->range('query');
        $hit_length = $hit->length;
        @hit_coords=$hsp->range('hit');
        $coverage_length=abs($query_coords[1]-
$query_coords[0]);
        $coverage_percentage=$coverage_length/$query_length;
```

```
        $extension_percentage=$query_length/$hit_length;
        print OUT $result->query_name,"\t",$query_length,"\t",
$query_coords[0],"-",$query_coords[1],"\t",$hit->name,"\t",
$hit_length,"\t",$hit_coords[0],"-",$hit_coords[1],"\t",
$coverage_length,"\t",int($coverage_percentage*100),"%\t",int(
$extension_percentage*100),"%\t",int($hsp-
>frac_identical*100),"%\t",int($hsp->frac_conserved*100),
"%\n";
        last if ($count2 == $hspcount);
      }

      last if ($count == $numHits);
    }

        print OUT "----------\n";

  }
}
```

APÊNDICE 3 – Conteúdo do arquivo *tcruzi_proteins_from_TcruziSylvioX10-vs-353153.protein.sequences.v10-best_hits.tsv*, com os *best hits* do *BLAST* utilizados como entrada da busca no *STRING,* aqui representado na forma de quadro para melhor visualização do conteúdo.

Query name	Q.length (pb)	Q.range	Subject name	S.length (pb)	S.range	Coverage length	Coverage	Extension	Identity	Similarity
TCSYLVIO_010762	232	1-232	353153.XP_805059.1	543	125-356	231	99%	42%	90%	94%
TCSYLVIO_008689	259	1-259	353153.XP_806860.1	377	119-377	258	99%	68%	91%	95%
TCSYLVIO_004793	835	1-835	353153.XP_817126.1	835	1-835	834	99%	100%	98%	99%
TCSYLVIO_004114	610	1-607	353153.XP_809273.1	1003	1-622	606	99%	60%	90%	92%
TCSYLVIO_007621	490	1-490	353153.XP_804170.1	543	54-543	489	99%	90%	94%	96%
TCSYLVIO_009850	268	1-268	353153.XP_807130.1	342	1-268	267	99%	78%	91%	94%
TCSYLVIO_001850	514	1-514	353153.XP_811157.1	619	106-619	513	99%	83%	98%	99%
TCSYLVIO_004013	298	17-295	353153.XP_803308.1	284	1-279	278	93%	104%	90%	92%
TCSYLVIO_006124	308	1-308	353153.XP_820940.1	766	459-766	307	99%	40%	98%	99%
TCSYLVIO_002420	576	1-576	353153.XP_810255.1	576	1-576	575	99%	100%	95%	96%
TCSYLVIO_008205	314	1-314	353153.XP_808883.1	446	1-314	313	99%	70%	94%	96%
TCSYLVIO_006975	291	1-291	353153.XP_805953.1	358	2-292	290	99%	81%	93%	95%
TCSYLVIO_000604	209	1-209	353153.XP_808829.1	3520	533-741	208	99%	5%	93%	94%
TCSYLVIO_011104	200	1-200	353153.XP_805061.1	920	362-560	199	99%	21%	93%	96%
TCSYLVIO_002521	421	1-421	353153.XP_807213.1	421	1-421	420	99%	100%	98%	98%
TCSYLVIO_000129	518	1-518	353153.XP_807157.1	518	1-518	517	99%	100%	97%	98%
TCSYLVIO_001905	333	1-333	353153.XP_816569.1	333	1-333	332	99%	100%	96%	98%
TCSYLVIO_008058	505	1-505	353153.XP_819384.1	737	233-737	504	99%	68%	92%	95%
TCSYLVIO_009378	476	1-476	353153.XP_811516.1	476	1-476	475	99%	100%	98%	99%
TCSYLVIO_006143	1259	1-1259	353153.XP_807687.1	1461	211-1461	1258	99%	86%	96%	97%
TCSYLVIO_002437	644	1-644	353153.XP_803464.1	644	1-644	643	99%	100%	97%	98%

APÊNDICE 4 – Conteúdo do arquivo *tcruzi_proteins-string_id-regulation.tsv*, que apresenta a associação da regulação de RNA/proteína com o *STRING id,* aqui representado na forma de quadro para melhor visualização do conteúdo.

Subject name	Regulation
353153.XP_803464.1	rna down/ptn down
353153.XP_805059.1	rna down/ptn down
353153.XP_807130.1	rna down/ptn down
353153.XP_808829.1	rna down/ptn down
353153.XP_808883.1	rna down/ptn down
353153.XP_810255.1	rna down/ptn down
353153.XP_811516.1	rna down/ptn down
353153.XP_816569.1	rna down/ptn down
353153.XP_819384.1	rna down/ptn down
353153.XP_803308.1	rna down/ptn up
353153.XP_806860.1	rna down/ptn up
353153.XP_809273.1	rna down/ptn up
353153.XP_811157.1	rna up/ptn down
353153.XP_820940.1	rna up/ptn down
353153.XP_804170.1	rna up/ptn up
353153.XP_805061.1	rna up/ptn up
353153.XP_805953.1	rna up/ptn up
353153.XP_807157.1	rna up/ptn up
353153.XP_807213.1	rna up/ptn up
353153.XP_807687.1	rna up/ptn up
353153.XP_817126.1	rna up/ptn up

APÊNDICE 5 – Programa em *awk* que faz a união, na mesma linha, de todos os tipos de interação relacionadas a cada uma das ligações entre proteínas da rede.

```
#!/usr/bin/awk -f

#################################################################
#                                                               #
# Nome: split_proteins_actions.awk                              #
#                                                               #
# Descrição: separa os tipos de interação presentes no          #
#            arquivo tcruzi_proteins-string_interactions-       #
#            actions_unified.tsv em colunas distintas e pré-    #
#            definidas.                                         #
#                                                               #
# Utilização:                                                   #
#                                                               #
# split_proteins_actions.awk <arquivo_com_tipos_de_interação>#
#                                                               #
# Este código pode ser utilizado livremente desde que seja      #
# citada a fonte. E pode ser incorporado a outros códigos,      #
# desde que esses também sejam disponibilizados livremente.     #
#                                                               #
#################################################################

BEGIN {
       FS="\t"
      }

{
 if ($1=="Interaction") {
 # Impressão do Cabeçalho
 print
"Interaction\tBinding\tReaction\tCatalysis\tExpression\tActiva
tion\tPTMod"

} else {
    # Conjunto de condicionais para procurar cada um dos tipos
de interação
    # e armazená-los em variáveis distintas
    if ($2 ~ "binding") {
    bi = "binding"
    } else {
        bi = ""
    }

    if ($2 ~ "reaction") {
    re = "reaction"
    } else {
```

```
        re = ""
    }

    if ($2 ~ "catalysis") {
    ca = "catalysis"
    } else {
        ca = ""
    }

    if ($2 ~ "expression") {
    ex = "expression"
    } else {
        ex = ""
    }

    if ($2 ~ "activation") {
    ac = "activation"
    } else {
        ac = ""
    }

    if ($2 ~ "ptmod") {
    pt = "ptmod"
    } else {
        pt = ""
    }

    print $1"\t"bi"\t"re"\t"ca"\t"ex"\t"ac"\t"pt

    }

}
```

APÊNDICE 6 – Conteúdo do arquivo *tcruzi_proteins-string_interactions-actions_unified.tsv*, que apresenta a associação entre as interações e todos os tipos associados a elas, aqui representado na forma de quadro para melhor visualização do conteúdo.

Interaction	Action
353153.XP_821226.1 (interacts with) 353153.XP_813201.1	binding catalysis reaction
353153.XP_818122.1 (interacts with) 353153.XP_817146.1	binding
353153.XP_820940.1 (interacts with) 353153.XP_808084.1	binding catalysis ptmod reaction
353153.XP_821624.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_812569.1	binding reaction
353153.XP_820503.1 (interacts with) 353153.XP_817266.1	binding
353153.XP_814744.1 (interacts with) 353153.XP_814609.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_809183.1	binding reaction
353153.XP_818985.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_813966.1	binding reaction
353153.XP_820202.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_811126.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_811126.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_806008.1 (interacts with) 353153.XP_804043.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_814775.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_802255.1	binding reaction
353153.XP_811157.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_820334.1 (interacts with) 353153.XP_816734.1	binding catalysis reaction
353153.XP_820202.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_821226.1 (interacts with) 353153.XP_820958.1	binding catalysis reaction
353153.XP_819691.1 (interacts with) 353153.XP_817266.1	binding reaction
353153.XP_814775.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_818122.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_814734.1 (interacts with) 353153.XP_813863.1	binding catalysis reaction
353153.XP_804193.1 (interacts with) 353153.XP_802255.1	binding
353153.XP_820940.1 (interacts with) 353153.XP_804193.1	binding catalysis ptmod reaction
353153.XP_814744.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_811157.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_804855.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_819681.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_817146.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_810168.1	binding

353153.XP_810205.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_811157.1 (interacts with) 353153.XP_803300.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_810205.1	binding
353153.XP_812569.1 (interacts with) 353153.XP_809183.1	binding
353153.XP_819681.1 (interacts with) 353153.XP_817266.1	binding
353153.XP_811157.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_817146.1 (interacts with) 353153.XP_814775.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_813106.1	activation binding expression
353153.XP_808084.1 (interacts with) 353153.XP_804100.1	binding
353153.XP_820202.1 (interacts with) 353153.XP_819882.1	binding catalysis reaction
353153.XP_817266.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_814552.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_805045.1	binding reaction
353153.XP_813201.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_817266.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_804043.1 (interacts with) 353153.XP_802255.1	binding reaction
353153.XP_820503.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_811029.1	binding
353153.XP_811485.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_817266.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_813700.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_809667.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_814609.1	binding reaction
353153.XP_819681.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_808659.1	binding
353153.XP_811126.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_805045.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_802147.1	binding reaction
353153.XP_818122.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_817266.1 (interacts with) 353153.XP_809183.1	binding reaction
353153.XP_817266.1 (interacts with) 353153.XP_817146.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_811485.1	binding
353153.XP_820334.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_817266.1 (interacts with) 353153.XP_802147.1	binding reaction
353153.XP_814734.1 (interacts with) 353153.XP_813201.1	binding catalysis reaction
353153.XP_821226.1 (interacts with) 353153.XP_816734.1	binding catalysis reaction
353153.XP_820958.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_820940.1 (interacts with) 353153.XP_809667.1	binding catalysis ptmod reaction

353153.XP_804100.1 (interacts with) 353153.XP_802241.1	binding
353153.XP_814744.1 (interacts with) 353153.XP_810205.1	binding binding reaction
353153.XP_819882.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_811485.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_819820.1 (interacts with) 353153.XP_804043.1	binding
353153.XP_818122.1 (interacts with) 353153.XP_813700.1	binding reaction
353153.XP_820503.1 (interacts with) 353153.XP_810205.1	binding
353153.XP_820334.1 (interacts with) 353153.XP_818899.1	binding catalysis reaction
353153.XP_821624.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_821226.1 (interacts with) 353153.XP_818899.1	binding catalysis reaction
353153.XP_819681.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_813863.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_811157.1	binding catalysis reaction
353153.XP_814552.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_820940.1 (interacts with) 353153.XP_804100.1	binding catalysis ptmod reaction
353153.XP_810582.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_817266.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_804193.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_811126.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_819681.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_820958.1 (interacts with) 353153.XP_819882.1	binding catalysis reaction
353153.XP_812569.1 (interacts with) 353153.XP_802147.1	binding
353153.XP_814744.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_819820.1 (interacts with) 353153.XP_817266.1	binding
353153.XP_811126.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_808659.1	binding
353153.XP_814744.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_814775.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_806008.1 (interacts with) 353153.XP_804855.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_817266.1	binding
353153.XP_820940.1 (interacts with) 353153.XP_812569.1	binding catalysis ptmod reaction
353153.XP_810582.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_811126.1 (interacts with) 353153.XP_803300.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_810168.1	binding
353153.XP_810582.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_814775.1 (interacts with) 353153.XP_806008.1	binding

353153.XP_821488.1 (interacts with) 353153.XP_818985.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_814775.1	binding
353153.XP_820940.1 (interacts with) 353153.XP_802255.1	binding catalysis ptmod reaction
353153.XP_816312.1 (interacts with) 353153.XP_814552.1	binding
353153.XP_812569.1 (interacts with) 353153.XP_802241.1	binding
353153.XP_818899.1 (interacts with) 353153.XP_814190.1	binding catalysis reaction
353153.XP_821226.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_809183.1	binding reaction
353153.XP_819820.1 (interacts with) 353153.XP_810205.1	binding
353153.XP_820202.1 (interacts with) 353153.XP_813201.1	binding catalysis reaction
353153.XP_811485.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_813445.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_821800.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_819820.1 (interacts with) 353153.XP_819681.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_820958.1 (interacts with) 353153.XP_813863.1	binding catalysis reaction
353153.XP_818826.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_814190.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_821800.1 (interacts with) 353153.XP_821624.1	binding reaction
353153.XP_817266.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_810168.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_819681.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_820503.1	binding
353153.XP_818899.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_806008.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_819681.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_818826.1 (interacts with) 353153.XP_814552.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_819681.1 (interacts with) 353153.XP_814775.1	binding
353153.XP_808084.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_802241.1	binding reaction
353153.XP_819681.1 (interacts with) 353153.XP_814552.1	binding reaction
353153.XP_816734.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_816312.1 (interacts with) 353153.XP_811185.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_808084.1	binding reaction
353153.XP_819820.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_819691.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_817266.1 (interacts with) 353153.XP_802241.1	binding reaction
353153.XP_820940.1 (interacts with) 353153.XP_805045.1	binding catalysis ptmod reaction

353153.XP_817266.1 (interacts with) 353153.XP_808084.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_810168.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_811485.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_814190.1 (interacts with) 353153.XP_813445.1	binding catalysis reaction
353153.XP_818899.1 (interacts with) 353153.XP_814734.1	binding catalysis reaction
353153.XP_814744.1 (interacts with) 353153.XP_814552.1	binding reaction
353153.XP_809183.1 (interacts with) 353153.XP_804193.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_814775.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_803300.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_814552.1	binding
353153.XP_810582.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_820958.1 (interacts with) 353153.XP_820334.1	binding catalysis reaction
353153.XP_821800.1 (interacts with) 353153.XP_820503.1	activation binding expression
353153.XP_811185.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_816734.1 (interacts with) 353153.XP_814734.1	binding catalysis reaction
353153.XP_817146.1 (interacts with) 353153.XP_813700.1	binding
353153.XP_811485.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_819681.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_804193.1	binding reaction
353153.XP_818122.1 (interacts with) 353153.XP_814744.1	binding reaction
353153.XP_819691.1 (interacts with) 353153.XP_814775.1	binding reaction
353153.XP_819681.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_809018.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_814744.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_820958.1 (interacts with) 353153.XP_813201.1	binding catalysis reaction
353153.XP_812569.1 (interacts with) 353153.XP_808084.1	binding
353153.XP_817266.1 (interacts with) 353153.XP_804193.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_802147.1	binding reaction
353153.XP_813106.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_816734.1 (interacts with) 353153.XP_814190.1	binding catalysis reaction
353153.XP_821800.1 (interacts with) 353153.XP_814744.1	binding reaction
353153.XP_806008.1 (interacts with) 353153.XP_803300.1	binding
353153.XP_813700.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_819681.1 (interacts with) 353153.XP_817146.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_802575.1	binding

353153.XP_819681.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_813966.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_819820.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_814734.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_804975.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_811029.1	binding
353153.XP_820202.1 (interacts with) 353153.XP_816734.1	binding catalysis reaction
353153.XP_807687.1 (interacts with) 353153.XP_804127.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_811485.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_810582.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_810168.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_819820.1 (interacts with) 353153.XP_814775.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_817146.1	binding
353153.XP_807687.1 (interacts with) 353153.XP_806042.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_814775.1 (interacts with) 353153.XP_802241.1	binding reaction
353153.XP_819820.1 (interacts with) 353153.XP_814552.1	binding
353153.XP_813700.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_814775.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_814744.1	binding
353153.XP_820202.1 (interacts with) 353153.XP_818899.1	binding catalysis reaction
353153.XP_813106.1 (interacts with) 353153.XP_804043.1	activation binding expression
353153.XP_813700.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_814552.1	binding reaction
353153.XP_814552.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_820958.1 (interacts with) 353153.XP_820202.1	binding catalysis reaction
353153.XP_813700.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_809667.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_819820.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_821488.1	binding
353153.XP_819820.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_819820.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_818899.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_821624.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_813201.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_804100.1	binding reaction

353153.XP_820334.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_821624.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_814552.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_819820.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_819882.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_817266.1 (interacts with) 353153.XP_814744.1	binding
353153.XP_814552.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_818122.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_814609.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_811157.1 (interacts with) 353153.XP_802575.1	activation binding ptmod
353153.XP_812569.1 (interacts with) 353153.XP_809667.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_804100.1 (interacts with) 353153.XP_802255.1	binding
353153.XP_820958.1 (interacts with) 353153.XP_816734.1	binding catalysis reaction
353153.XP_819820.1 (interacts with) 353153.XP_817146.1	binding
353153.XP_820202.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_813863.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_821800.1 (interacts with) 353153.XP_818122.1	binding reaction
353153.XP_818985.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_803300.1	binding
353153.XP_820940.1 (interacts with) 353153.XP_819691.1	binding catalysis ptmod reaction
353153.XP_811185.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_819681.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_814609.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_809183.1	binding reaction
353153.XP_811157.1 (interacts with) 353153.XP_804975.1	activation binding ptmod
353153.XP_821624.1 (interacts with) 353153.XP_817146.1	binding
353153.XP_820334.1 (interacts with) 353153.XP_814734.1	binding catalysis reaction
353153.XP_821624.1 (interacts with) 353153.XP_813966.1	binding reaction
353153.XP_810168.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_820334.1 (interacts with) 353153.XP_813445.1	binding catalysis reaction
353153.XP_818826.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_813700.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_819882.1 (interacts with) 353153.XP_814734.1	binding catalysis reaction
353153.XP_814775.1 (interacts with) 353153.XP_808084.1	binding reaction
353153.XP_820958.1 (interacts with) 353153.XP_818899.1	binding catalysis reaction
353153.XP_819681.1 (interacts with) 353153.XP_818826.1	binding reaction
353153.XP_819681.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_804100.1 (interacts with) 353153.XP_804043.1	binding reaction

353153.XP_811202.1 (interacts with) 353153.XP_807157.1	binding
353153.XP_804193.1 (interacts with) 353153.XP_802147.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_802255.1	binding reaction
353153.XP_820503.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_809183.1 (interacts with) 353153.XP_804100.1	binding
353153.XP_818122.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_813863.1 (interacts with) 353153.XP_813445.1	binding catalysis reaction
353153.XP_813445.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_817266.1 (interacts with) 353153.XP_809667.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_819882.1 (interacts with) 353153.XP_814190.1	binding catalysis reaction
353153.XP_814190.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_820503.1 (interacts with) 353153.XP_813700.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_814609.1	binding
353153.XP_811157.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_818985.1 (interacts with) 353153.XP_811157.1	activation binding ptmod
353153.XP_814190.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_813966.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_821226.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_813700.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_814775.1 (interacts with) 353153.XP_804193.1	binding reaction
353153.XP_811485.1 (interacts with) 353153.XP_804193.1	binding reaction
353153.XP_814552.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_804043.1 (interacts with) 353153.XP_802147.1	binding reaction
353153.XP_816734.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_813700.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_817266.1 (interacts with) 353153.XP_804100.1	binding reaction
353153.XP_818826.1 (interacts with) 353153.XP_804193.1	binding reaction
353153.XP_810168.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_813700.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_805045.1	binding reaction
353153.XP_818899.1 (interacts with) 353153.XP_813863.1	binding catalysis reaction
353153.XP_812569.1 (interacts with) 353153.XP_802255.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_814744.1	binding
353153.XP_813700.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_820958.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_817266.1 (interacts with) 353153.XP_805045.1	binding reaction
353153.XP_818122.1 (interacts with) 353153.XP_814609.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_804043.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_814552.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_812569.1	binding reaction
353153.XP_821226.1 (interacts with) 353153.XP_814734.1	binding catalysis reaction

353153.XP_813106.1 (interacts with) 353153.XP_810205.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_802147.1	binding reaction
353153.XP_821226.1 (interacts with) 353153.XP_813445.1	binding catalysis reaction
353153.XP_813966.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_804193.1	binding reaction
353153.XP_817266.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_817266.1 (interacts with) 353153.XP_812569.1	binding reaction
353153.XP_818826.1 (interacts with) 353153.XP_813700.1	binding reaction
353153.XP_814552.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_820627.1 (interacts with) 353153.XP_807157.1	binding
353153.XP_819820.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_814552.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_819681.1 (interacts with) 353153.XP_813700.1	binding reaction
353153.XP_810205.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_811126.1 (interacts with) 353153.XP_802575.1	binding
353153.XP_817266.1 (interacts with) 353153.XP_802255.1	binding reaction
353153.XP_819820.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_820503.1 (interacts with) 353153.XP_814744.1	binding
353153.XP_814552.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_818985.1 (interacts with) 353153.XP_816312.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_814734.1 (interacts with) 353153.XP_813445.1	binding catalysis reaction
353153.XP_811126.1 (interacts with) 353153.XP_804975.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_817266.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_818826.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_811157.1 (interacts with) 353153.XP_811029.1	binding
353153.XP_814744.1 (interacts with) 353153.XP_813700.1	binding reaction
353153.XP_813106.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_809667.1	binding reaction
353153.XP_811029.1 (interacts with) 353153.XP_808659.1	binding
353153.XP_813106.1 (interacts with) 353153.XP_808659.1	binding
353153.XP_821800.1 (interacts with) 353153.XP_804043.1	binding expression
353153.XP_821624.1 (interacts with) 353153.XP_802241.1	binding reaction
353153.XP_814609.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_814609.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_814744.1	binding
353153.XP_814552.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_817266.1 (interacts with) 353153.XP_804043.1	binding

353153.XP_814190.1 (interacts with) 353153.XP_813201.1	binding catalysis reaction
353153.XP_804193.1 (interacts with) 353153.XP_802241.1	binding
353153.XP_814734.1 (interacts with) 353153.XP_814190.1	binding catalysis reaction
353153.XP_814609.1 (interacts with) 353153.XP_803300.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_814775.1 (interacts with) 353153.XP_804100.1	binding reaction
353153.XP_813863.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_814609.1 (interacts with) 353153.XP_813966.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_819681.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_814744.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_817266.1	binding
353153.XP_816734.1 (interacts with) 353153.XP_813863.1	binding catalysis reaction
353153.XP_820940.1 (interacts with) 353153.XP_809183.1	binding catalysis ptmod reaction
353153.XP_820503.1 (interacts with) 353153.XP_819820.1	binding
353153.XP_820202.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_819681.1 (interacts with) 353153.XP_814744.1	binding reaction
353153.XP_808659.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_811157.1 (interacts with) 353153.XP_804043.1	binding catalysis reaction
353153.XP_812569.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_804193.1	binding reaction
353153.XP_804043.1 (interacts with) 353153.XP_802241.1	binding reaction
353153.XP_814775.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_812569.1	binding reaction
353153.XP_818122.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_802575.1	binding
353153.XP_809183.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_804975.1	binding
353153.XP_820334.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_819820.1 (interacts with) 353153.XP_813700.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_802255.1	binding reaction
353153.XP_813106.1 (interacts with) 353153.XP_810168.1	binding
353153.XP_820202.1 (interacts with) 353153.XP_813445.1	binding catalysis reaction
353153.XP_821624.1 (interacts with) 353153.XP_813700.1	binding reaction
353153.XP_819882.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_820334.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_813700.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_821226.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_810205.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_808084.1	binding reaction
353153.XP_820503.1 (interacts with) 353153.XP_818122.1	binding
353153.XP_818122.1 (interacts with) 353153.XP_802485.1	binding

353153.XP_818122.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_804855.1	binding
353153.XP_809667.1 (interacts with) 353153.XP_804193.1	binding
353153.XP_813863.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_813201.1 (interacts with) 353153.XP_809018.1	binding catalysis reaction
353153.XP_821800.1 (interacts with) 353153.XP_819691.1	binding reaction
353153.XP_818122.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_811157.1 (interacts with) 353153.XP_804855.1	activation binding ptmod
353153.XP_811485.1 (interacts with) 353153.XP_804100.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_810205.1	binding
353153.XP_811126.1 (interacts with) 353153.XP_811029.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_804100.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_808659.1 (interacts with) 353153.XP_803300.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_820503.1	binding
353153.XP_820334.1 (interacts with) 353153.XP_819882.1	binding catalysis reaction
353153.XP_811157.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_813106.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_805045.1	binding reaction
353153.XP_821226.1 (interacts with) 353153.XP_819882.1	binding catalysis reaction
353153.XP_814552.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_820940.1 (interacts with) 353153.XP_802147.1	binding catalysis ptmod reaction
353153.XP_821624.1 (interacts with) 353153.XP_821488.1	binding
353153.XP_811185.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_811485.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_811029.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_818122.1	binding reaction
353153.XP_813106.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_814744.1 (interacts with) 353153.XP_804100.1	binding reaction
353153.XP_819820.1 (interacts with) 353153.XP_814744.1	binding
353153.XP_814609.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_818826.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_818826.1 (interacts with) 353153.XP_812569.1	binding reaction
353153.XP_819681.1 (interacts with) 353153.XP_818122.1	binding reaction
353153.XP_821488.1 (interacts with) 353153.XP_808659.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_814744.1	binding reaction
353153.XP_806042.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_819681.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_821800.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_811157.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_821226.1 (interacts with) 353153.XP_804127.1	binding catalysis reaction
353153.XP_820202.1 (interacts with) 353153.XP_814734.1	binding catalysis reaction

353153.XP_813966.1 (interacts with) 353153.XP_811157.1	binding reaction
353153.XP_811126.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_820940.1 (interacts with) 353153.XP_802241.1	binding catalysis ptmod reaction
353153.XP_812569.1 (interacts with) 353153.XP_811485.1	binding reaction
353153.XP_820334.1 (interacts with) 353153.XP_813201.1	binding catalysis reaction
353153.XP_810205.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_811126.1	binding reaction
353153.XP_814744.1 (interacts with) 353153.XP_812569.1	binding reaction
353153.XP_806008.1 (interacts with) 353153.XP_802575.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_814775.1 (interacts with) 353153.XP_804043.1	binding
353153.XP_820958.1 (interacts with) 353153.XP_814190.1	binding catalysis reaction
353153.XP_818122.1 (interacts with) 353153.XP_814552.1	binding reaction
353153.XP_808659.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_810582.1	binding
353153.XP_819882.1 (interacts with) 353153.XP_813863.1	binding catalysis reaction
353153.XP_806008.1 (interacts with) 353153.XP_804975.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_814609.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_811029.1	binding
353153.XP_809667.1 (interacts with) 353153.XP_804100.1	binding
353153.XP_813863.1 (interacts with) 353153.XP_813201.1	binding catalysis reaction
353153.XP_813106.1 (interacts with) 353153.XP_811157.1	binding
353153.XP_817146.1 (interacts with) 353153.XP_810205.1	binding
353153.XP_813445.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_814734.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_817266.1 (interacts with) 353153.XP_802485.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_809667.1	binding reaction
353153.XP_821624.1 (interacts with) 353153.XP_819820.1	binding
353153.XP_814190.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_816312.1 (interacts with) 353153.XP_814609.1	binding
353153.XP_818899.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_818122.1 (interacts with) 353153.XP_810582.1	binding reaction
353153.XP_818122.1 (interacts with) 353153.XP_813106.1	binding
353153.XP_811126.1 (interacts with) 353153.XP_804855.1	binding
353153.XP_820503.1 (interacts with) 353153.XP_804043.1	activation binding expression
353153.XP_816734.1 (interacts with) 353153.XP_807687.1	binding
353153.XP_808084.1 (interacts with) 353153.XP_804193.1	binding
353153.XP_821624.1 (interacts with) 353153.XP_804100.1	binding reaction
353153.XP_814609.1 (interacts with) 353153.XP_813700.1	binding reaction
353153.XP_817146.1 (interacts with) 353153.XP_808659.1	binding
353153.XP_821488.1 (interacts with) 353153.XP_814775.1	binding

353153.XP_818122.1 (interacts with) 353153.XP_806008.1	binding
353153.XP_818826.1 (interacts with) 353153.XP_814609.1	binding reaction
353153.XP_811126.1 (interacts with) 353153.XP_810205.1	binding reaction
353153.XP_811157.1 (interacts with) 353153.XP_808659.1	binding reaction
353153.XP_816312.1 (interacts with) 353153.XP_804043.1	binding
353153.XP_820958.1 (interacts with) 353153.XP_813445.1	binding catalysis reaction
353153.XP_804100.1 (interacts with) 353153.XP_802147.1	binding
353153.XP_819681.1 (interacts with) 353153.XP_814609.1	binding reaction
353153.XP_810205.1 (interacts with) 353153.XP_810168.1	binding reaction
353153.XP_819691.1 (interacts with) 353153.XP_804043.1	binding reaction
353153.XP_819820.1 (interacts with) 353153.XP_811126.1	binding
353153.XP_809018.1 (interacts with) 353153.XP_806042.1	binding catalysis reaction
353153.XP_821800.1 (interacts with) 353153.XP_806008.1	binding

APÊNDICE 7 – Programa *split_string_proteins_details.awk*, utilizado para extrair diversas informações do resultado do *BLAST*.

```
#!/usr/bin/awk -f

#################################################################
#                                                               #
# Nome: split_string_proteins_details.awk                       #
#                                                               #
# Descrição: reestrutura o arquivo de anotações de proteínas    #
#            gerado na aba "Tables / Exports" do STRING v.10    #
#            (opção "... protein annotations"), colocando a     #
#            proteína na primeira coluna, a sua anotação        #
#            funcional na segunda e cada um dos nomes ou        #
#            informações associadas na terceira, uma            #
#            informação por linha.                              #
#                                                               #
# Utilização:                                                   #
#                                                               #
# split_string_proteins_details.awk <arquivo_de_anotações>      #
#                                                               #
# Este código pode ser utilizado livremente desde que seja      #
# citada a fonte. E pode ser incorporado a outros códigos,      #
# desde que esses também sejam disponibilizados livremente.     #
#################################################################

BEGIN {
       FS="\t"
       # Imprime a linha do cabeçalho
       print "string_id\tfunctional_annotation\tprotein_info"
      }

{
 # Condicional para eliminar o cabeçalho original
 if (substr($1,1,1) != "#") {
   # Quebra o campo com a lista de proteínas,gerando um array
   split(substr($0, index($0,$5)),aliases,"\t")
   # Condicional executada para cada elemento do array
   for (id in aliases)
     print $2"\t"$4"\t"aliases[id]
 }
}
```

APÊNDICE 8 – Conteúdo do arquivo *tcruzi_proteins-string_proteins_annotations-ec_numbers.tsv*, que relaciona as proteínas presentes na rede ao seu *EC number*, caso exista uma associação. O arquivo está representado na forma de quadro para melhor visualização do conteúdo.

string_id	functional_annotation	protein_info
353153.XP_804043.1	dual specificity protein phosphatase	EC:3.1.3.48
353153.XP_804170.1	surface protease GP63	EC:3.4.24.36
353153.XP_805059.1	surface protease GP63	EC:3.4.24.36
353153.XP_806008.1	hypothetical protein	EC:6.3.2.19
353153.XP_808084.1	DNA-directed RNA polymerase I largest subunit	EC:2.7.7.6
353153.XP_809667.1	DNA-directed RNA polymerase I largest subunit	EC:2.7.7.6
353153.XP_811202.1	pyruvate kinase 2	EC:2.7.1.40
353153.XP_812569.1	RNA polymerase IIA largest subunit	EC:2.7.7.6
353153.XP_816312.1	hypothetical protein	EC:6.3.2.19
353153.XP_816569.1	cysteine peptidase C (CPC)	EC:3.4.22.-
353153.XP_818575.1	hydroxyacylglutathione hydrolase	EC:3.1.2.6
353153.XP_818985.1	serine/arginine-rich protein specific kinase SRPK	EC:2.7.11.1
353153.XP_819691.1	DNA-directed RNA polymerase III largest subunit	EC:2.7.7.6
353153.XP_820171.1	hydroxyacylglutathione hydrolase	EC:3.1.2.6
353153.XP_820627.1	pyruvate kinase 2	EC:2.7.1.40
353153.XP_821800.1	dual specificity protein phosphatase	EC:3.1.3.48

APÊNDICE 9 – Conteúdo do arquivo *tcruzi_proteins-string_proteins_annotations-ec_numbers_details.dat*, contendo os detalhes de cada um dos *EC numbers* encontrados na rede. Para reduzir a quantidade de linhas a serem exibidas neste trabalho (2520 no arquivo original), foram retiradas as informações de referência cruzada com o *Swiss-Prot*.

```
ID   2.7.11.1
DE   Non-specific serine/threonine protein kinase.
AN   Protein phosphokinase.
AN   Protein serine kinase.
AN   Protein serine-threonine kinase.
AN   Protein-serine kinase.
AN   Serine kinase.
AN   Serine protein kinase.
AN   Serine(threonine) protein kinase.
AN   Serine-specific protein kinase.
AN   Serine/threonine protein kinase.
AN   Threonine-specific protein kinase.
CA   ATP + a protein = ADP + a phosphoprotein.
CC   -!- This is a heterogeneous group of serine/threonine protein kinases
CC       that do not have an activating compound and are either non-specific
CC       or their specificity has not been analyzed to date.
CC   -!- Formerly EC 2.7.1.37 and EC 2.7.1.70.
PR   PROSITE; PDOC00100;
//
ID   2.7.1.40
DE   Pyruvate kinase.
AN   Phosphoenol transphosphorylase.
AN   Phosphoenolpyruvate kinase.
CA   ATP + pyruvate = ADP + phosphoenolpyruvate.
CC   -!- UTP, GTP, CTP, ITP and dATP can also act as donors.
CC   -!- Also phosphorylates hydroxylamine and fluoride in the presence of
CC       CO(2).
PR   PROSITE; PDOC00101;
//
ID   2.7.7.6
DE   DNA-directed RNA polymerase.
AN   DNA-dependent RNA polymerase.
AN   RNA nucleotidyltransferase (DNA-directed).
AN   RNA polymerase I.
AN   RNA polymerase II.
AN   RNA polymerase III.
CA   Nucleoside triphosphate + RNA(n) = diphosphate + RNA(n+1).
CC   -!- Catalyzes DNA-template-directed extension of the 3'-end of an RNA
CC       strand by one nucleotide at a time.
CC   -!- Can initiate a chain de novo.
CC   -!- In eukaryotes three forms of the enzyme have been distinguished on
CC       the basis of sensitivity of alpha-amanitin, and the type of RNA
CC       synthesized.
CC   -!- See also EC 2.7.7.19 and EC 2.7.7.48.
PR   PROSITE; PDOC00106;
PR   PROSITE; PDOC00410;
PR   PROSITE; PDOC00411;
PR   PROSITE; PDOC00790;
PR   PROSITE; PDOC00854;
PR   PROSITE; PDOC00855;
PR   PROSITE; PDOC00856;
PR   PROSITE; PDOC00889;
PR   PROSITE; PDOC00896;
//
```

```
ID   3.1.2.6
DE   Hydroxyacylglutathione hydrolase.
AN   Glyoxalase II.
CA   S-(2-hydroxyacyl)glutathione + H(2)O = glutathione + a 2-hydroxy
CA   carboxylate.
CC   -!- Also hydrolyzes S-acetoacetylglutathione, more slowly.
CC   -!- Formerly EC 3.1.2.8.
//
ID   3.1.3.48
DE   Protein-tyrosine-phosphatase.
AN   Phosphotyrosine phosphatase.
AN   PTPase.
CA   Protein tyrosine phosphate + H(2)O = protein tyrosine + phosphate.
CC   -!- Dephosphorylates O-phosphotyrosine groups in phosphoproteins, such as
CC       the products of EC 2.7.10.2.
PR   PROSITE; PDOC00323;
//
ID   3.4.24.36
DE   Leishmanolysin.
AN   Glycoprotein gp63.
AN   Promastigote surface endopeptidase.
CA   Preference for hydrophobic residues at P1 and P1' and basic residues at
CA   P2' and P3'. A model nonapeptide is cleaved at -Ala-Tyr-|-Leu-Lys-Lys-.
CF   Ca(2+); Zn(2+).
CC   -!- A membrane-bound glycoprotein found on the promastigote of various
CC       species of Leishmania protozoans.
CC   -!- Contains consensus sequence for a zinc binding site; Z-Tyr-Leu-NHOH
CC       is a strong inhibitor.
CC   -!- The enzyme can activate its proenzyme by cleavage of the 100-
CC       Val-|-Val-101 bond.
CC   -!- An acidic pH optimum is found with certain protein substrates.
CC   -!- Belongs to peptidase family M8.
PR   PROSITE; PDOC00129;
//
ID   6.3.2.19
DE   Transferred entry: 2.3.2.23, 2.3.2.27 and 6.2.1.45.
//
```

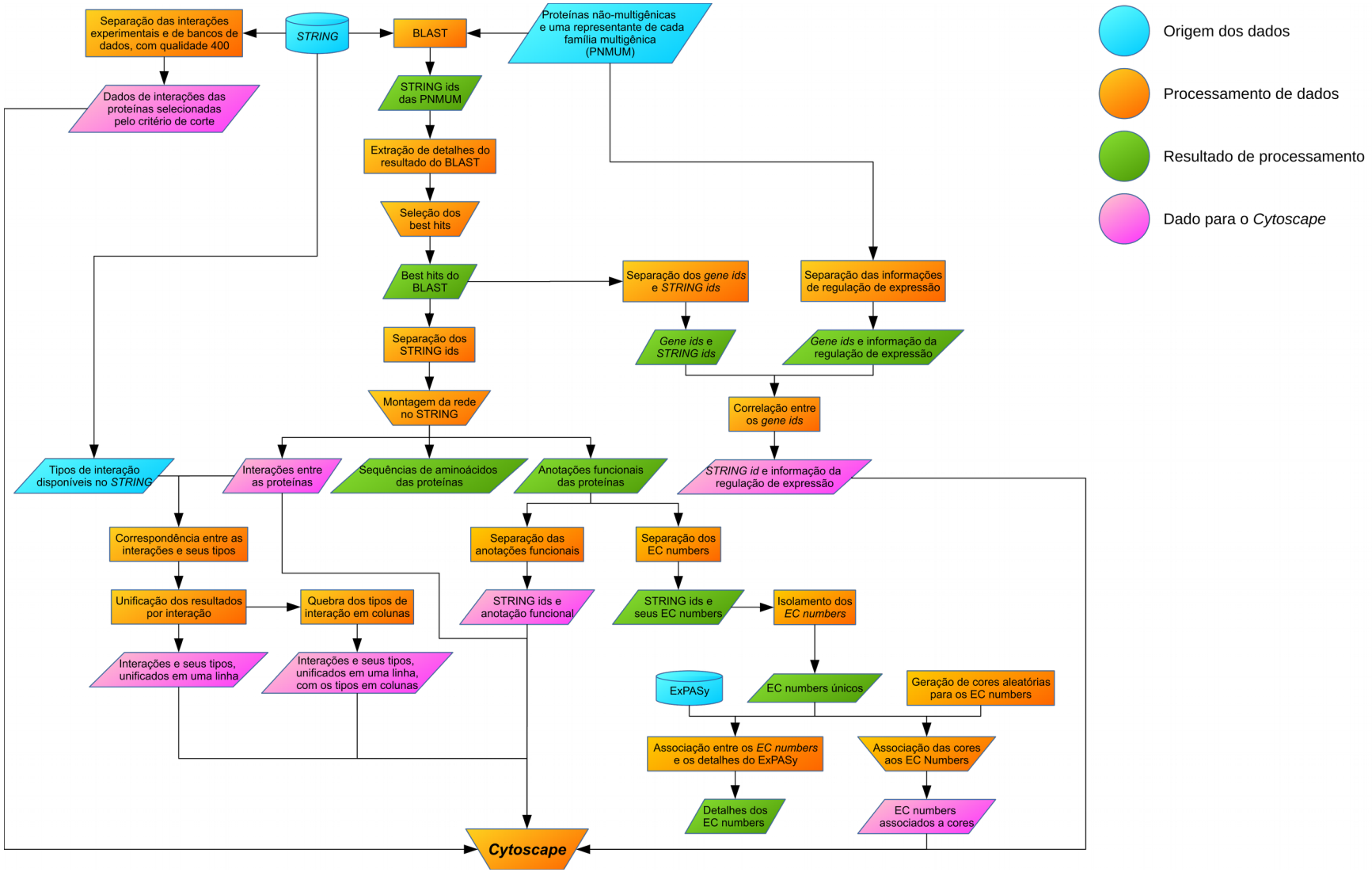
APÊNDICE 10 – Conteúdo do arquivo *tcruzi_proteins-string_proteins_annotations-functional_annotation.tsv*, gerado na seção 3.7.4, e que contém as anotações funcionais de todas as proteínas presentes na rede, aqui representado na forma de quadro para melhor visualização do conteúdo.

string_id	functional_annotation
353153.XP_802143.1	hypothetical protein
353153.XP_802147.1	DNA-directed RNA polymerase I largest subunit
353153.XP_802241.1	DNA-directed RNA polymerase I largest subunit
353153.XP_802255.1	DNA-directed RNA polymerase I largest subunit
353153.XP_802485.1	hypothetical protein
353153.XP_802575.1	protein kinase
353153.XP_803300.1	branch point binding protein
353153.XP_803308.1	hypothetical protein
353153.XP_803464.1	lysosomal/endosomal membrane protein p67
353153.XP_804043.1	dual specificity protein phosphatase
353153.XP_804100.1	RNA polymerase IIA largest subunit
353153.XP_804127.1	clathrin coat assembly protein AP19
353153.XP_804170.1	surface protease GP63
353153.XP_804193.1	RNA polymerase IIA largest subunit
353153.XP_804855.1	protein kinase
353153.XP_804975.1	protein kinase
353153.XP_805045.1	DNA-directed RNA polymerase III largest subunit
353153.XP_805059.1	surface protease GP63
353153.XP_805953.1	metacaspase
353153.XP_806008.1	hypothetical protein
353153.XP_806042.1	mu-adaptin 4
353153.XP_806860.1	UDP-Gal or UDP-GlcNAc-dependent glycosyltransferase
353153.XP_807130.1	lipase
353153.XP_807157.1	actin interacting protein-like protein
353153.XP_807213.1	hypothetical protein
353153.XP_807687.1	hypothetical protein
353153.XP_808084.1	DNA-directed RNA polymerase I largest subunit
353153.XP_808659.1	hypothetical protein
353153.XP_808883.1	trans-sialidase
353153.XP_809018.1	clathrin coat assembly protein AP19
353153.XP_809183.1	DNA-directed RNA polymerase I largest subunit
353153.XP_809273.1	hypothetical protein
353153.XP_809667.1	DNA-directed RNA polymerase I largest subunit
353153.XP_810168.1	U2 small nuclear ribonucleoprotein B
353153.XP_810205.1	spliceosome-associated protein

353153.XP_810255.1	hypothetical protein
353153.XP_810582.1	splicing factor 3A
353153.XP_811029.1	branch point binding protein
353153.XP_811126.1	hypothetical protein
353153.XP_811157.1	RNA-binding protein
353153.XP_811185.1	snoRNP protein gar1
353153.XP_811202.1	pyruvate kinase 2
353153.XP_811485.1	RNA-binding protein
353153.XP_811516.1	amino acid transporter
353153.XP_812569.1	RNA polymerase IIA largest subunit
353153.XP_813106.1	hypothetical protein
353153.XP_813201.1	mu-adaptin 4
353153.XP_813445.1	clathrin coat assembly protein
353153.XP_813700.1	small nuclear ribonucleoprotein Sm-F
353153.XP_813863.1	beta-adaptin 4
353153.XP_813966.1	nucleolar RNA-binding protein
353153.XP_814190.1	beta-adaptin
353153.XP_814552.1	small nuclear ribonucleoprotein Sm-F
353153.XP_814609.1	hypothetical protein
353153.XP_814734.1	clathrin assembly sigma-adaptin protein complex 4
353153.XP_814744.1	splicing factor 3B subunit 1
353153.XP_814775.1	hypothetical protein
353153.XP_816050.1	hypothetical protein
353153.XP_816312.1	hypothetical protein
353153.XP_816569.1	cysteine peptidase C (CPC)
353153.XP_816734.1	mu-adaptin 1
353153.XP_817126.1	hypothetical protein
353153.XP_817146.1	hypothetical protein
353153.XP_817266.1	hypothetical protein
353153.XP_818122.1	U2 small nuclear ribonucleoprotein B
353153.XP_818575.1	hydroxyacylglutathione hydrolase
353153.XP_818826.1	RNA-binding protein
353153.XP_818899.1	mu-adaptin 1
353153.XP_818985.1	serine/arginine-rich protein specific kinase SRPK
353153.XP_819384.1	trans-sialidase
353153.XP_819681.1	splicing factor 3A
353153.XP_819691.1	DNA-directed RNA polymerase III largest subunit
353153.XP_819820.1	hypothetical protein
353153.XP_819882.1	clathrin coat assembly protein
353153.XP_820171.1	hydroxyacylglutathione hydrolase
353153.XP_820202.1	beta-adaptin 1
353153.XP_820334.1	beta-adaptin

353153.XP_820503.1	hypothetical protein
353153.XP_820627.1	pyruvate kinase 2
353153.XP_820940.1	hypothetical protein
353153.XP_820958.1	clathrin assembly sigma-adaptin protein complex 4
353153.XP_821226.1	adaptin-related protein-like
353153.XP_821488.1	hypothetical protein
353153.XP_821624.1	PRP8 protein homologue
353153.XP_821800.1	dual specificity protein phosphatase

APÊNDICE 11 – Fluxograma resumido de todas as etapas de tratamento dos dados integrados no *Cytoscape*. As cores indicam a categoria de cada um dos elementos, segundo a legenda ao lado do fluxograma.



APÊNDICE 12 – Conteúdo do arquivo de estilo *t.cruzi-ec_number-expression-interaction.xml*, utilizado na formatação da rede apresentada na FIGURA 19.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<vizmap documentVersion="3.0" id="VizMap-2017_05_27-18_21">
    <visualStyle name="t.cruzi-ec-expression-interaction">
        <network>
            <visualProperty name="NETWORK_NODE_SELECTION" default="true"/>
            <visualProperty name="NETWORK_SCALE_FACTOR" default="1.0"/>
            <visualProperty name="NETWORK_DEPTH" default="0.0"/>
            <visualProperty name="NETWORK_HEIGHT" default="400.0"/>
            <visualProperty name="NETWORK_CENTER_X_LOCATION" default="0.0"/>
            <visualProperty name="NETWORK_WIDTH" default="550.0"/>
            <visualProperty name="NETWORK_TITLE" default=""/>
            <visualProperty name="NETWORK_BACKGROUND_PAINT" default="#FFFFFF"/>
            <visualProperty name="NETWORK_CENTER_Y_LOCATION" default="0.0"/>
            <visualProperty name="NETWORK_SIZE" default="550.0"/>
            <visualProperty name="NETWORK_CENTER_Z_LOCATION" default="0.0"/>
            <visualProperty name="NETWORK_EDGE_SELECTION" default="true"/>
        </network>
        <node>
            <dependency name="nodeCustomGraphicsSizeSync" value="true"/>
            <dependency name="nodeSizeLocked" value="true"/>
            <visualProperty name="NODE_LABEL_POSITION" default="N,S,c,0.00,0.00"/>
            <visualProperty name="NODE_DEPTH" default="0.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_8" default="50.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_5"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_2"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_8"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_9"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_BORDER_WIDTH" default="0.0">
                <discreteMapping attributeType="string" attributeName="Regulation">
                    <discreteMappingEntry value="5.0" attributeValue="rna up/ptn
down"/>
                    <discreteMappingEntry value="5.0" attributeValue="rna up/ptn
up"/>
                </discreteMapping>
            </visualProperty>
            <visualProperty name="NODE_CUSTOMGRAPHICS_9"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_3" default="50.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_2"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_TOOLTIP" default=""/>
            <visualProperty name="NODE_SELECTED" default="false"/>
            <visualProperty name="NODE_X_LOCATION" default="0.0"/>
            <visualProperty name="NODE_CUSTOMPAINT_6"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_6, name=Node Custom
Paint 6)"/>
            <visualProperty name="NODE_CUSTOMPAINT_2"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_2, name=Node Custom
Paint 2)"/>
            <visualProperty name="NODE_SIZE" default="25.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_9" default="50.0"/>
            <visualProperty name="NODE_PAINT" default="#787878"/>
```

```
            <visualProperty name="NODE_CUSTOMPAINT_7"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_7, name=Node Custom
Paint 7)"/>
            <visualProperty name="NODE_VISIBLE" default="true"/>
            <visualProperty name="NODE_CUSTOMPAINT_4"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_4, name=Node Custom
Paint 4)"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_3"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_SELECTED_PAINT" default="#FFFF00"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_5" default="50.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_1"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_7"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_3"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_LABEL_FONT_SIZE" default="10"/>
            <visualProperty name="NODE_HEIGHT" default="30.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_6"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_LABEL_WIDTH" default="200.0"/>
            <visualProperty name="NODE_CUSTOMPAINT_1"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_1, name=Node Custom
Paint 1)"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_5"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_BORDER_TRANSPARENCY" default="255"/>
            <visualProperty name="NODE_Z_LOCATION" default="0.0"/>
            <visualProperty name="NODE_CUSTOMPAINT_5"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_5, name=Node Custom
Paint 5)"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_6" default="50.0"/>
            <visualProperty name="COMPOUND_NODE_SHAPE" default="ROUND_RECTANGLE"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_6"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_1"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_SHAPE" default="ELLIPSE">
                <discreteMapping attributeType="string" attributeName="Regulation">
                    <discreteMappingEntry value="DIAMOND" attributeValue="rna
up/ptn down"/>
                    <discreteMappingEntry value="TRIANGLE" attributeValue="rna
up/ptn up"/>
                </discreteMapping>
            </visualProperty>
            <visualProperty name="NODE_LABEL_TRANSPARENCY" default="255"/>
            <visualProperty name="NODE_LABEL_COLOR" default="#333333"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_7"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_LABEL_FONT_FACE"
default="Dialog.plain,plain,12"/>
            <visualProperty name="NODE_Y_LOCATION" default="0.0"/>
            <visualProperty name="NODE_TRANSPARENCY" default="255"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_1" default="50.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_8"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_WIDTH" default="70.0"/>
            <visualProperty name="NODE_FILL_COLOR" default="#CCCCCC">
                <passthroughMapping attributeType="string"
attributeName="ec_number_color"/>
```

```
            </visualProperty>
            <visualProperty name="COMPOUND_NODE_PADDING" default="10.0"/>
            <visualProperty name="NODE_CUSTOMPAINT_3"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_3, name=Node Custom
Paint 3)"/>
            <visualProperty name="NODE_BORDER_PAINT" default="#000000"/>
            <visualProperty name="NODE_BORDER_STROKE" default="SOLID"/>
            <visualProperty name="NODE_CUSTOMPAINT_9"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_9, name=Node Custom
Paint 9)"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_4" default="50.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_7" default="50.0"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_4"
default="org.cytoscape.ding.customgraphics.NullCustomGraphics,0,[ Remove
Graphics ],"/>
            <visualProperty name="NODE_LABEL" default="">
                <passthroughMapping attributeType="string"
attributeName="ec_number"/>
            </visualProperty>
            <visualProperty name="NODE_CUSTOMGRAPHICS_POSITION_4"
default="C,C,c,0.00,0.00"/>
            <visualProperty name="NODE_NESTED_NETWORK_IMAGE_VISIBLE"
default="true"/>
            <visualProperty name="NODE_CUSTOMPAINT_8"
default="DefaultVisualizableVisualProperty(id=NODE_CUSTOMPAINT_8, name=Node Custom
Paint 8)"/>
            <visualProperty name="NODE_CUSTOMGRAPHICS_SIZE_2" default="50.0"/>
        </node>
        <edge>
            <dependency name="arrowColorMatchesEdge" value="false"/>
            <visualProperty name="EDGE_LABEL_FONT_FACE"
default="Dialog.plain,plain,10"/>
            <visualProperty name="EDGE_TARGET_ARROW_UNSELECTED_PAINT"
default="#000000"/>
            <visualProperty name="EDGE_SOURCE_ARROW_SELECTED_PAINT"
default="#FFFF00"/>
            <visualProperty name="EDGE_TOOLTIP" default=""/>
            <visualProperty name="EDGE_TARGET_ARROW_SELECTED_PAINT"
default="#FFFF00"/>
            <visualProperty name="EDGE_LABEL" default=""/>
            <visualProperty name="EDGE_SOURCE_ARROW_SHAPE" default="NONE"/>
            <visualProperty name="EDGE_TRANSPARENCY" default="200"/>
            <visualProperty name="EDGE_BEND" default=""/>
            <visualProperty name="EDGE_SELECTED_PAINT" default="#FF0000"/>
            <visualProperty name="EDGE_WIDTH" default="1.0"/>
            <visualProperty name="EDGE_SOURCE_ARROW_UNSELECTED_PAINT"
default="#000000"/>
            <visualProperty name="EDGE_UNSELECTED_PAINT" default="#404040"/>
            <visualProperty name="EDGE_SELECTED" default="false"/>
            <visualProperty name="EDGE_CURVED" default="true"/>
            <visualProperty name="EDGE_STROKE_SELECTED_PAINT" default="#FF0000"/>
            <visualProperty name="EDGE_LINE_TYPE" default="SOLID">
                <discreteMapping attributeType="string"
attributeName="interaction">
                    <discreteMappingEntry value="SOLID" attributeValue="pp"/>
                    <discreteMappingEntry value="LONG_DASH" attributeValue="pd"/>
                </discreteMapping>
            </visualProperty>
            <visualProperty name="EDGE_VISIBLE" default="true"/>
            <visualProperty name="EDGE_TARGET_ARROW_SHAPE" default="NONE"/>
            <visualProperty name="EDGE_PAINT" default="#323232"/>
            <visualProperty name="EDGE_LABEL_FONT_SIZE" default="10"/>
            <visualProperty name="EDGE_LABEL_WIDTH" default="200.0"/>
            <visualProperty name="EDGE_LABEL_TRANSPARENCY" default="255"/>
            <visualProperty name="EDGE_STROKE_UNSELECTED_PAINT" default="#CCCCCC">
```

```
                <discreteMapping attributeType="string" attributeName="Actions">
                    <discreteMappingEntry value="#3130E4" attributeValue="binding
catalysis reaction"/>
                    <discreteMappingEntry value="#00D2D8" attributeValue="binding
reaction"/>
                    <discreteMappingEntry value="#00CC00" attributeValue="binding
expression"/>
                    <discreteMappingEntry value="#FFAA00"
attributeValue="binding"/>
                    <discreteMappingEntry value="#006666"
attributeValue="activation binding expression"/>
                    <discreteMappingEntry value="#990099"
attributeValue="activation binding ptmod"/>
                    <discreteMappingEntry value="#FF9AC3" attributeValue="binding
catalysis ptmod reaction"/>
                </discreteMapping>
            </visualProperty>
            <visualProperty name="EDGE_LABEL_COLOR" default="#333333"/>
        </edge>
    </visualStyle>
</vizmap>
```

ANEXO

ANEXO 1 – Artigo publicado, que utilizou parte da metodologia construída neste trabalho. O artigo, na íntegra, encontra-se nas páginas seguintes. Ele também está disponível em http://www.mdpi.com/1422-0067/18/2/371.

Immunoinformatics Features Linked to Leishmania Vaccine Development: Data Integration of Experimental and In Silico Studies

Rory C. F. Brito, Frederico G. Guimarães, João P. L. Velloso, Rodrigo Corrêa-Oliveira, Jeronimo C. Ruiz, Alexandre B. Reis and Daniela M. Resende

(This article belongs to the Special Issue Reverse Vaccinology)

Abstract

Leishmaniasis is a wide-spectrum disease caused by parasites from *Leishmania* genus. There is no human vaccine available and it is considered by many studies as apotential effective tool for disease control. To discover novel antigens, computational programs have been used in reverse vaccinology strategies. In this work, we developed a validation antigen approach that integrates prediction of B and T cell epitopes, analysis of Protein-Protein Interaction (PPI) networks and metabolic pathways. We selected twenty candidate proteins from *Leishmania* tested in murine model, with experimental outcome published in the literature. The predictions for CD4+ and CD8+ T cell epitopes were correlated with protection in experimental outcomes. We also mapped immunogenic proteins on PPI networks in order to find Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with them. Our results suggest that non-protective antigens have lowest frequency of predicted T CD4+ and T CD8+ epitopes, compared with protective ones. T CD4+ and T CD8+ cells are more related to leishmaniasis protection in experimental outcomes than B cell predicted epitopes. Considering KEGG analysis, the proteins considered protective are connected to nodes with few pathways, including those associated with ribosome biosynthesis and purine metabolism.

Keywords: immunoinformatics; epitope prediction; pathways; protein–protein interaction networks; reverse vaccinology; leishmaniasis

International Journal of
Molecular Sciences

MDPI

Article

Immunoinformatics Features Linked to *Leishmania* Vaccine Development: Data Integration of Experimental and In Silico Studies

Rory C. F. Brito [1,2], Frederico G. Guimarães [3], João P. L. Velloso [3], Rodrigo Corrêa-Oliveira [4,5], Jeronimo C. Ruiz [3,6], Alexandre B. Reis [1,2,5,*] and Daniela M. Resende [3,6]

1 Laboratório de Pesquisas Clínicas, Programa de Pós-graduação em Ciências Farmacêuticas/CiPharma, Escola de Farmácia, Campus Morro do Cruzeiro, Universidade Federal de Ouro Preto, Bauxita, 35.400-000 Ouro Preto, Minas Gerais, Brazil; rorybrito@gmail.com

2 Laboratório de Imunopatologia, Núcleo de Pesquisas em Ciências Biológicas, Campus Morro do Cruzeiro, Universidade Federal de Ouro Preto, Bauxita, 35.400-000 Ouro Preto, Minas Gerais, Brazil

3 Grupo Informática de Biossistemas e Genômica, Programa de Pós-graduação em Ciências da Saúde, Centro de Pesquisas René Rachou, Fiocruz Minas, Av. Augusto de Lima, 1715, Barro Preto, 30.190-002 Belo Horizonte, Minas Gerais, Brazil; frederico.guimaraes@cpqrr.fiocruz.br (F.G.G.); jpvlinhares@gmail.com (J.P.L.V.); jeronimo@cpqrr.fiocruz.br (J.C.R.); dmresende@cpqrr.fiocruz.br (D.M.R.)

4 Grupo Imunologia Celular e Molecular, Programa de Pós-graduação em Ciências da Saúde, Centro de Pesquisas René Rachou, Fiocruz Minas, Av. Augusto de Lima, 1715, Barro Preto, 30.190-002 Belo Horizonte, Minas Gerais, Brazil; correa@cpqrr.fiocruz.br

5 Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais (INCT-DT), Campus Morro do Cruzeiro, Universidade Federal de Ouro Preto, Bauxita, 35.400-000 Ouro Preto, Minas Gerais, Brazil

6 Programa de Pós-graduação em Biologia Computacional e Sistemas, Instituto Oswaldo Cruz, Fiocruz, Av. Brasil, 4.365, Pavilhão Arthur Neiva, Manguinhos, 21.040-360 Rio de Janeiro, Rio de Janeiro, Brazil

* Correspondence: alexreis@nupeb.ufop.br; Tel.: +55-31-3559-1694

Academic Editor: Christopher Woelk
Received: 28 December 2016; Accepted: 3 February 2017; Published: 10 February 2017

Abstract: Leishmaniasis is a wide-spectrum disease caused by parasites from *Leishmania* genus. There is no human vaccine available and it is considered by many studies as apotential effective tool for disease control. To discover novel antigens, computational programs have been used in reverse vaccinology strategies. In this work, we developed a validation antigen approach that integrates prediction of B and T cell epitopes, analysis of Protein-Protein Interaction (PPI) networks and metabolic pathways. We selected twenty candidate proteins from *Leishmania* tested in murine model, with experimental outcome published in the literature. The predictions for $CD4^+$ and $CD8^+$ T cell epitopes were correlated with protection in experimental outcomes. We also mapped immunogenic proteins on PPI networks in order to find Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with them. Our results suggest that non-protective antigens have lowest frequency of predicted T $CD4^+$ and T $CD8^+$ epitopes, compared with protective ones. T $CD4^+$ and T $CD8^+$ cells are more related to leishmaniasis protection in experimental outcomes than B cell predicted epitopes. Considering KEGG analysis, the proteins considered protective are connected to nodes with few pathways, including those associated with ribosome biosynthesis and purine metabolism.

Keywords: immunoinformatics; epitope prediction; pathways; protein–protein interaction networks; reverse vaccinology; leishmaniasis

1. Introduction

Leishmaniasis is a wide-spectrum disease caused by parasites from *Leishmania* genus. It is prevalent in Americas, Europe, Africa and Asia. Overall, human infection is caused by at least 20

species whose vectors are phlebotomine sandflies [1]. Although being considered by many studies one of the best possible alternatives for this disease control, there is no human vaccine available [2].

In the advent of reverse vaccinology, in the latest years, a great effort has been made by bioinformaticians in order to provide epitopes predictors programs. Currently, it is possible to scan entire genomes searching for immunogenic epitopes and then select promising proteins for vaccine development. The bottleneck in this workflow analysis is the validation of predictions for protozoan parasites. Many predictors are available for B cells, T $CD4^+$ cells and T $CD8^+$ cells epitopes and subcellular localization. They are valuable in a pre-screening evaluation for vaccine targets and searching for diagnostic markers.

The building of protein-protein interaction (PPI) networks may give some insights to understand the biological role of these targets, and so might be a valuable asset in vaccine development. These networks are constituted by nodes that correspond to proteins, connected by edges, representing the interactions between two connected proteins. With PPI networks, we can have an overview of protein relationships and notice those with high connections (also referred as "hubs"). Hub proteins tend to have essential role in the parasite metabolism and might be good candidates to vaccinal and drug target [3,4].

To support *Leishmania* vaccine research, we developed an approach that integrates prediction of B and T cell epitopes, analysis of PPI networks and metabolic pathways. With the aim of validating this methodology, we selected *Leishmania* proteins tested as vaccine candidates in murine model, with experimental outcome (EO) published in the literature. After predicting epitopes in the selected proteins using specific computational programs, we correlated the predictions for T $CD4^+$ and T $CD8^+$ cells with protection in EO. Finally, we mapped the immunogenic proteins on PPI networks in order to find Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways associated with them.

2. Results

2.1. Leishmania Proteins Dataset Selection

Through the use of text mining technics from Pubmed website that included, but was not restricted to, categorization and entity extraction, we were able to identify and select 20 proteins from six different *Leishmania* species that were used in studies aiming the vaccine development against these parasites.

It is important to highlight that, for each one of those proteins, a specific MySQL ID was assigned to link GI accession number and TriTrypDB specific ID. Based on the results published, the EO was categorized into: (a) "no protection" (nine proteins); (b) "partial protection" (five proteins); and (c) "protection" (six proteins). The accession numbers of these proteins are depicted in Table 1.

2.2. Epitope and Subcellular Localization Predictions

With the purpose of selecting potential immunogenic epitopes in the selected experimental dataset, Structured Query Language (SQL) statements were used. The results obtained in terms of number of predicted binding Major Histocompatibility Complex (MHC) class I and II epitopes, B cell epitopes and subcellular locations are detailed in Table 2. Interestingly, the majority of the proteins within the "protection" group were predicted as extracellular and the proteins belonging to "no protection" group were predicted as located in nuclear and cytoplasmic compartments.

Table 1. Selected candidate antigens from dermatotropic and visceratropic *Leishmania* species to leishmaniasis vaccine development. Proteins in literature tested in mice model were selected randomly.

Leishmania Tropism	Geographic Area	Specie	Candidate Antigen	Function	MySQL ID	NCBI Sequence Acession	Animal	Experimental Outcome Indicative	Reference
Dermatotropic *Leishmania* species	New world	*L. brazilinensis*	Thiol-specific-antioxidant (TSA)	Tryparedoxin peroxidase	LbrM.15.1080	gi 154334618	mice	No protection	[5]
			LeiF	*Leishmania* putative eukaryotic initiation factor	LbrM.25.0580	gi 154338682	mice	No protection	[5]
			LACK	*Leishmania* homolog of receptors for activated C-kinase	LbrM.28.2950	gi 154340729	mice	Partial protection	[5]
		L. amazonensis	P4 nuclease partial	Endonuclease activity	A16600	gi 29165287	mice	No protection	[6]
			Cysteine proteinase	Cysteine-type peptidase activity	A22180	gi 30142572	mice	Partial protection	[7]
			HSP20	Heat shock protein	A38570	gi 513044555	mice	No protection	[8]
			GP46	Membrane glycoprotein	A64110	gi 159321	mice	Protection	[9]
		L. mexicana	GP63	Metalloendopeptidase activity	LmxM.10.0465	gi 401416782	mice	Protection	[10]
	Old world	*L. major*	LmTSI	Stress-induced protein sti1	LmjF.08.1110	gi 68124434	mice	Protection	[11]
			GP63	Metalloendopeptidase activity	LmjF.10.0470	gi 157865341	mice	Protection	[12]
			PSA 2	Promastigote surface antigen protein 2	LmjF.12.1000	gi 68124979	mice	No protection	[13]
			TSA	Thiol-specific-antioxidant—Tryparedoxin peroxidase	LmjF.15.1080	gi 68125473	mice	No protection	[14]
			Histone H1	DNA binding	LmjF.27.1190	gi 4008565	mice	No protection	[15]
			LACK	*Leishmania* homolog of receptors for activated C-kinase	LmjF.28.2740	gi 157872022	mice	Partial protection	[16]
Viscerotropic *Leishmania* species	New world	*L. infantum*	H2A	DNA binding	LinJ.21.1160	gi 339898105	mice	No protection	[17]
			LiCY1	Peptidylprolyl isomerase	LinJ.25.0940	gi 146088699	mice	Partial protection	[18]
			Histone H1	DNA binding	LinJ.27.1070	gi 78146500	mice	No protection	[19]
			CPC	Cysteine-type peptidase activity	LinJ.29.0860	gi 146092987	mice	Protection	[20]
	Old world	*L. donovani*	NH36	Hydrolase activity	LdBPK_181570.1	gi 19697561	mice	Partial protection	[21]
			A2	Amastigote-specific protein—stress response protein	LdBPK_220560.1	gi 12382244	mice	Protection	[22]

Table 2. Number of binding Major Histocompatibility Complex (MHC) epitopes, B cell epitopes and subcellular localization predicted by different computational programs.

MySQL ID	Prediction of Binding MHC Epitopes			Prediction of B Cells Epitopes			EO [1]	Prediction of Subcelular Localization
	Binding MHC Class I Epitopes		Binding MHC Class II Epitopes	AAP12	BCPred12	BepiPred		
	NetMHC	NetCTL	NetMHCII					
LbrM.15.1080	7	74	132	97	32	1	No protection	cyt
LbrM.25.0580	6	31	121	75	23	2	No protection	cyt
LbrM.28.2950	14	67	298	130	15	2	Partial protection	nuc
A16600	4	18	63	21	9	2	No protection	cyt
A22180	15	105	403	214	65	14	Partial protection	ext
A38570	10	46	146	52	0	5	No protection	ext
A64110	28	149	739	193	20	10	Protection	ext
LmxM.10.0465	31	196	747	302	91	36	Protection	ext
LmjF.08.1110	29	177	369	291	79	16	Protection	cyt
LmjF.10.0470	27	177	668	317	52	19	Protection	pla
LmjF.12.1000	23	100	475	226	102	12	No protection	ext
LmjF.15.1080	9	77	199	81	35	5	No protection	cyt
LmjF.27.1190	1	27	89	20	20	2	No protection	nuc
LmjF.28.2740	16	64	301	172	18	7	Partial protection	nuc
LinJ.21.1160	5	52	130	33	30	2	No protection	nuc
LinJ.25.0940	8	26	116	98	69	2	Partial protection	cyt
LinJ.27.1070	1	36	53	80	58	2	No protection	nuc
LinJ.29.0860	21	99	331	201	85	4	Protection	ext
LdBPK_181570.1	14	89	356	161	63	2	Partial protection	ext
LdBPK_220560.1	35	159	669	200	165	12	Protection	pla

[1] EO = Experimental outcome.

Specifically regarding the epitopes capacity to bind MHC class I, MHC class II and epitopes for B cell activation, considering a path from "no protection" to "protection" groups, a gradual increase of the number of predicted epitopes for T cells and B cells was observed.

2.3. Predicted Epitopes and Experimental Outcome Correlation

To evaluate the possible association between the number of predicted epitopes (NPE) for B and T cells, and the EO of selected proteins, the following consensus predictions were produced: (a) prediction for T $CD8^+$ epitopes obtained from NetMHC and NetCTL; and (b) prediction for B cell epitopes obtained from AAP12, BCPred12 and BepiPred. The consensus predictions were obtained overlapping identical predictions made by different methodologies.

To graphically depict the results, Box Plots and Correspondence Maps (CM) approaches were applied to visualize the potential associations determined through Spearman r and Chi-square distance, respectively. Firstly, the disease (leishmaniasis) was stratified into cutaneous leishmaniasis (CL) and visceral leishmaniasis (VL) and the EO of antigens from *Leishmania* that cause CL and VL were correlated with NPE (Figure 1a). Regarding VL analyses, significant correlation was observed only with EO and predicted epitopes for $CD8^+$ T cells ($p < 0.05$) (data not shown). On the other hand, for CL analyses, it was observed significant correlation between EO and NPE for T $CD4^+$, T $CD8^+$ and B cells, as shown in Figure 1a. After that, analyses were performed concerning the disease without any stratification. As can be observed from Figure 1b (NPE and EO correlation for T $CD4^+$, T $CD8^+$ and B cells), a significant correlation exists between NPE specific to $CD4^+$ and $CD8^+$ T cells with $r = 0.752/p < 0.05$ and $r = 0.793/p < 0.05$. In addition, a weak association with B cell predicted epitopes ($r = 0.515/p < 0.05$) was observed. In other words, non-protective antigens have lowest frequency of predicted T $CD4^+$ and T $CD8^+$ epitopes, compared with protective ones.In regards to CM analysis (Figure 1c), considering the adopted variables (antigens EO versus NPE for T and B cells), the grouping outcome, which is related with data correlation, shows the same strong association above mentioned for leishmaniasis with no stratification.

As the last analysis layer used to validate data correlation, the Chi-square results confirmed the significant association between EO and the predicted epitopes for T $CD4^+$ and T $CD8^+$, $p < 0.05$ (Tables S1 and S2) and the weak one between EO and predicted epitopes for B cells ($p < 0.05$, see Table S3).

2.4. Number of Alleles (NA) and Experimental Outcome (EO) Correlation

To hypothesize possible reasons linked with vaccine success or failure, an evaluation of allele-specific affinity (MHC I and II) was investigated. As illustrated in Figure 2a, the amounts of epitopes binding MHC haplotype *d* (BALB/c MHC alleles) and haplotype *b* binders (C57BL/6 MHC alleles) identified in the "protection" group were superior to those ones identified in the "no protection" group. In summary, our results indicate that epitopes for MHC class I and II haplotype *b* and *d* are more frequent in the success antigens tested for vaccine development. In addition, a detailed analysis in which MHC class I and II haplotypes were individually investigated revealed a strong association between NPE from MHC class I haplotype *d* and EO ($p < 0.05$ and $r = 0.855$) that is not observed for MHC class II haplotype *d* and *b*.

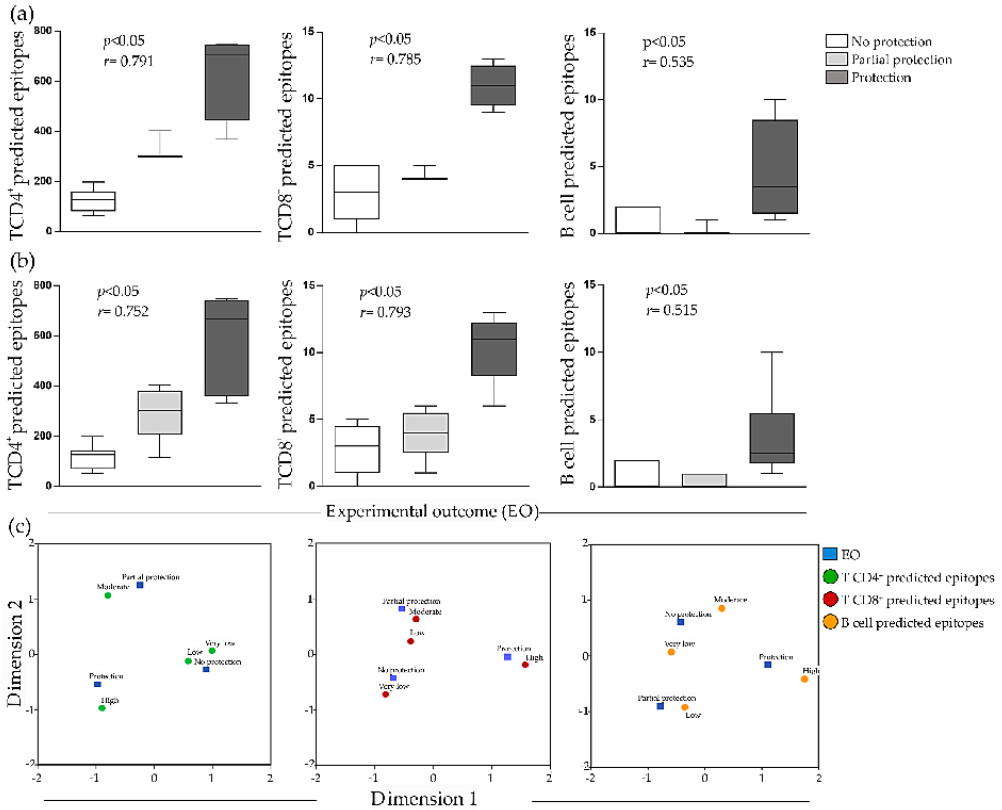


Figure 1. Correlation analysis: (**a**) Box plots of the relationships between T CD4$^+$, T CD8$^+$ and B cell epitopes and experimental outcome of candidate antigens taking into account cutaneous leishmaniasis (CL); (**b**) box plots of the relationships between T CD4$^+$, T CD8$^+$ and B cell epitopes and experimental outcome of candidate antigens concerning leishmaniasis diseases with no stratification; and (**c**) correspondence map showing the association between experimental outcome and T CD4$^+$, T CD8$^+$ and B cell predicted epitopes for leishmaniasis with no stratification.

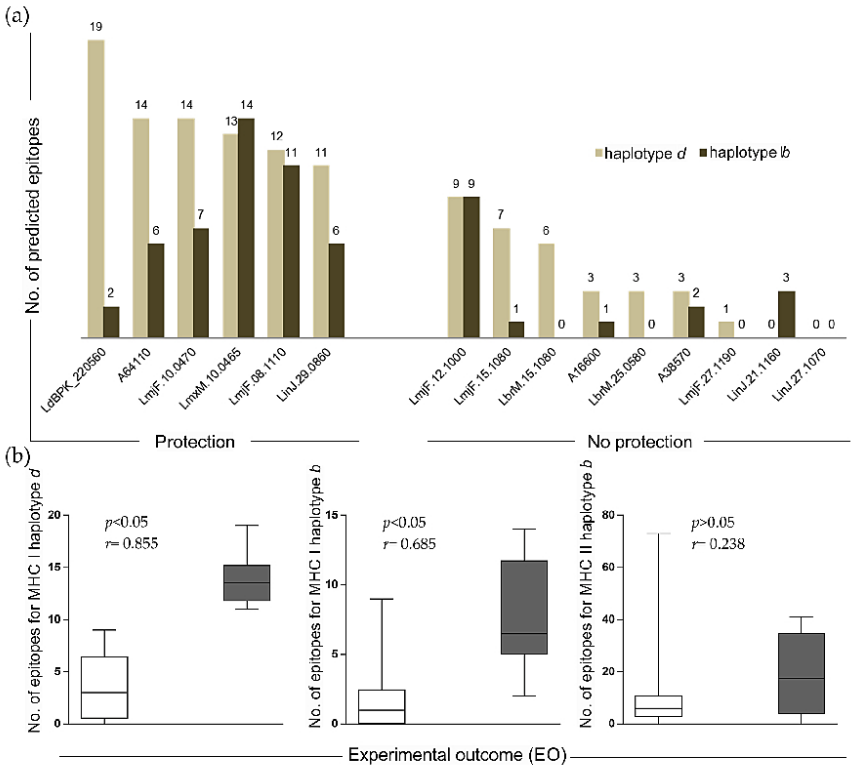


Figure 2. Evaluation of epitopes amount and relationship with experimental outcome: (**a**) bar graph showing number of epitopes for MHC class I and II (haplotype *d* and *b*) in the selected antigens classified in "protection" and "no protection" groups; and (**b**) box plot of the relationships between $CD4^+$ and $CD8^+$ T cell epitopes and experimental outcome of candidate antigens.

2.5. Mapping Immunogenic Proteins on Protein-Protein Interaction Networks (PPI Networks)

We chose Cytoscape to model PPI networks using data from STRING v.10. Figure 3a presents these networks, annotated according to their related KEGG pathways. Analyses of enriched pathways are shown in Figure 3b, according to their False Discovery Rate. Considering the "no protection" group, the most common pathways found were: ribosome (58 genes involved), glutathione metabolism (17 genes), RNA degradation (15 genes), protein processing in endoplasmic reticulum (15 genes), peroxisome (sixgenes) and homologous recombination (sevengenes). On the other hand, for "protection" group, we identified pathways as ribosome (195 genes involved), metabolic pathways (12 genes), purine metabolism (sixgenes) and protein processing in endoplasmic reticulum (seven genes). Interestingly, target proteins from "no protection" group are connected with nodes from many different pathways. In contrast, we observed that proteins of the "protection" group are connected to nodes of few pathways. Thus, there is a strong negative correlation ($r = -8.55$) between the number of connected pathways and EO of the selected proteins (Figure 3c).

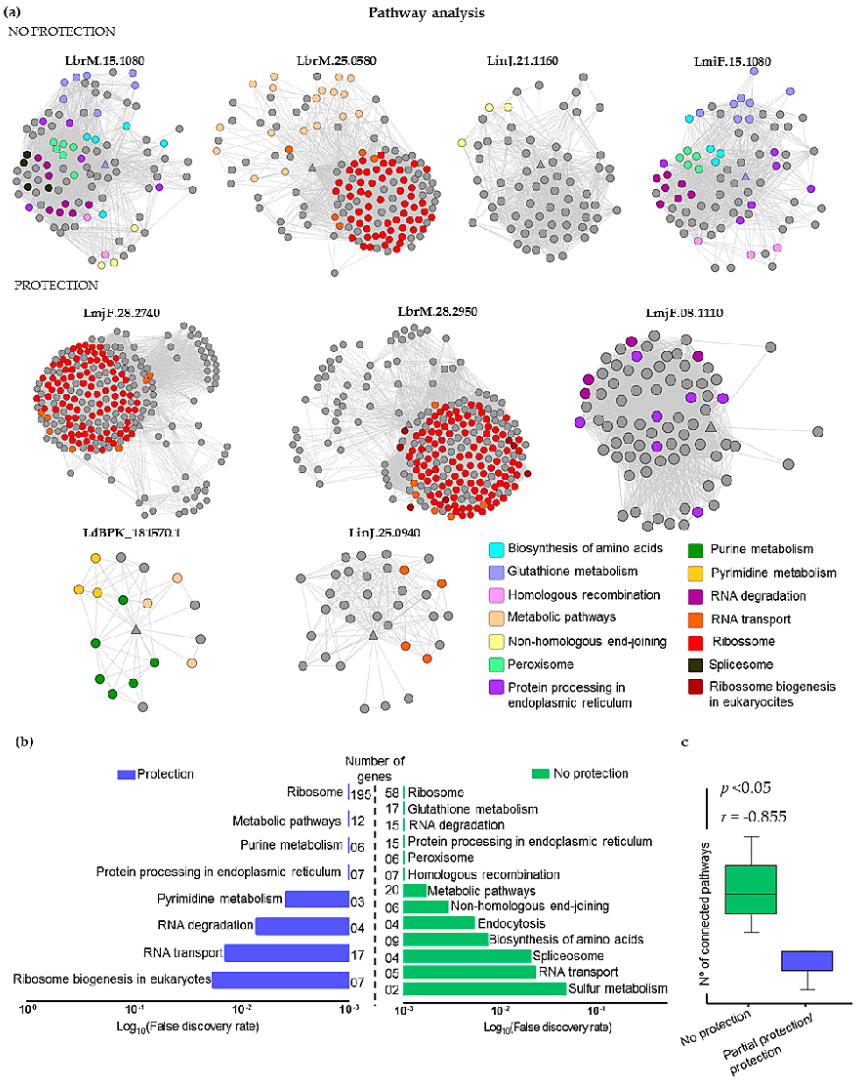


Figure 3. Immunogenic proteins mapped in Protein-Protein Interaction (PPI) networks: (**a**) PPI networks constructed starting with the target proteins ("no protection" and "partial protection/protection") represented by triangles and specific pathways associated with each node (circles) using KEGG database. (**b**) Analysis of KEGG enriched pathways was performed by False Discovery Rate. For both "no protection" and "partial protection/protection", the bar shows the fold-enrichment of the pathways. (**c**) Significant correlation between number of pathways connected with the target proteins of "no protection" and "partial protection/protection" groups ($p = 0.007$).

3. Discussion

Nowadays, many computational methodologies have been described for epitope predictions of bacterial, fungal and others microorganisms. For protozoan (specifically *Leishmania* species), there are not strong and validated platforms to identify promising antigens for *Leishmania* vaccines [23]. Herein, we developed a sturdy and complete platform with potential of identifying candidates for vaccines against leishmaniasis. This platform integrates prediction of B and T cell epitopes, analysis of PPI networks and signaling pathways.

The first step was to select the input antigens to validate the platform. In this concern, we did an extensive search in the literature regarding antigens tested in murine model. It was difficult to categorize the 20 selected antigens for this work because there is a lack of consensus concerning the model, challenge inoculum, mechanisms of protective immunity (response induction, parasite burden reduction), so there is no standardization to appoint if a vaccine indicates protection or not [24]. We tried to choose proteins with no interference of adjuvant since it can entirely modify the antigen response [25].

Regarding the platform validation, we performed the epitope mapping using ad-hoc algorithms. Our results suggest that antigens with more predicted epitopes for T $CD4^+$ and T $CD8^+$ cells could be associated with protection in EO. In this regard, our results revealed that there are strong correlation and association between predicted epitopes and the EO. The inertia values show the powerful association between the variables. The CM dimension 1 represents the highest inertia allowing the interpretation of the results in the first dimension. T $CD4^+$ and T $CD8^+$ predicted epitopes versus EO show higher inertia values when compared to B cells predicted epitopes emphasizing that epitopes for $CD4^+$ and $CD8^+$ T cells are crucial for *Leishmania* vaccines success. We used all available human and mouse alleles to restrict the epitopes allowing enhance the assertive prediction. This analysis is useful to identify conserved epitopes that can bind various alleles of MHC appointing for rare and promising epitopes. The in silico analyses and in vivo validation of epitopes demonstrates that some algorithms may be important tools for the identification of epitopes, and consequently of immunogenic proteins. The algorithm NetCTL version 1.2 makes prediction of peptide–MHC class I binding, proteasomal C terminal cleavage, both using artificial neural networks, and TAP transport efficiency using weight matrix. The tree predictions are then integrated [26]. Another predictor also used for MHC class I binding peptides was NetMHC version 3.0. It predicts binding of peptides to different HLA alleles using artificial neural networks and weight matrices. For peptide–MHC class II binding prediction NetMHCII, version 1.0, was used. It predicts binding of peptides to 14 different HLA-DR alleles, including human and mouse, using position specific weight matrices [27]. To perform B-cell epitopes predictions, we used only methods that predict continuous epitopes. We used first BepiPred, version 1.0, that predicts linear B-cell epitopes using a combination of Hidden Markov model and a propensity scale method [28]. Then we used BCPREDS server comprising the AAP12 and BCPred12 predictors. The first one is based on the finding that B-cell epitopes favor particular amino acid pair, and it was trained using support vector machine classifier. The second uses subsequence kernel trained using support vector machine classifiers with 701 linear B-cell epitopes, extracted from Bcipep database, and 701 non-epitopes, randomly extracted from SwissProt sequences [29–31]. Finally, we used WoLF PSORT predictor, which is an amino acid sequence predictor of subcellular localization sites of proteins. It uses known sorting signal motifs and some correlative sequence features [32]. The integration of these predictors could reveal proteins that are secreted or presented in parasite membrane, capable of eliciting B and T cells responses.

Herrera-Najera et al. [33] performed a large-scale prediction of T cell epitopes in the whole genome of *L. major*, obtaining 26 potential epitopes through prediction consensus. Fourteen of them revealed to be immunogenic epitopes that were capable to stimulate T cells to produce IFN-γ. Other studies employing computational predictions in specific *Leishmania* proteins have shown that it is quite possible to use combined algorithms in epitopes searching that could be validated by in vivo experiments [34]. Duarte et al. [35] developed a combined epitope prediction platform in order to investigate T $CD8^+$

epitopes in 63 *L. braziliensis* proteins, demonstrating a cytotoxic activity of some predicted epitopes in *Leishmania* infected mice. Recently, in silico methods for linear epitope predictions (NetMHC, NetCTL, and NetMHCII) were combined with molecular modeling to identify potential epitopes in the whole *L. braziliensis* proteome. Therefore, the pipeline was validated based on stimulation of human peripheral blood mononuclear cells (PBMCs) proliferation. The results obtained after the in vitro assays showed that six of ten selected epitopes could be classified as potentially immunogenic [36].

The role of B cell epitopes still hasroom for discussion since there is no consensus if immunoglobulins could be associated with resistance or susceptibility in leishmaniasis [37,38]. On the other hand, our results demonstrate a correlation between protection and specific B cell epitopes. In this context, we suggest that an effective vaccine should have epitopes capable of eliciting a strong T cell response and B cells too. In addition to the epitopes for B cells and MHC class I and II, another important feature is the subcellular localization of the antigen. It is known that extracellular *Leishmania* proteins are more immunogenic and considered better targets for vaccine development [39,40]. This fact is indeed corroborated by our findings that show the majority of the antigens from "protection group" are linked with extracellular compartmentalization. In silico approaches have limitations regarding the proteome annotation (e.g., the data of *L. amazonensis* used in this work) and the large number of linear epitopes. Nevertheless, our results of epitope prediction indicate a higher assertive and successful prediction, so it can be a useful approach for vaccine development against leishmaniasis.

To better understand the biological importance of vaccine candidates, we proposed the use of PPI networks enriched with KEGG pathways information. It is well known that some proteins are essential for specific biological processes of *Leishmania* spp. [41,42]. In this context, we proposed to analyze antigen pathways through modeled PPI and its relation with protection and no protection of vaccine candidates after challenging with infective *Leishmania* parasites. Our analyses showed that many of the selected antigens do not have any KEGG pathway associated to them, but, instead, are connected to proteins that are part of some pathway. Pathways associated with ribosome biosynthesis, purine metabolism and metabolic processes are present in "protection" group networks. Ribosome related proteins were considered relevant molecules during infection, since in some circumstances they can modulate cell activities and cytokine release. Many works associated these pathways to immune response [43]. Cordeiro-Da-Silva and collaborators in 2001 characterized a *Leishmania major* gene considered to be homologous to the mammalian ribosomal protein S3a. This ribosomal protein can be found in many other *Leishmania* species such as *L. infantum*, *L. amazonensis*, and *L. mexicana*. The article authors suggested that this protein could participate in the Th1/Th2 immune response balance during leishmaniasis [44]. Soto and collaborators in 1993 using sera from dogs affected by visceral leishmaniasis identified high antigenic *Leishmania* acidic ribosomal proteins, also called P-type proteins [45]. Another work by Soto and collaborators in 2000 showed that intraperitoneal administration in BALB/c mice of the acidic ribosomal protein LiP2a, without adjuvants, elicited a strong humoral response and was capable of stimulating production of IFN-γ in cultured splenocytes from LiP2a-immunized mice [46]. Our findings also match with results obtained in the secretome of *L. donovani*, where the majority of virulent proteins (secreted proteins) belong mainly to metabolic and biosynthesis processes [47]. To check the importance of some proteins related to metabolic process, *Leishmania* knocked-out for protein kinases and phosphatases possible involved in parasite metabolism regulation were generated. After this process, in many cases highly attenuated or completely avirulent parasites could be observed [48]. Naderer and collaborators generated a *Leishmania major* mutant lacking the regulatory subunit of the Ca^{2+}/calmodulin-dependent serine/threonine-specific phosphatase. This modified *Leishmania* grew normally at 27 °C. However, this parasite lost viability when exposed to 34 °C [49]. Target of rapamycin (TOR) kinases are involved in some regulatory pathways related to cell growth and structure in eukaryotes. Silva and collaborators generated TOR3 knocked-out *Leishmania major* parasites. These knocked-out parasites exhibited slower growth than wild-type parasites and were unable to survive or replicate in macrophages in vitro. These parasites were not capable of inducing disease or establish infection in mice in vivo [50]. In addition,

McConville and Naderer [48] have shown that metabolic pathways are important to *Leishmania* virulence, since down regulation of metabolic genes causes latency of many *Leishmania* species. One possible application of these attenuated or avirulent parasites could be in whole parasites vaccines. Carter et al. [51] have noticed that purine metabolism is vital for *Leishmania* survival. Surprisingly, our findings show that proteins associated to protection are connected to few pathways when compared to proteins that are not protective.

In summary, in this work, we proposed and validated a computational approach regarding epitope prediction, topological structure and pathway analyses to drive a rational vaccine design against leishmaniasis.

4. Materials and Methods

4.1. Selection of the Leishmania Antigens

Leishmania proteins tested in murine model, with EO published in the literature, were selected. Studies describing vaccine effectiveness after challenge with *Leishmania* spp. were preferentially chosen. Bearing in mind that there is no standardization of the protection concept and that there is variation of results in the literature, it was necessary to create three categories as already described. According to our categorization, "no protection" group includes antigens that promote no or slight reduction of parasite burden or lesions after *Leishmania* challenge. The "protection" group includes antigens that promoted significantly strong reduction of the parasite burden or lesions showing strong immune response to *Leishmania* antigens. Thus, the term partial protection was used to classify antigens that are between a potential protection and no protection at all. The "partial protection" group comprises proteins that could slightly reduce the parasite burden and/or lesions more than "no protection" group. In addition, these proteins can elicit some immune response which results in ineffective protection. It is important to highlight that the influence of adjuvants was not taken into account for the categorization, thus only the response of antigens tested alone was chosen to categorize the groups. Twenty candidates from Old and New World *Leishmania* species were categorized according with their experimental results in the following groups: (a) "no protection"; (b) "partial protection"; and (c) "protection". The selected experimentally validated data included information from proteins of *L. amazonensis*, *L. braziliensis*, *L. major*, *L. mexicana*, *L. donovani* and *L. infantum* that were subsequently used in this study to corroborate the in silico bioinformatics predictions. The selected antigens are described in Table 1.

4.2. Leishmania Proteome Data

The predicted proteome sequences of dermatotropic and viscerotropic *Leishmania* species were obtained from TriTrypDB (Kinetoplastid Genomics Resource) and the sequence of *L. amazonensis* was downloaded from http://bioinfo08.ibi.unicamp.br/leishmania/ [52]. Detailed information about the predicted proteomes versions used in this work can be found in Table 3.

Table 3. *Leishmania* predicted proteomes used in the study. The version and number of predicted proteins of each species are shown.

Leishmania Specie	Version of Proteome	Predicted Proteins
L. braziliensis	3.1	8357
L. amazonensis	- [1]	8168
L. mexicana	9.0	8250
L. major	9.0	8400
L. donovani	8.0	8083
L. infantum	3.2	8241

[1] This is a draft version. This proteome still has many annotation errors.

4.3. Epitope and Subcellular Localization Predictions

All proteomic data used in this work were screened in order to predict T CD4$^+$ and T CD8$^+$ epitopes, B cell epitopes and subcellular localization of proteins.

For T CD8$^+$ epitope prediction (MHC class I binding epitopes), algorithms NetCTL [26,53,54] and NetMHC [55–57] were used. Regarding T CD4$^+$ epitopes, NetMHCII [27,58] was used to predict MHC class II binding epitopes. For B cell epitopes, BepiPred [28] and BCPREDS (AAP12 and BCPred12 models) [29–31] were used to predict epitopes. Finally, the protein subcellular localization was predicted using WoLF PSORT [32,55–57]. The algorithms choice was made taking into account their viability for local stand-alone server installation, the number of citations found in literature, and the results previously published by Resende, Rezende, Oliveira, Batista, Correa-Oliveira, Reis and Ruiz [23] describing algorithms specificity, sensitivity and accuracy with parasite data obtained from UniProt (http://www.uniprot.org/) and IEDB (Immune Epitope Database and Analysis Resourse) (http://www.iedb.org/). The analytical workflow used in this study is presented in Figure 4.

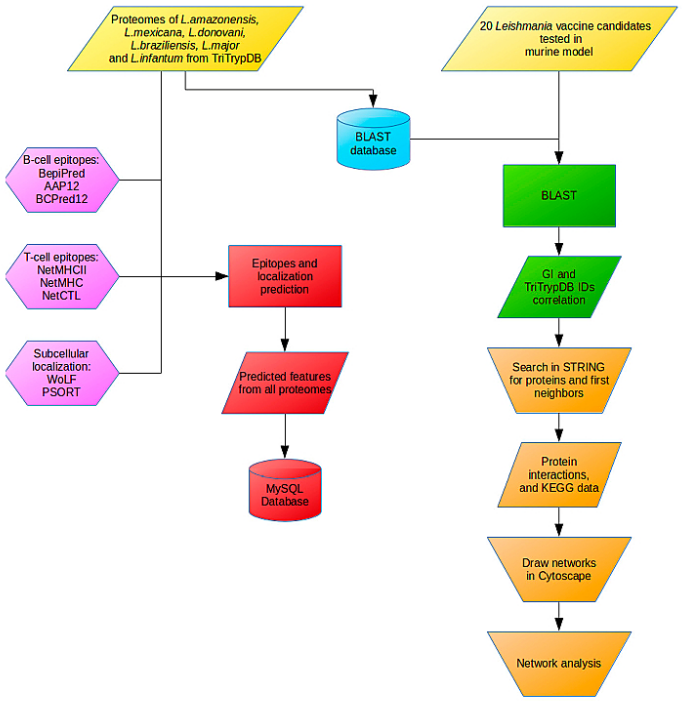


Figure 4. Workflow of analysis showing the steps followed along this work.

To carry out the predictions, the algorithms were parameterized for eukaryotic genomes. For MHC class I binding epitopes, predictions for 12 human supertypes and seven mice alleles were performed. The following alleles were used: A1, A2, A3, A24, A26, B7, B8, B27, B39, B44, B58, B62, H2-Db, Dk-H2, H2-Dd, H2-Kb, H2-Kd, Kk-H2, and H2-Ld. Concerning MHC class II binding epitopes, we used 14 human alleles and three mice alleles bringing the total of different alleles to 17, as follows: HLA-DRB1*01:01, HLA-DRB1*03:01, HLA-DRB1*04:01, HLA-DRB1*04:04, HLA-DRB1*04:05, HLA-DRB1*07:01, HLA-DRB1*08:02, HLA-DRB1*09:01, HLA-DRB1*11:01,

HLA-DRB1*13:02, HLA-DRB1*15:01, HLA-DRB3*01:01, HLA-DRB4*01:01, HLA-DRB5*01:01, H2-IAs, H2-IAd and H2-IAb.

4.4. Development of Relational Database

Taking into account the great amount of data generated by the algorithms, we constructed a relational database using MySQL as Relational Database Management System (RDBMS) (http://www.mysql.com). The use of a database system in this work represents a crucial step that allows the integration of the results from all predictors and a way of getting a data receptacle or conceptual repository from which is possible to extract data correlation, helping in the identification of target proteins. 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 relational model was built in MySQL Workbench (http://wb.mysql.com). To extract, parse and load data into database, specific Perl scripts were developed using DBI (the Perl interface to databases) and BioPerl modules. The relational schema is presented in Figure 5.

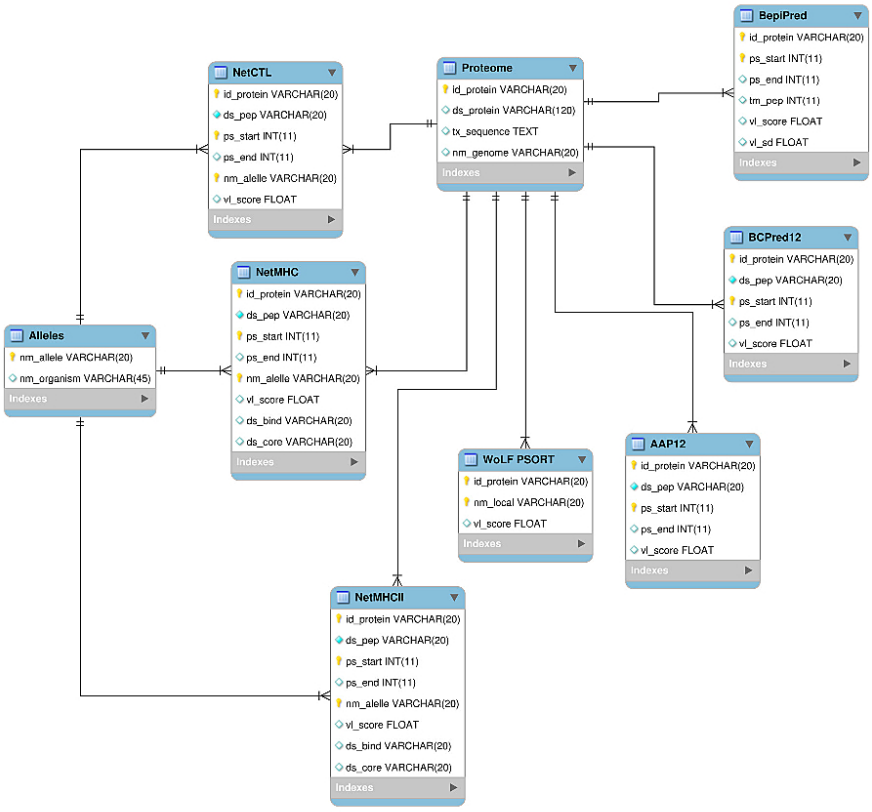


Figure 5. MySQL relational database scheme developed to integrate data from predictions.

4.5. Mapping Immunogenic Proteins on Protein-Protein Interaction Networks (PPI Networks)

In this topic, two groups were created: P (protection), aggregating proteins classified as "partial protection" and "protection", and NP (no protection) for proteins classified as "no protection".

We entered these proteins into STRING v.10 [59] looking for interaction networks associated with them, using *Leishmania* as reference organism. The active interaction sources used were text mining, experiments, database, co-expression, neighborhood, gene fusion and co-occurrence, with medium confidence score (0.400). The networks were built using the target proteins as central nodes and expanding them with the first neighbors, using 500 as maximum value of interactions. From STRING, we obtained the PPI network data and KEGG [60] pathway functional enrichments ($p < 0.05$) of proteins involved in these networks. Then we built the protein networks using Cytoscape [61], adding KEGG information over them.

4.6. Statistical Analysis

The analyses were performed using SPSS version 20 (SPSS Inc., Chicago, IL, USA). Association coefficients were determined using Spearman two tailed test and correspondence analysis were determined by Chi-squared, statistical significance was considered when $p < 0.05$.

5. Conclusions

In this work, we validated a computational approach regarding epitope prediction, topological structure and pathway analyses to drive a rational vaccine design against leishmaniasis, using antigens tested in murine model described in literature. Our results suggest that $CD4^+$ and $CD8^+$ T cells are more related to leishmaniasis protection in EO than B cells. For a deeper analysis, we also used PPI networks enriched with KEGG pathways information. According to our results, proteins associated to protection are connected to few pathways when compared to proteins classified as "no protection". In addition, analysis of PPIs and KEGG pathways associated to proteins from "protection" group corroborate the idea already published in the literature that reverse vaccinology approaches are able to identify proteins related to pathogenicity of infectious agents, helping researchers to understand virulence mechanisms and how immune responses from hosts are able to fight them. Obtained results may be helpful in discovering new potential antigens using computational approaches.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/18/2/371/s1.

Acknowledgments: The authors acknowledge the Brazilian agencies Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) (MCTI/CNPq/CT-BIOTEC-GENOPROT Nº 21/2010 560943/2010-5–Alexandre B. Reis, MCTI/CNPq/CT-BIOTEC Nº 27/2013 402688/2013-9–Alexandre B. Reis, PQ 2015 301526/2015-0–Jeronimo C. Ruiz MCTI/CNPq Nº 14/2015 301652/2015-0–Jeronimo C. Ruiz and MCTI/CNPq Nº 14/2013 486618/2013-7–Jeronimo C. Ruiz), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (PRONEX APQ-01373-14–Alexandre B. Reis, FAPEMIG 01/2013 APQ-01661-13–Jeronimo C. Ruiz and PPM IX 02/2015 PPM-00710-15–Jeronimo C. Ruiz), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Universidade Federal de Ouro Preto (UFOP) and Fundação Oswaldo Cruz (FIOCRUZ) for financial support. Rory C. F. Brito, Daniela M. Resende and João P. L. Velloso are grateful for CAPES scholarships, Alexandre B. Reis, Rodrigo Corrêa-Oliveira and Jeronimo C. Ruiz are also grateful for CNPq fellowships.

Author Contributions: Rory C. F. Brito participated with ideas and suggestions linked with the development and implementation of the bioinformatics methodology; also involved in the selection of candidate proteins to vaccine development against leishamaniasis in the literature, epitopes and subcelullar localization predictions and formal analysis of the data. Frederico G. Guimarães was responsable for PPI networks construction and analysis, and in mapping KEGG pathways associated to immunogenic proteins in PPI networks. Rory C. F. Brito, Frederico G. Guimarães and João P. L. Velloso participated in drafting the article and/or revising it critically for important intellectual content, and also created the figures. Rodrigo Corrêa-Oliveira, Jeronimo C. Ruiz, Alexandre B. Reis and Daniela M. Resende contributed materials and analysis tools. Jeronimo C. Ruiz and Daniela M. Resende participated in the design of the bioinformatics work. Jeronimo C. Ruiz, Alexandre B. Reis and Daniela M. Resende participated in the study conception, project administration, critical revision of the article, supervision and funding acquisition.

Conflicts of Interest: The authors declare that there is no conflict of interest.

References

1. Peacock, C.S.; Seeger, K.; Harris, D.; Murphy, L.; Ruiz, J.C.; Quail, M.A.; Peters, N.; Adlem, E.; Tivey, A.; Aslett, M.; et al. Comparative genomic analysis of three *Leishmania* species that cause diverse human disease. *Nat. Genet.* **2007**, *39*, 839–847. [CrossRef] [PubMed]
2. Gillespie, P.M.; Beaumier, C.M.; Strych, U.; Hayward, T.; Hotez, P.J.; Bottazzi, M.E. Status of vaccine research and development of vaccines for leishmaniasis. *Vaccine* **2016**, *34*, 2992–2995. [CrossRef] [PubMed]
3. Rezende, A.M.; Folador, E.L.; Resende, D.M.; Ruiz, J.C. Computational prediction of protein–protein interactions in *Leishmania* predicted proteomes. *PLoS ONE* **2012**, *7*, e51304. [CrossRef] [PubMed]
4. Xu, C.; Ye, B.; Han, Z.; Huang, M.; Zhu, Y. Comparison of transcriptional profiles between $CD4^+$ and $CD8^+$ T cells in HIV type 1-infected patients. *AIDS Res. Hum. Retrovir.* **2014**, *30*, 134–141. [CrossRef] [PubMed]
5. Salay, G.; Dorta, M.L.; Santos, N.M.; Mortara, R.A.; Brodskyn, C.; Oliveira, C.I.; Barbieri, C.L.; Rodrigues, M.M. Testing of four *Leishmania* vaccine candidates in a mouse model of infection with Leishmania (Viannia) braziliensis, the main causative agent of cutaneous leishmaniasis in the New World. *Clin. Vaccine Immunol. CVI* **2007**, *14*, 1173–1181. [CrossRef] [PubMed]
6. Campbell, K.; Diao, H.; Ji, J.; Soong, L. DNA immunization with the gene encoding P4 nuclease of Leishmania amazonensis protects mice against cutaneous Leishmaniasis. *Infect. Immun.* **2003**, *71*, 6270–6278. [CrossRef] [PubMed]
7. Fedeli, C.E.; Ferreira, J.H.; Mussalem, J.S.; Longo-Maugeri, I.M.; Gentil, L.G.; dos Santos, M.R.; Katz, S.; Barbieri, C.L. Partial protective responses induced by a recombinant cysteine proteinase from *Leishmania* (Leishmania) amazonensis in a murine model of cutaneous leishmaniasis. *Exp. Parasitol.* **2010**, *124*, 153–158. [CrossRef] [PubMed]
8. Montalvo-Alvarez, A.M.; Folgueira, C.; Carrion, J.; Monzote-Fidalgo, L.; Canavate, C.; Requena, J.M. The Leishmania HSP20 is antigenic during natural infections, but, as DNA vaccine, it does not protect BALB/c mice against experimental *L. amazonensis* infection. *J. Biomed. Biotechnol.* **2008**, *2008*, 695432. [CrossRef] [PubMed]
9. Champsi, J.; McMahon-Pratt, D. Membrane glycoprotein M-2 protects against Leishmania amazonensis infection. *Infect. Immun.* **1988**, *56*, 3272–3279. [PubMed]
10. Gonzalez, C.R.; Noriega, F.R.; Huerta, S.; Santiago, A.; Vega, M.; Paniagua, J.; Ortiz-Navarrete, V.; Isibasi, A.; Levine, M.M. Immunogenicity of a Salmonella typhi CVD 908 candidate vaccine strain expressing the major surface protein gp63 of Leishmania mexicana mexicana. *Vaccine* **1998**, *16*, 1043–1052. [CrossRef]
11. Mendez, S.; Gurunathan, S.; Kamhawi, S.; Belkaid, Y.; Moga, M.A.; Skeiky, Y.A.; Campos-Neto, A.; Reed, S.; Seder, R.A.; Sacks, D. The potency and durability of DNA- and protein-based vaccines against Leishmania major evaluated using low-dose, intradermal challenge. *J. Immunol.* **2001**, *166*, 5122–5128. [CrossRef] [PubMed]
12. Rivier, D.; Bovay, P.; Shah, R.; Didisheim, S.; Mauel, J. Vaccination against Leishmania major in a CBA mouse model of infection: Role of adjuvants and mechanism of protection. *Parasite Immunol.* **1999**, *21*, 461–473. [CrossRef] [PubMed]
13. Sjolander, A.; Baldwin, T.M.; Curtis, J.M.; Bengtsson, K.L.; Handman, E. Vaccination with recombinant Parasite Surface Antigen 2 from Leishmania major induces a Th1 type of immune response but does not protect against infection. *Vaccine* **1998**, *16*, 2077–2084. [CrossRef]
14. Webb, J.R.; Campos-Neto, A.; Ovendale, P.J.; Martin, T.I.; Stromberg, E.J.; Badaro, R.; Reed, S.G. Human and murine immune responses to a novel Leishmania major recombinant protein encoded by members of a multicopy gene family. *Infect. Immun.* **1998**, *66*, 3279–3289. [PubMed]
15. Solioz, N.; Blum-Tirouvanziam, U.; Jacquet, R.; Rafati, S.; Corradin, G.; Mauel, J.; Fasel, N. The protective capacities of histone H1 against experimental murine cutaneous leishmaniasis. *Vaccine* **1999**, *18*, 850–859. [CrossRef]
16. Soussi, N.; Milon, G.; Colle, J.H.; Mougneau, E.; Glaichenhaus, N.; Goossens, P.L. Listeria monocytogenes as a short-lived delivery system for the induction of type 1 cell-mediated immunity against the p36/LACK antigen of Leishmania major. *Infect. Immun.* **2000**, *68*, 1498–1506. [CrossRef] [PubMed]
17. Carrion, J.; Folgueira, C.; Alonso, C. Immunization strategies against visceral leishmaniosis with the nucleosomal histones of Leishmania infantum encoded in DNA vaccine or pulsed in dendritic cells. *Vaccine* **2008**, *26*, 2537–2544. [CrossRef] [PubMed]

18. Santos-Gomes, G.M.; Rodrigues, A.; Teixeira, F.; Carreira, J.; Alexandre-Pires, G.; Carvalho, S.; Santos-Mateus, D.; Martins, C.; Vale-Gato, I.; Marques, C.; et al. Immunization with the Leishmania infantum recombinant cyclophilin protein 1 confers partial protection to subsequent parasite infection and generates specific memory T cells. *Vaccine* **2014**, *32*, 1247–1253. [CrossRef] [PubMed]
19. Agallou, M.; Smirlis, D.; Soteriadou, K.P.; Karagouni, E. Vaccination with Leishmania histone H1-pulsed dendritic cells confers protection in murine visceral leishmaniasis. *Vaccine* **2012**, *30*, 5086–5093. [CrossRef] [PubMed]
20. Khoshgoo, N.; Zahedifard, F.; Azizi, H.; Taslimi, Y.; Alonso, M.J.; Rafati, S. Cysteine proteinase type III is protective against Leishmania infantum infection in BALB/c mice and highly antigenic in visceral leishmaniasis individuals. *Vaccine* **2008**, *26*, 5822–5829. [CrossRef] [PubMed]
21. Aguilar-Be, I.; da Silva Zardo, R.; Paraguai de Souza, E.; Borja-Cabrera, G.P.; Rosado-Vallado, M.; Mut-Martin, M.; Garcia-Miss Mdel, R.; Palatnik de Sousa, C.B.; Dumonteil, E. Cross-protective efficacy of a prophylactic Leishmania donovani DNA vaccine against visceral and cutaneous murine leishmaniasis. *Infect. Immun.* **2005**, *73*, 812–819. [CrossRef]
22. Ghosh, A.; Zhang, W.W.; Matlashewski, G. Immunization with A2 protein results in a mixed Th1/Th2 and a humoral response which protects mice against Leishmania donovani infections. *Vaccine* **2001**, *20*, 59–66. [CrossRef]
23. Resende, D.M.; Rezende, A.M.; Oliveira, N.J.; Batista, I.C.; Correa-Oliveira, R.; Reis, A.B.; Ruiz, J.C. An assessment on epitope prediction methods for protozoa genomes. *BMC Bioinform.* **2012**, *13*, 309. [CrossRef] [PubMed]
24. Costa, C.H.; Peters, N.C.; Maruyama, S.R.; de Brito, E.C., Jr.; Santos, I.K. Vaccines for the leishmaniases: Proposals for a research agenda. *PLoS Negl. Trop. Dis.* **2011**, *5*, e943.
25. Reed, S.G.; Coler, R.N.; Campos-Neto, A. Development of a leishmaniasis vaccine: The importance of MPL. *Expert Rev. Vaccines* **2003**, *2*, 239–252. [CrossRef] [PubMed]
26. Larsen, M.V.; Lundegaard, C.; Lamberth, K.; Buus, S.; Lund, O.; Nielsen, M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinform.* **2007**, *8*, 424. [CrossRef] [PubMed]
27. Nielsen, M.; Lundegaard, C.; Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinform.* **2007**, *8*, 238. [CrossRef] [PubMed]
28. Larsen, J.E.; Lund, O.; Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immun. Res.* **2006**, *2*, 2. [CrossRef] [PubMed]
29. Chen, J.; Liu, H.; Yang, J.; Chou, K.C. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids* **2007**, *33*, 423–428. [CrossRef] [PubMed]
30. El Manzalawy, Y.; Dobbs, D.; Honavar, V. Predicting linear B-cell epitopes using string kernels. *J. Mol. Recognit.* **2008**, *21*, 243–255. [CrossRef] [PubMed]
31. El Manzalawy, Y.; Dobbs, D.; Honavar, V. Predicting flexible length linear B-cell epitopes. *Comput. Syst. Bioinform. Conf.* **2008**, *7*, 121–132.
32. Horton, P.; Park, K.J.; Obayashi, T.; Fujita, N.; Harada, H.; Adams-Collier, C.J.; Nakai, K. WoLF PSORT: Protein localization predictor. *Nucleic Acids Res.* **2007**, *35*, W585–W587. [CrossRef] [PubMed]
33. Herrera-Najera, C.; Pina-Aguilar, R.; Xacur-Garcia, F.; Ramirez-Sierra, M.J.; Dumonteil, E. Mining the Leishmania genome for novel antigens and vaccine candidates. *Proteomics* **2009**, *9*, 1293–1301. [CrossRef] [PubMed]
34. Agallou, M.; Athanasiou, E.; Koutsoni, O.; Dotsika, E.; Karagouni, E. Experimental Validation of Multi-Epitope Peptides Including Promising MHC Class I- and II-Restricted Epitopes of Four Known Leishmania infantum Proteins. *Front. Immunol.* **2014**, *5*, 268. [CrossRef] [PubMed]
35. Duarte, A.; Queiroz, A.T.; Tosta, R.; Carvalho, A.M.; Barbosa, C.H.; Bellio, M.; de Oliveira, C.I.; Barral-Netto, M. Prediction of $CD8^+$ Epitopes in Leishmania braziliensis Proteins Using EPIBOT: In Silico Search and In Vivo Validation. *PLoS ONE* **2015**, *10*, e0124786. [CrossRef] [PubMed]
36. Freitas e Silva, R.; Ferreira, L.F.; Hernandes, M.Z.; de Brito, M.E.; de Oliveira, B.C.; da Silva, A.A.; de-Melo-Neto, O.P.; Rezende, A.M.; Pereira, V.R. Combination of In Silico Methods in the Search for Potential CD4(+) and CD8(+) T Cell Epitopes in the Proteome of Leishmania braziliensis. *Front. Immunol.* **2016**, *7*, 327.
37. Day, M.J. Immunoglobulin G subclass distribution in canine leishmaniosis: A review and analysis of pitfalls in interpretation. *Vet. Parasitol.* **2007**, *147*, 2–8. [CrossRef] [PubMed]

38. De Almeida Silva, L.; Romero, H.D.; Prata, A.; Costa, R.T.; Nascimento, E.; Carvalho, S.F.; Rodrigues, V. Immunologic tests in patients after clinical cure of visceral leishmaniasis. *Am. J. Trop. Med. Hyg.* **2006**, *75*, 739–743. [PubMed]
39. Rosa, R.; Marques, C.; Rodrigues, O.R.; Santos-Gomes, G.M. Immunization with Leishmania infantum released proteins confers partial protection against parasite infection with a predominant Th1 specific immune response. *Vaccine* **2007**, *25*, 4525–4532. [CrossRef] [PubMed]
40. Lemesre, J.L.; Holzmuller, P.; Goncalves, R.B.; Bourdoiseau, G.; Hugnet, C.; Cavaleyra, M.; Papierok, G. Long-lasting protection against canine visceral leishmaniasis using the LiESAp-MDP vaccine in endemic areas of France: Double-blind randomised efficacy field trial. *Vaccine* **2007**, *25*, 4223–4234. [CrossRef] [PubMed]
41. Alcolea, P.J.; Alonso, A.; Sanchez-Gorostiaga, A.; Moreno-Paz, M.; Gomez, M.J.; Ramos, I.; Parro, V.; Larraga, V. Genome-wide analysis reveals increased levels of transcripts related with infectivity in peanut lectin non-agglutinated promastigotes of Leishmania infantum. *Genomics* **2009**, *93*, 551–564. [CrossRef] [PubMed]
42. Benitez, D.; Medeiros, A.; Fiestas, L.; Panozzo-Zenere, E.A.; Maiwald, F.; Prousis, K.C.; Roussaki, M.; Calogeropoulou, T.; Detsi, A.; Jaeger, T.; et al. Identification of Novel Chemical Scaffolds Inhibiting Trypanothione Synthetase from Pathogenic Trypanosomatids. *PLoS Negl. Trop. Dis.* **2016**, *10*, e0004617. [CrossRef] [PubMed]
43. Iborra, S.; Parody, N.; Abanades, D.R.; Bonay, P.; Prates, D.; Novais, F.O.; Barral-Netto, M.; Alonso, C.; Soto, M. Vaccination with the Leishmania major ribosomal proteins plus CpG oligodeoxynucleotides induces protection against experimental cutaneous leishmaniasis in mice. *Microb. Inf. Inst. Pasteur* **2008**, *10*, 1133–1141. [CrossRef] [PubMed]
44. Cordeiro-Da-Silva, A.; Borges, M.C.; Guilvard, E.; Ouaissi, A. Dual role of the Leishmania major ribosomal protein S3a homologue in regulation of T- and B-cell activation. *Infect. Immun.* **2001**, *69*, 6588–6596. [CrossRef] [PubMed]
45. Soto, M.; Requena, J.M.; Garcia, M.; Gomez, L.C.; Navarrete, I.; Alonso, C. Genomic organization and expression of two independent gene arrays coding for two antigenic acidic ribosomal proteins of Leishmania. *J. Biol. Chem.* **1993**, *268*, 21835–21843. [PubMed]
46. Soto, M.; Alonso, C.; Requena, J.M. The Leishmania infantum acidic ribosomal protein LiP2a induces a prominent humoral response in vivo and stimulates cell proliferation in vitro and interferon-gamma (IFN-γ) production by murine splenocytes. *Clin. Exp. Immunol.* **2000**, *122*, 212–218. [CrossRef] [PubMed]
47. Silverman, J.M.; Chan, S.K.; Robinson, D.P.; Dwyer, D.M.; Nandan, D.; Foster, L.J.; Reiner, N.E. Proteomic analysis of the secretome of Leishmania donovani. *Genome Biol.* **2008**, *9*, R35. [CrossRef] [PubMed]
48. McConville, M.J.; Naderer, T. Metabolic pathways required for the intracellular survival of Leishmania. *Annu. Rev. Microbiol.* **2011**, *65*, 543–561. [CrossRef] [PubMed]
49. Naderer, T.; Dandash, O.; McConville, M.J. Calcineurin is required for Leishmania major stress response pathways and for virulence in the mammalian host. *Mol. Microbiol.* **2011**, *80*, 471–480. [CrossRef] [PubMed]
50. Madeira da Silva, L.; Beverley, S.M. Expansion of the target of rapamycin (TOR) kinase family and function in Leishmania shows that TOR3 is required for acidocalcisome biogenesis and animal infectivity. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 11965–11970. [CrossRef] [PubMed]
51. Carter, N.S.; Yates, P.; Arendt, C.S.; Boitz, J.M.; Ullman, B. Purine and pyrimidine metabolism in Leishmania. *Adv. Exp. Med. Biol.* **2008**, *625*, 141–154. [PubMed]
52. Real, F.; Vidal, R.O.; Carazzolle, M.F.; Mondego, J.M.; Costa, G.G.; Herai, R.H.; Wurtele, M.; de Carvalho, L.M.; Carmona e Ferreira, R.; Mortara, R.A.; et al. The genome sequence of *Leishmania* (*Leishmania*) *amazonensis*: Functional annotation and extended analysis of gene models. *DNA Res.* **2013**, *20*, 567–581. [CrossRef] [PubMed]
53. Peters, B.; Bulik, S.; Tampe, R.; Van Endert, P.M.; Holzhutter, H.G. Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors. *J. Immunol.* **2003**, *171*, 1741–1749. [CrossRef] [PubMed]
54. 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*, 33–41. [CrossRef] [PubMed]

55. Buus, S.; Lauemoller, S.L.; 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*, 378–384. [CrossRef] [PubMed]
56. Nielsen, M.; Lundegaard, C.; Worning, P.; Hvid, C.S.; 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*, 1388–1397. [CrossRef] [PubMed]
57. Nielsen, M.; Lundegaard, C.; Worning, P.; Lauemoller, S.L.; 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*, 1007–1017. [CrossRef] [PubMed]
58. Nielsen, M.; Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinform.* **2009**, *10*, 296. [CrossRef] [PubMed]
59. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P.; et al. STRING v10: Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–D452. [CrossRef] [PubMed]
60. Kanehisa, M. KEGG Bioinformatics Resource for Plant Genomics and Metabolomics. *Methods Mol. Biol.* **2016**, *1374*, 55–70. [PubMed]
61. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]