Integração de Dados na Detecção de Alvos para Fármacos de *Schistosoma mansoni*

Francimary P. Garcia[1], **Kele Teixeira Belloze**[1]

[1]Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ

francigarciaoliveira@gmail.com, kele.belloze@cefet-rj.br

Abstract. *Schistosomiasis mansoni caused by Schistosoma mansoni organism is a major neglected disease occurring in the world. However, there is a single drug recommended by the World Health Organization for its treatment. Therefore, searching for alternative drug targets in the fight against the disease is important. This work aims to identify possible new drug targets for S. mansoni. The methodology adopts an approach based on orthology and homology making use of essential proteins of model organisms and proteins already known as drug targets, and integration of these data. As a preliminary result, a list of 91 candidate proteins for drug targets was found.*

Resumo. *A esquistossomose causada pelo organismo Schistosoma mansoni é uma doença negligenciada importante pela ocorrência no mundo. Contudo, existe um único medicamento recomendado pela Organização Mundial de Saúde para o seu tratamento. Logo, pesquisas por alvos para fármacos alternativos no combate à doença são importantes. Este trabalho tem como objetivo identificar possíveis novos alvos para fármacos de S. mansoni. A metodologia adota uma abordagem baseada em ortologia e homologia fazendo uso de proteínas essenciais de organismos modelo e proteínas já conhecidas como alvos de fármacos, e integração destes dados. Resultados preliminares apontam uma lista de 91 proteínas candidatas a alvos para fármaco.*

1. Introdução

A esquistossomose é uma doença negligenciada causada por organismos helmintos (vermes parasitários) do gênero *Schistosoma*. De acordo com a Organização Mundial de Saúde (OMS), a doença afeta quase 240 milhões de pessoas em todo o mundo [WHO 2018]. A ocorrência da doença prevalece em áreas tropicais e subtropicais, em comunidades pobres sem água potável e saneamento adequado. Mais de 700 milhões de pessoas vivem em áreas com essas características no mundo. O *S. mansoni*, causador da maioria das infecções em humanos, é a única espécie do gênero descrita no Brasil [Souza et al. 2011]. De acordo com dados do Ministério da Saúde, o número de casos da doença na área endêmica do Brasil, a qual engloba principalmente a região Nordeste, é de quase 20 mil [MS 2017].

Para o combate à doença é utilizado o medicamento praziquantel (PZQ) recomendado pela OMS, que tem sido usado na prática clínica há quase quatro décadas [Neves et al. 2016]. No entanto, devido à alta incidência de reinfecção, o uso generalizado e repetido deste medicamento em áreas endêmicas, suscita preocupações sobre o

desenvolvimento de resistência ao medicamento pelo helminto. Este problema é enfatizado ainda mais pela falta de eficácia do PZQ contra os vermes juvenis, o que é uma causa potencial de falha no tratamento em áreas endêmicas [Caffrey et al. 2009]. Por esta razão, pesquisas por alvos para fármacos alternativos no combate à esquistossomose são urgentes.

Diante do cenário apresentado, o objetivo deste trabalho é identificar possíveis novos alvos (proteínas) para fármacos de *S. mansoni* no combate à doença, tendo como foco os atributos de essencialidade e drogabilidade das proteínas. Para a metodologia desse trabalho, é feita uma integração de dados obtidos por meio da aplicação de uma abordagem baseada em ortologia e homologia na qual são utilizados: i) dados sobre as proteínas essenciais de organismos modelo, para trabalhar o atributo da essencialidade e, ii) dados sobre proteínas já classificadas como alvos de fármacos desenvolvidos e comercializados, para trabalhar o atributo da drogabilidade. Como resultado foi encontrada uma lista de 91 proteínas de *S. mansoni* candidatas a alvos para fármacos.

Além dessa introdução, este artigo está organizado nas seguintes seções: a seção 2 descreve os conceitos que embasam este trabalho; a seção 3 apresenta os trabalhos relacionados; a seção 4 detalha a metodologia usada na condução da pesquisa; a seção 5 apresenta os resultados obtidos e a seção 6 descreve as considerações finais sobre o artigo.

2. Alvos para Fármacos e Homologia

A descoberta de fármacos baseada em alvo é uma técnica comumente utilizada, porque pode reduzir os custos de algumas experiências laboratoriais necessárias para o processo inicial de desenvolvimento de fármacos [Guido et al. 2010]. Contudo, essa é uma tarefa não trivial a partir de dados experimentais. Sendo assim, as análises *in silico* apoiam essa descoberta levantando características consideradas desejáveis em um alvo para fármaco, como a essencialidade (se ausente, causa a morte da célula biológica), a drogabilidade (se as moléculas semelhantes a fármacos são suscetíveis de interagir com o alvo), a especificidade/seletividade (potencial para inibir o patógeno sem prejudicar o hospedeiro) e a importância das fases do ciclo de vida do patógeno relevantes para a saúde humana [Crowther et al. 2010].

Entre as estratégias existentes para a descoberta de fármacos para tratar doenças tropicais negligenciadas, podemos citar a abordagem baseada em homologia. Essa abordagem é usada para inferir relações biológicas e características da evolução entre os organismos que estão sendo comparados [Morrison et al. 2015].

Homologia é a relação de ancestralidade entre duas ou mais entidades (e.g. genes ou proteínas), ou seja, significa dizer que as mesmas compartilham um ancestral comum [Koonin 2005]. Sendo assim, a homologia é um termo qualitativo [Moreira 2015]. A similaridade, por sua vez, corresponde ao grau de proximidade entre duas ou mais sequências moleculares, geralmente expresso em porcentagem (%). Portanto, a similaridade é um termo quantitativo. Sequências ou estruturas similares podem ou não compartilhar de um ancestral comum. Por exemplo, podemos dizer que dois genes homólogos são 90% similares no nível da sequência de nucleotídeos. Porém, estes genes não podem ser referidos como 90% homólogos [Moreira 2015].

Em um conceito mais específico, a ortologia é definida como um tipo de homologia onde genes de diferentes espécies descendem de um único gene no último ancestral

comum, a partir de um processo de especiação [Fitch 1970]. Os genes ortólogos são importantes para a compreensão da genômica e da biologia molecular. Isso se deve ao fato que conhecida a função de um determinado gene ortólogo A que se apresenta ortólogo a um gene B recém-sequenciado ou com função desconhecida, é possível inferir, ao menos de forma provisória, a função do gene B por meio de sua alta similaridade e conservação com esse gene bem conhecido (gene A) [Moreira 2015].

3. Trabalhos Relacionados

TDR Targets [Agüero et al. 2008] é um trabalho de destaque quando o assunto é a identificação de alvos para fármacos em patógenos de doenças negligenciadas. TDR Targets é um banco de dados que foi criado para facilitar as análises focadas em alvo para esses patógenos, os quais são priorizados pelo Programa Especial de Pesquisa e Treinamento em Doenças Tropicais (TDR) da Organização Mundial de Saúde. O banco de dados pode ser usado para duas tarefas científicas gerais: i) análise de proteínas individuais, encontrando informações relacionadas ao seu potencial como alvo para fármacos e; ii) triagem e classificação de múltiplas proteínas como candidatas alvo para fármacos de acordo com os critérios especificados pelo usuário.

Os trabalhos de [Crowther et al. 2010] e [Caffrey et al. 2009] utilizam análise *in silico* para identificar alvos para fármacos. Crowther e colaboradores (2010) apresentaram uma abordagem para priorizar as proteínas dos patógenos observando se as mesmas atendem aos critérios considerados desejáveis em um alvo para fármacos. Esses critérios são baseados em ambas as informações derivadas da sequência (por exemplo, massa molecular) e dos dados funcionais sobre a expressão, essencialidade, fenótipos, vias metabólicas e drogabilidade. Esta abordagem também destaca o fato que para muitos critérios relevantes faltam dados em patógenos menos estudados (por exemplo, helmintos), sendo demonstrado como essa questão pode ser parcialmente vencida utilizando-se do mapeamento de dados de genes homólogos em organismos bem estudados.

Caffrey e colaboradores (2009) empregaram uma abordagem de química comparativa utilizando o genoma do *S. mansoni* de forma a identificar genes essenciais putativos com base em semelhança com genes/proteínas essenciais identificados por determinação experimental em dois organismos modelo, o nematoide *Caenorhabditis elegans* e a mosca da fruta *Drosophila melanogaster*. Em seguida, definiram um subconjunto de possíveis alvos para fármacos para os quais as informações estruturais de proteínas alvo são conhecidas, incluindo ligantes.

O presente trabalho adotou, de maneira similar aos trabalhos apresentados, os atributos de essencialidade e drogabilidade para o levantamento de proteínas candidatas a alvo para fármaco de *S. mansoni*. Por meio de uma abordagem baseada em ortologia e homologia, utiliza os dados de proteínas bem anotadas e conhecidas como as proteínas dos organismos modelo e de um banco de dados de alvos para fármacos para levantar as proteínas do *S. mansoni* que podem conter os atributos de essencialidade e drogabilidade. O diferencial em relação aos demais trabalhos é a integração dos dados realizada, na qual encontra inicialmente as proteínas candidatas essenciais do *S. mansoni* e a partir destas, encontra as proteínas com características de drogabilidade, resultando assim em conjunto de proteínas com características de essencialidade e drogabilidade ao mesmo tempo.

4. Metodologia

A metodologia aplicada na condução desta pesquisa baseou-se no trabalho de [Belloze 2013] que propôs a utilização dos conceitos de homologia e atributos de essencialidade e drogabilidade da proteína para apoiar a priorização de alvos no combate a doenças tropicais negligenciadas causadas por protozoários. O presente trabalho se diferencia no organismo de estudo, na ferramenta adotada para a busca de proteínas ortólogas (Atividade 1 descrita a seguir) e na integração de dados realizada. Assim, a metodologia aplicada considerando suas modificações em relação ao trabalho supracitado é apresentada na Figura 1 e detalhada nas atividades 1 e 2 descritas a seguir.

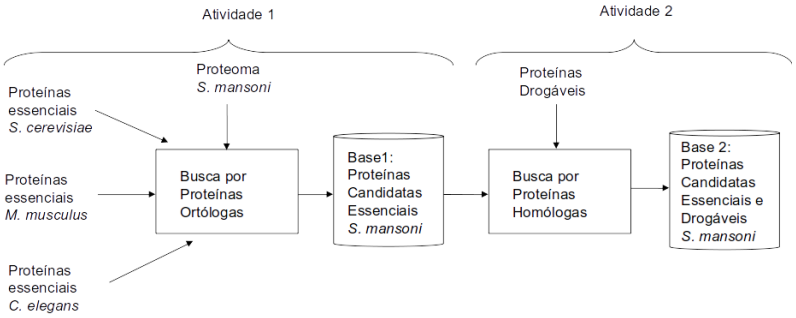

Figura 1. Busca de proteínas ortólogas, entre as proteínas do *S. mansoni* e as proteínas essenciais dos três organismos modelo apresentados e com este resultado, a busca de proteínas homólogas, contra proteínas drogáveis, resultando na base de proteínas candidatas essenciais e drogáveis do *S. mansoni*.

Atividade 1: inicialmente foi realizada a identificação de proteínas ortólogas entre as proteínas do *S. mansoni* e as proteínas essenciais de três organismos modelo eucarióticos: *Caenorhabditis elegans* (nematódeo), *Saccharomyces cerevisiae* (levedura) e *Mus musculus* (camundongo), baseando-se no conceito de essencialidade. De acordo com o conceito de ortologia, duas proteínas ortólogas podem ter a mesma função. Sendo a proteína do organismo modelo uma proteína essencial, foi pretendido então, por ortologia, sugerir o atributo de essencialidade às proteínas do *S. mansoni*. As proteínas ortólogas obtidas foram submetidas a um critério de corte considerando a ocorrência repetida nos quatro organismos. Desta maneira, foi obtida a base intermediária (Base 1) de proteínas candidatas a essenciais do *S. mansoni*.

O proteoma do *S. mansoni* foi obtido a partir da base *Ensembl Metazoa* [Kersey et al. 2017], enquanto as proteínas essenciais dos organismos modelo foram obtidas a partir da base de genes essenciais DEG (Database of Essential Gene) [Zhang et al. 2004].

Para a busca de ortologia foi utilizada a ferramenta *Orthofinder* [Emms and Kelly 2015], selecionada por aplicar um método que infere grupos de ortólogos (ortogrupos) de genes codificadores de proteínas. Na busca por sequências ortólogas, o *Orthofinder* executa as seguintes atividades:

1- Realiza busca BLAST [Altschul et al. 1990] *all-versus-all* (utilizando *e-value* padrão 1e-3); 2- Compara comprimento do gene e realiza normalização filogenética da distância do parâmetro *Score bit* (mede a similaridade de sequência, independente do tamanho da sequência de consulta e do tamanho do banco de dados) do BLAST, para que

os melhores resultados entre todas as espécies alcancem as mesmas pontuações, independentemente do comprimento da sequência ou da distância filogenética; 3- Delimita limiares de similaridade de sequência de ortogrupos usando RBNHs (*Reciprocal Best Length-Normalised Hit*); 4- Constrói um gráfico de ortogrupos para entrada no MCL (*Markov Cluster Algorithm*) [Enright et al. 2002]; 5- Agrupa genes em ortogrupos usando o MCL.

Atividade 2: em seguida, foi conduzido o processo de identificação de proteínas homólogas entre as proteínas candidatas a essenciais do *S. mansoni* levantadas na atividade 1 (Base1) e proteínas drogáveis (alvos para fármacos) disponibilizadas publicamente no banco de dados *DrugBank* [Wishart et al. 2017]. Nesta atividade, foi considerada apenas a homologia entre as sequências, pois duas proteínas homólogas possuem alta similaridade. Logo, se uma proteína que já é um alvo para fármaco e, portanto, possui características de drogabilidade, for altamente similar a uma proteína do *S. mansoni*, podemos sugerir que esta última pode conter características de drogabilidade também, não importando a função. As proteínas já consideradas alvos para fármacos foram obtidas dos conjuntos de dados das categorias *Approved* e *Small Molecule* do banco de dados *DrugBank*. A ferramenta BLAST foi utilizada para identificação das proteínas homólogas. Esta atividade resultou em uma base de dados (Base 2) composta por proteínas candidatas essenciais e drogáveis do organismo estudado, representadas por sequências primárias.

5. Resultados

Para a identificação das proteínas ortólogas foram utilizadas as sequências de proteínas dos quatro organismos (*S. mansoni* e os três organismos modelo) na mesma execução, possibilitando desta forma, a construção da árvore filogenética dos organismos e posterior identificação de ortólogos. O número de proteínas ortólogas encontradas é mostrado na Figura 2, na qual, por meio de uma representação de conjuntos, estão destacadas as quantidades de proteínas ortólogas entre o *S. mansoni* e cada um dos organismos modelo e as proteínas em comum a cada dois e três organismos modelo. A ortologia entre o *S. mansoni* e o *C. elegans*, por exemplo, resultou em 169 proteínas que só ocorreram nessa combinação, 118 proteínas que ocorreram também na ortologia entre o *S. mansoni* e o *S. cerevisae*, 111 proteínas que ocorreram também na ortologia entre o *S. mansoni* e o *M. musculus* e 138 proteínas que ocorreram nos três processos de ortologia.

Para continuidade da pesquisa, foram utilizadas as sequências de proteínas que representaram a interseção do resultado da ortologia entre as proteínas de *S. mansoni* e as proteínas essenciais dos três organismos modelo, ou seja, 138 proteínas do *S. mansoni* também encontradas na base de proteínas essenciais dos três organismos modelo, representando assim uma maior chance de se caracterizarem como essenciais.

Para a identificação das proteínas homólogas, a base de proteínas drogáveis foi composta de 7.172 sequências, das quais 2.683 sequências pertencentes à categoria *Approved* e 4.489 sequências pertencentes à categoria *Small Molecule*. Foi executada a ferramenta BLAST, na sua modalidade BLASTp (comparação entre sequências de proteínas) utilizando como entrada (proteína *query*), o arquivo com 138 sequências do *S. mansoni* candidatas essenciais e a base de proteínas drogáveis citada anteriormente.

Algumas execuções do BLASTp foram realizadas a fim de identificar os melhores valores para os parâmetros *evalue, best_hit_score_edge e best_hit_overhang*, além de consultas à literatura. Os parâmetros usados realizam as seguintes restrições às combinações

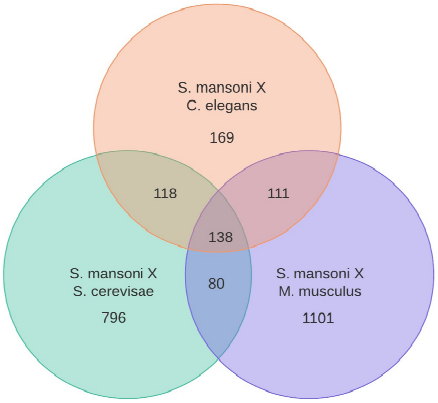

Figura 2. Análise quantitativa da ortologia realizada entre o proteoma do *S. mansoni* e cada conjunto de proteínas essenciais dos três organismos modelo e suas interseções.

obtidas: *e-value* indica o número de alinhamentos que seriam esperados apresentando valores de escore iguais ou melhores que o encontrado por acaso, dado o tamanho do banco de dados; *best_hit_score_edge* restringe os resultados aos melhores hits encontrados para cada *query* dentro do valor de e-value escolhido e *best_hit_overhang* controla quando um HSP (High-scoring Segment Pair) é considerado suficientemente curto para ser filtrado devido à presença de outro HSP.

Os valores usados foram: *e-value* = 1e-10 , *best_hit_score_edge* = 0.05 e *best_hit_overhang* = 0.25. Como resultado, foi obtida uma lista com 91 proteínas candidatas à alvos para fármacos. Uma lista com as 10 proteínas que apresentaram maiores percentuais (%) de identidade é mostrada na Tabela 1, na qual são apresentadas as seguintes informações: identificador da proteína do *S. mansoni*, nome da proteína, identificador da proteína homóloga do *DrugBank*, os valores de *e-value*, *bit score* e percentual (%) de identidade (Ident), obtidos no processo de homologia do BLAST. É apresentada também a informação sobre qual a categoria do *DrugBank* ocorreu a homologia, se *Approved* ou *Small Molecule*.

6. Conclusão

A pesquisa por alvos para fármacos que combatam a esquistossomose tem caráter importante devido ao elevado número de pessoas expostas a condições de pobreza que favorecem a ocorrência da doença nos países subdesenvolvidos. A dificuldade de investimentos da iniciativa privada neste setor e a existência de apenas um medicamento e que ainda pode vir a desenvolver resistência pelo parasita, representam preocupações que confirmam a necessidade de pesquisas por fármacos alternativos ao *S. mansoni*.

A metodologia proposta neste trabalho foca nos atributos de essencialidade e drogabilidade das proteínas na busca por candidatas a alvos que possam ser utilizados em pesquisa por novos fármacos, reduzindo o tempo e o custo envolvidos no processo de desenvolvimento de fármacos. Após as etapas realizadas nas atividades 1 e 2 da metodologia, foi obtida uma lista com 91 proteínas do organismo estudado, consideradas candidatas essenciais e drogáveis para acompanhamento experimental em testes de bancada a fim de verificar possíveis novos alvos para fármacos.

Tabela 1. Da lista de 91 proteínas candidatas essenciais e drogáveis do *S. mansoni*, são apresentadas as dez proteínas com maiores percentuais de identidade. A=Approved, S=Small.

S. mansoni	Nome da Proteína	ID Drugbank	Evalue	Bit Score	%Ident	A /S
Smp_026560.1	Putative calmodulin	P0DP25	4E-072	214	99.07	A
Smp_026560.2	Putative calmodulin	P0DP25	3E-103	295	97.99	A
Smp_203130.1	Putative uncharacterized protein	P63261	2E-133	380	96.72	A
Smp_183710.1	Putative actin	P63261	0.0	762	95.99	A
Smp_046600.1	Actin-1	P63261	0.0	762	95.99	A
Smp_161920.1	Putative actin	P63261	0.0	731	93.58	A
Smp_202970.1	Putative uncharacterized protein	P63261	0.0	727	91.15	A
Smp_018240.2	Cell division control protein 48 aaa family protein	P55072	0.0	1092	85.85	S
Smp_018240.1	Cell division control protein 48 aaa family protein	P55072	0.0	1021	84.63	S
Smp_067980.1	Ubiquitin conjugating enzyme E2, putative	P62837	6E-043	137	83.12	S

Este é um trabalho em andamento e como próximos passos desta pesquisa, estão previstas mais três etapas. Primeiro, o cruzamento da lista de proteínas final obtida com uma lista de proteínas essenciais do *Homo sapiens*, de modo a garantir que não incluiremos nesta nenhuma proteína essencial ao ser humano. Em seguida será realizada a identificação de informações sobre as estruturas secundárias destas proteínas, de forma a enriquecer a base de dados concebida. Finalmente, será feita a obtenção de um índice de drogabilidade para a lista de proteínas obtida, utilizando-se de modelos de padrões frequentes como Apriori, de modo a identificar comportamentos consistentes entre as proteínas candidatas e pesos obtidos por meio da análise de características da lista de proteínas.

Referências

Agüero, F., Al-Lazikani, B., Aslett, M., Berriman, M., Buckner, F. S., Campbell, R. K., Carmona, S., Carruthers, I. M., Chan, A. E., Chen, F., et al. (2008). Genomic-scale prioritization of drug targets: the tdr targets database. *Nature reviews Drug discovery*, 7(11):900.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Belloze, K. T. (2013). *Priorização de alvos para fármacos no combate a doenças tropicais negligenciadas causadas por protozoários.* Doutorado em biologia computacional e sistemas, Instituto Oswaldo Cruz, Fiocruz, Rio de Janeiro.

Caffrey, C. R., Rohwer, A., Oellien, F., Marhöfer, R. J., Braschi, S., Oliveira, G., McKerrow, J. H., and Selzer, P. M. (2009). A comparative chemogenomics strategy to predict potential drug targets in the metazoan pathogen, schistosoma mansoni. *PloS one*, 4(2):e4413.

Crowther, G. J., Shanmugam, D., Carmona, S. J., Doyle, M. A., Hertz-Fowler, C., Berriman, M., Nwaka, S., Ralph, S. A., Roos, D. S., Van Voorhis, W. C., et al. (2010).

Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. *PLoS neglected tropical diseases*, 4(8):e804.

Emms, D. M. and Kelly, S. (2015). Orthofinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome biology*, 16(1):157.

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30(7):1575–1584.

Fitch, W. M. (1970). Further improvements in the method of testing for evolutionary homology among proteins. *Journal of molecular biology*, 49(1):1–14.

Guido, R. V. C., Andricopulo, A. D., and Oliva, G. (2010). Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. *Estudos Avançados*, 24:81 – 98.

Kersey, P. J., Allen, J. E., Allot, A., Barba, M., Boddu, S., Bolt, B. J., Carvalho-Silva, D., Christensen, M., Davis, P., Grabmueller, C., et al. (2017). Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. *Nucleic acids research*, 46(D1):D802–D808.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338.

Moreira, L. (2015). Ciências genômicas: fundamentos e aplicações. *Ribeirão Preto: Sociedade Brasileira de Genética*.

Morrison, D. A., Morgan, M. J., and Kelchner, S. A. (2015). Molecular homology and multiple-sequence alignment: an analysis of concepts and practice. *Australian Systematic Botany*, 28(1):46–62.

MS (2017). Ministério da saúde. Disponível em: `http://portalms.saude.gov.br/saude-de-a-z/esquistossomose/situacao-epidemiologica`. Data do Acesso: 21 de Março de 2018.

Neves, B. J., Dantas, R. F., Senger, M. R., Melo-Filho, C. C., Valente, W. C., De Almeida, A. C., Rezende-Neto, J. M., Lima, E. F., Paveley, R., Furnham, N., et al. (2016). Discovery of new anti-schistosomal hits by integration of qsar-based virtual screening and high content screening. *Journal of medicinal chemistry*, 59(15):7075–7088.

Souza, F., Vitorino, R. R., Costa, A., Faria Jr, F., Santana, L., and Gomes, A. P. (2011). Esquistossomose mansônica: aspectos gerais, imunologia, patogênese e história natural. *Rev Bras Clin Med*, 9(4):300–7.

WHO (2018). Shistosomiasis. Disponível em: `http://www.who.int/schistosomiasis/en/`. Data do Acesso: 01 de Março de 2018.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082.

Zhang, R., Ou, H.-Y., and Zhang, C.-T. (2004). Deg: a database of essential genes. *Nucleic acids research*, 32(suppl_1):D271–D272.