CEFET/RJ
DIRETORIA DE PESQUISA E PÓS-GRADUAÇÃO

INTEGRAÇÃO DE DADOS NA DETECÇÃO DE ALVOS PARA FÁRMACOS DE *SCHISTOSOMA MANSONI*

Francimary Procopio Garcia de Oliveira

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Orientador(a): Kele Teixeira Belloze

Rio de Janeiro,

Janeiro 2020

INTEGRAÇÃO DE DADOS NA DETECÇÃO DE ALVOS PARA FÁRMACOS DE *SCHISTOSOMA MANSONI*

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação, do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, CEFET/RJ, como parte dos requisitos necessários à obtenção do título de mestre.

Francimary Procopio Garcia de Oliveira

Banca Examinadora:

Presidente, Professor D.Sc. Kele Teixeira Belloze (CEFET/RJ)

Professor D.Sc. Rafaelli de Carvalho Coutinho (CEFET/RJ)

Professor D.Sc. Ana Carolina Ramos Guimarães (FIOCRUZ)

Rio de Janeiro,

Janeiro 2020

Ficha catalográfica elaborada pela Biblioteca Central do CEFET/RJ

O48 Oliveira, Francimary Procopio Garcia de.
Integração de dados na detecção de alvos para fármacos de Schistosoma mansoni / Francimary Procopio Garcia de Oliveira – 2020.
69f. + apêndice : il. color., grafs., tabs. ; enc.

Dissertação (Mestrado) Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, 2020.
Bibliografia: f. 59-65.
Orientador: Kele Teixeira Belloze.

1. Integração de dados (Computação). 2. Aprendizado do computador. 3. Farmacologia. 4. Schistosoma mansoni. I. Belloze, Kele Teixeira (Orient.). II. Título.

CDD 005.74

Elaborada pela bibliotecária Vanessa Suane de Souza CRB-7/6753

DEDICATÓRIA

Dedico este trabalho antes de tudo a força de Deus que nos guia e orienta a cada dia e à minha família, Raymundo, Mateus e Frederico, que me servem de inspiração e motivação.

AGRADECIMENTOS

O presente trabalho foi desenvolvido com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Agradeço a Profa. Kele Teixeira Belloze pela orientação, sempre presente e me apontando a direção correta, e também pelas ideias, pela paciência, conselhos, pela amizade e momentos únicos até aqui. Ela personifica o significado da palavra orientadora.

Agradeço ao espetacular time de professores do CEFET/RJ, em especial aos professores Eduardo Ogasawara, Gustavo Guedes, Eduardo Bezerra, Rafael Machado e Joel dos Santos, pelos momentos de aprendizado, colaboração, orientação e amizade durante as aulas, os trabalhos, as provas e os desafios que um mestrado apresenta.

Agradeço aos funcionários da secretaria de pós-graduação do CEFET/RJ, sempre prestativos e atentos às necessidades do time de alunos.

Agradeço às professoras Rafaelli Coutinho e Carolina Guimarães por terem aceito o convite para participar da banca avaliadora e por todas as sugestões e comentários tão oportunos e enriquecedores.

Agradeço aos colegas de mestrado pelo apoio nas trocas de experiências, pela amizade, companheirismo e descontração tão necessária em alguns momentos. Destaco o Flávio Damasceno, o Yukio Okuno e o Carlos Guedes pela amizade que certamente ficará de legado.

Por fim agradeço a Deus por ter me guiado para esta oportunidade tão rica para minha vivência, e agradeço a minha família pelo apoio, paciência e carinho neste trajeto até aqui.

RESUMO

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

A esquistossomose causada pelo organismo *Schistosoma mansoni* é uma doença negligenciada importante por sua 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 na integração de dados biológicos que se encontram dispersos por diversos bancos de dados públicos e na utilização dos conceitos de ortologia e homologia para identificação dos atributos de essencialidade e drogabilidade das proteínas. Também foi abordado o método de aprendizado de máquina para identificação do atributo de essencialidade das proteínas do *S. mansoni*, a partir das bases de proteínas essenciais e não-essenciais dos organismos modelo.

Considerando as etapas da metodologia que utilizaram abordagem baseada em homologia, para as quais as características de essencialidade e drogabilidade foram pretendidas, os resultados apresentam uma lista de 15 proteínas candidatas a alvos para fármaco do *S. mansoni*. Em relação ao método de aprendizado de máquina, o classificador *Random Forest* foi indicado como o de melhor desempenho, apresentando um percentual de 79% de acurácia nas atividades de aprendizado. A partir do classificador, 1.412 proteínas foram indicadas como essenciais na atividade de predição das proteínas do *S. mansoni*. A análise comparativa entre os dois métodos, baseado em homologia e baseado em aprendizado de máquina, foi realizada e uma lista das seis proteínas melhor ranqueadas foi apresentada.

Palavras-chave: Integração de Dados; Alvos para Fármacos; *S. mansoni*; Aprendizado de Máquina

ABSTRACT

Data Integration in Drug Target Detection for *Schistosoma mansoni*

Schistosomiasis caused by *Schistosoma mansoni* is a significant neglected disease for its occurrence in the world. However, there is only one 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 targets for *S. mansoni* drugs. The methodology takes an approach based on biological data integration, which is scattered across various public databases, in the use of homology and orthology concepts to identify essentiality and druggability proteins attributes. The machine learning method was also addressed to identify the essentiality attribute of *S. mansoni* proteins from the following essential and non-essential model organisms protein bases.

The intended homology-based methodology steps, which used essentiality and drugability characteristics, were achieved by obtaining a list of 15 *S. mansoni* drug target candidate proteins. The machine learning method indicated the *Random Forest* classifier as best performance, with 79% accuracy and 1,412 proteins reported as essential in the *S. mansoni* proteins prediction activity. The performed comparative analysis between two methods, based on homology and based on machine learning, presented a list of 6 proteins best rated.

Keywords: Data Integration; Drug Target; *S. mansoni*; Machine Learning

LISTA DE ILUSTRAÇÕES

Figura 1 – Etapas envolvidas no processo de descoberta e desenvolvimento de fármacos. Adaptado de Guido et al. (2010). 19

Figura 2 – Métodos computacionais utilizados na identificação e predição de genes essenciais. Adaptado de (Peng et al., 2017). 20

Figura 3 – Esquema resumido do ciclo biológico do parasita causador da esquistossomose. (SH, 2018) 22

Figura 4 – Estruturas homólogas entre diferentes espécies (BE, 2018). 23

Figura 5 – Exemplos de homologia, ortologia e paralogia. Adaptado de (Altenhoff and Dessimoz, 2012). 24

Figura 6 – Exemplo de estrutura primária de proteína, exibindo o conjunto de aminoácidos. 26

Figura 7 – Exemplos dos quatro níveis de estruturas das proteínas. 26

Figura 8 – Diferentes categorias de aprendizado de máquina. Adaptado de (Choy et al., 2018). 30

Figura 9 – Atividades 1, 2 e 3: busca por proteínas candidatas essenciais e drogáveis do *S. mansoni*. 42

Figura 10 – Parâmetros utilizados no método baseado em homologia. 44

Figura 11 – Atividades que compõem o método baseado em aprendizado de máquina para a busca pela característica de essencialidade do *S. mansoni*. 45

Figura 12 – Parâmetros utilizados no método baseado em aprendizado de máquina. 47

Figura 13 – Análise quantitativa da ortologia realizada entre o proteoma do *S. mansoni* e cada conjunto de proteínas essenciais dos três organismos modelo e suas interseções. 49

LISTA DE TABELAS

Tabela 1 – Trabalhos relacionados ao método baseado em homologia. 37

Tabela 2 – Lista das 10 proteínas candidatas essenciais e drogáveis do *S. mansoni* que apresentaram maiores percentuais de identidade. A=Approved, S=Small. 50

Tabela 3 – Lista final de 15 proteínas candidatas essenciais e drogáveis do *S. mansoni*. A=Approved, S=Small. 52

Tabela 4 – Comparação do desempenho de diferentes algoritmos de classificação utilizando as técnicas de sobreamostragem e subamostragem. 53

Tabela 5 – Proteínas do *S. mansoni* candidatas a essenciais de acordo com as abordagens baseadas em homologia e aprendizado de máquina. 55

Tabela 6 – Lista de 138 proteínas resultantes do processo de ortologia entre as proteínas essenciais dos organismos modelo selecionados e o proteoma do *S. mansoni*. 66

LISTA DE ABREVIATURAS E SIGLAS

AAC	*AminoAcid Composition*
ARFF	*Attribute Relation File Format*
AUC	*Area Under the Curve*
DEEPCNF	*Deep Convolutional Neural Fields*
DEG	*Database of Essential Gene*
DNA	Ácido Desoxirribonucleico
FN	*False Negative*
FP	*False Positive*
MCL	*Markov Cluster Algorithm*
MOMA	*Minimization Of Metabolic Adjustment*
OGEE	*Online GEne Essentiality*
OMS	Organização Mundial De Saúde
PPI	*Protein Protein Interactions*
PQ	Programação Quadrática
PR-AUC	*Precision Recall - Area Under the Curve*
PZQ	Praziquantel
RF	*Random Forest*
ROC	*Receiver Operating Characteristic*
ROC-AUC	*Receiver Operating Characteristic - Area Under the Curve*
SMO	*Sequential Minimal Optimization*
SMOTE	*Synthetic Minority Oversampling TEchnique*
SVM	*Support Vector Machine*
TN	*True Negative*
TP	*True Positive*

SUMÁRIO

1	**Introdução**	**14**
1.1	O Problema	14
1.2	Objetivo	15
1.3	Metodologia	15
1.4	Resultados Obtidos	16
1.5	Organização do Texto	17
2	**Referencial Teórico**	**18**
2.1	Processo de Desenvolvimento de Novos Fármacos	18
2.2	Alvos Moleculares para Desenvolvimento de Fármacos	19
2.3	*Schistosoma mansoni*	21
2.4	Homologia	22
2.5	Organismos Modelo	24
2.6	Comparação de Sequências	25
2.7	Características Baseadas na Sequência de Proteínas	27
2.8	Integração de Dados	28
2.9	Aprendizado de Máquina	30
2.9.1	Algoritmos de Aprendizado de Máquina	31
2.9.2	Validação Cruzada	33
2.9.3	Avaliação de Predição	34
3	**Trabalhos Relacionados**	**35**
3.1	Método Baseado em Homologia	35
3.2	Método Baseado em Aprendizado de Máquina	37
3.3	Considerações	39
4	**Metodologia**	**41**

4.1 Método Baseado em Homologia 41
4.2 Método Baseado em Aprendizado de Máquina 44
4.2.1 Coleta e Integração de Dados 45
4.2.2 Transformação dos Dados 46
4.2.3 Classificação e Predição 46

5 Resultados 48
5.1 Busca por Proteínas Ortólogas - Candidatas Essenciais 48
5.2 Busca por Proteínas Homólogas - Candidatas Drogáveis 49
5.3 Busca por Proteínas Ortólogas - Essenciais ao *Homo sapiens* 51
5.4 Treinamento dos Algoritmos de Classificação 52
5.5 Predição das Proteínas do *S. mansoni* 54
5.6 Comparação entre os Métodos baseados em Homologia e em Aprendizado de Máquina 54

6 Considerações Finais 56

Referências 59

A Lista de Proteínas Candidatas Essenciais do *S. mansoni* 66

1- Introdução

1.1- O Problema

A esquistossomose é uma doença tropical 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 *Schistosoma mansoni*, causador da maioria das infecções em humanos, é a única espécie do gênero descrita no Brasil (Souza et al., 2011). Estudos comparativos com cepas encontradas no Brasil, apresentaram como resultado que diferentes cepas do *S. mansoni* possuem diferenças acentuadas quanto à sensibilidade aos esquistossomicidas usados (Katz et al., 1973). 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).

Na ausência de uma vacina para o combate à doença, a estratégia de controle da esquistossomose depende de um único medicamento recomendado pela OMS, o Praziquantel (PZQ), que tem sido usado na prática clínica há quase quatro décadas (Neves et al., 2015). No entanto, devido à alta incidência de reinfecção, o uso 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.

1.2- Objetivo

Diante do cenário apresentado, o objetivo deste trabalho é identificar possíveis novos alvos (proteínas) para fármacos de *S. mansoni* no combate à doença. Para atingir este objetivo, esta pesquisa foca na integração dos dados biológicos dispersos em diferentes bancos de dados públicos associada com a identificação dos atributos de essencialidade e drogabilidade das proteínas (conceitos melhor detalhados na seção 2.2). A metodologia aplicada para o desenvolvimento da pesquisa faz uso de abordagens baseadas em homologia e em aprendizado de máquina. Considerando o atributo de essencialidade das proteínas, uma comparação entre resultados obtidos pelas duas abordagens foi conduzida.

Como resultado, uma lista de proteínas que tenham características dos atributos de essencialidade e drogabilidade é apresentada. Tais proteínas são identificadas como candidatas a novos alvos para o desenvolvimento de fármaco no combate ao *S. mansoni*. Desta forma, cada proteína pode ser então utilizada para validação experimental em bancada biológica[1] no intuito de agilizar o processo inicial de desenvolvimento de novos fármacos.

1.3- Metodologia

Esta pesquisa emprega uma metodologia de trabalho composta de duas etapas principais. A primeira etapa utiliza um método baseado em homologia, o qual faz uso de dados de organismos modelo, para identificação de proteínas candidatas essenciais e drogáveis do *S. mansoni*. A segunda etapa utiliza um método baseado em aprendizado de máquina para, a partir dos exemplos conhecidos de proteínas essenciais e não-essenciais de organismos modelo, possibilitar a predição do atributo de essencialidade no universo das proteínas do organismo de estudo.

A etapa da metodologia baseada em homologia visa a criação de uma base de

[1] Tipo de experimento científico que investiga os efeitos de uma substância em um órgão isolado ou em um organismo vivo.

dados integrada com informações do organismo estudado. Nesta etapa é aplicado um método baseado em ortologia e homologia no qual são utilizados:

- Dados validados em outros organismos, como os organismos modelo, para trabalhar o atributo da essencialidade;
- Dados sobre alvos para fármacos já desenvolvidos e comercializados, para trabalhar o atributo da drogabilidade.

A etapa baseada em aprendizado de máquina utiliza os exemplos conhecidos de proteínas essenciais e não-essenciais dos organismos modelo para o treinamento de algoritmos de classificação. Posteriormente, os modelos de classificação, possibilitam a predição do atributo de essencialidade no universo das proteínas do organismo estudado. Nesta etapa, o atributo da drogabilidade não foi considerado, por não se tratar de um atributo capaz de ser deduzido a partir das sequências das proteínas do organismo estudado.

1.4- Resultados Obtidos

Como resultados obtidos da primeira etapa baseada em homologia, foi encontrada uma lista de 15 proteínas do *S. mansoni* candidatas a alvos para fármacos, por possuírem as características de essencialidade e drogabilidade. Na segunda etapa do trabalho, os algoritmos de classificação foram utilizados para aprendizado em relação à característica de essencialidade das proteínas, utilizando as bases de proteínas essenciais e não-essenciais conhecidas dos organismos modelo. O algoritmo *Random Forest* apresentou o melhor desempenho na atividade de classificação, com um percentual de 79% de acurácia. Em seguida o modelo de classificação gerado a partir do *Random Forest*, foi utilizado para realizar a predição da característica de essencialidade no proteoma[2] do *S. mansoni*, resultando em uma lista de 1.412 proteínas do organismo estudado, indicadas como essenciais. O genoma total do organismo possui 25.599 proteínas.

No prosseguimento da pesquisa, foi realizada uma comparação dos resultados obtidos entre os dois métodos aplicados, considerando o atributo de essencialidade.

[2] É o conjunto de proteínas expressas em resposta a definidas condições ambientais ou temporais de um sistema.

No método baseado em homologia foram indicadas 138 proteínas como candidatas a essenciais do *S. mansoni*, enquanto no método baseado em aprendizado de máquina 1.412 proteínas foram indicadas como essenciais, ocorrendo 52 proteínas em comum entre os dois métodos estudados. As bases de dados utilizadas neste trabalho, assim como os resultados obtidos, estão publicados para eventuais consultas[3].

1.5- Organização do Texto

Além desta introdução, este trabalho está organizado nos seguintes capítulos: O capítulo 2 aborda o referencial necessário para o entendimento, embasamento e conceituação deste trabalho, apresentando os termos biológicos e de computação que norteiam a pesquisa. O capítulo 3 apresenta a relação de trabalhos que utilizam metodologias semelhantes para condução de pesquisas de busca de alvos para fármacos, e de utilização de algoritmos de aprendizado de máquina na identificação de proteínas essenciais, citando suas principais caraterísticas, semelhanças e distinções. O capítulo 4 mostra de forma detalhada a metodologia utilizada nesta pesquisa, explicando as etapas envolvidas, os dados de entrada e saída de cada atividade e as ferramentas e algoritmos utilizados. O capítulo 5 apresenta os resultados obtidos na pesquisa, parâmetros usados nas execuções das ferramentas, critérios utilizados para a integração dos dados e cortes realizados. O capítulo 6 aponta as principais contribuições do presente trabalho e apresentando os próximos passos para a continuidade da pesquisa.

[3]https://github.com/swrg-cefetrj/S.-mansoni-Research.

2- Referencial Teórico

Este capítulo aborda os conceitos, termos e informações necessárias para o correto entendimento e condução desta pesquisa, os quais englobam conceitos biológicos e computacionais relacionados ao tema proposto.

2.1- Processo de Desenvolvimento de Novos Fármacos

O processo de descoberta e desenvolvimento de fármacos, conforme ilustrado na Figura 1, é dividido em duas grandes fases: (i) descoberta, também conhecida como pré-clínica ou pesquisa básica (marcada em azul na figura) e (ii) desenvolvimento ou clínica (marcada em amarelo na figura). A fase de descoberta é composta pelas atividades de pesquisa para identificação do alvo a ser utilizado, enquanto a fase de desenvolvimento é caracterizada pelas atividades de experimentação em organismos vivos (Guido et al., 2010).

Nos estágios iniciais da fase de descoberta, as pesquisas se concentram geralmente na identificação e otimização de moléculas pequenas, o alvo molecular, capazes de representar novas entidades químicas com potencial de desenvolvimento clínico. A validação do alvo molecular selecionado é fundamental por uma série de razões que envolvem desde o estabelecimento de sua relevância no processo fisiopatológico em estudo até a caracterização do impacto de sua modulação seletiva no tratamento ou na cura de doenças ou disfunções em humanos (Guido et al., 2010). Por ser um processo longo, complexo e oneroso, atividades que possam reduzir o tempo e o custo envolvidos são importantes. O presente trabalho tem seu foco no início do fluxo apresentado, primeira atividade destacada - concepção do projeto e abordagem do problema (*in silico* e *in vitro*).

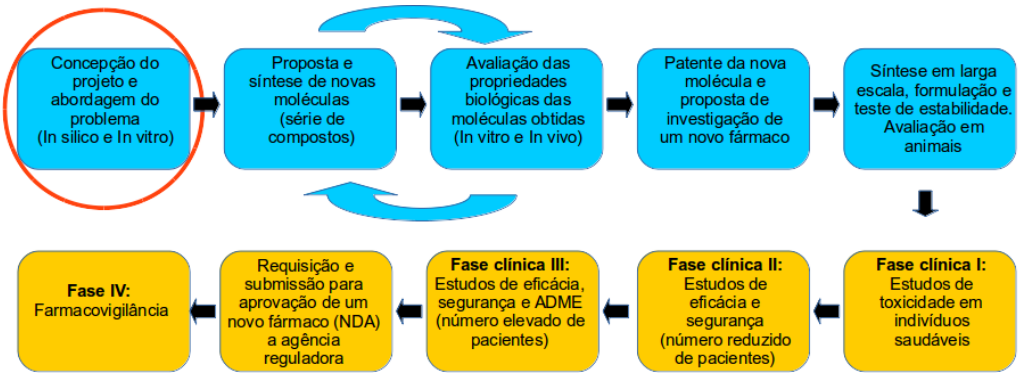

Figura 1 – Etapas envolvidas no processo de descoberta e desenvolvimento de fármacos. Adaptado de Guido et al. (2010).

2.2- Alvos Moleculares para Desenvolvimento de Fármacos

Um alvo molecular é uma proteína, peptídeo ou ácido nucleico com atividade que pode ser modulada por um composto (Gashaw et al., 2011). A descoberta de fármacos baseada em alvo é uma técnica comumente utilizada, onde o foco da pesquisa é direcionado ao estudo do conjunto de biomoléculas do patógeno (Crowther et al., 2010). Diferente das técnicas que estudam as células do organismo doente, neste caso é detalhado o causador da doença. A pesquisa de fármacos baseado em alvo se apoia na descoberta de características consideradas desejáveis em um alvo molecular para desenvolvimento de fármacos, ou simplesmente alvo para fármaco, como a essencialidade, a drogabilidade, a especificidade/seletividade e a importância das fases do ciclo de vida do patógeno relevantes para a saúde humana (Crowther et al., 2010). Para estas pesquisas, custos podem ser reduzidos devido à possibilidade de realização de algumas experiências laboratoriais a partir da utilização de análises *in silico*[1]

A essencialidade se refere aos genes essenciais de um organismo, que são genes indispensáveis para sustentar a vida celular, os quais constituem um conjunto mínimo de genes requerido para uma célula viva. As funções codificadas por genes essenciais são consideradas como uma fundação da vida. Sendo assim, a análise de genes essenciais pode ajudar a responder à questão de quais são as funções básicas necessárias para sustentar a vida celular (Zhang et al., 2004). Os produtos gênicos (ou seja, as proteínas) essenciais compreendem excelentes alvos para fármacos. Dessa

[1] Expressão usada no âmbito da simulação computacional e áreas correlatas para indicar algo ocorrido "em ou através de uma simulação computacional".

maneira, faz-se importante pesquisar sobre as proteínas essenciais daqueles organismos em que ainda há carência deste conhecimento, como o *S. mansoni*.

Na busca pela característica de essencialidade das proteínas, diversas pesquisas foram desenvolvidas e, por consequência, diversas bases de dados foram geradas e métodos computacionais baseados em diferentes abordagens foram testados. Na Figura 2 é apresentado um cenário com as principais bases de dados, atributos e abordagens computacionais utilizadas com o objetivo de identificar genes/proteínas essenciais. Na presente pesquisa, as bases de dados DEG e OGEE exibidas na Figura 2 foram utilizadas assim como os atributos baseados em homologia e baseados nas sequências. A presente pesquisa também aplicou as abordagens baseada em homologia e aprendizado de máquina.

Em relação aos atributos usados para predição da característica de essencialidade, os atributos baseados em homologia são aqueles relacionados à taxa de evolução dos organismos, dados de duplicação genética, entre outros. Os atributos baseados na sequência são dados como tamanho da proteína, uso de códon, uso de aminoácido, parâmetro Z-curve, entre outros. Os atributos dependentes do contexto são dados como interação proteína-proteína, localização da proteína, expressão genética, entre outros.

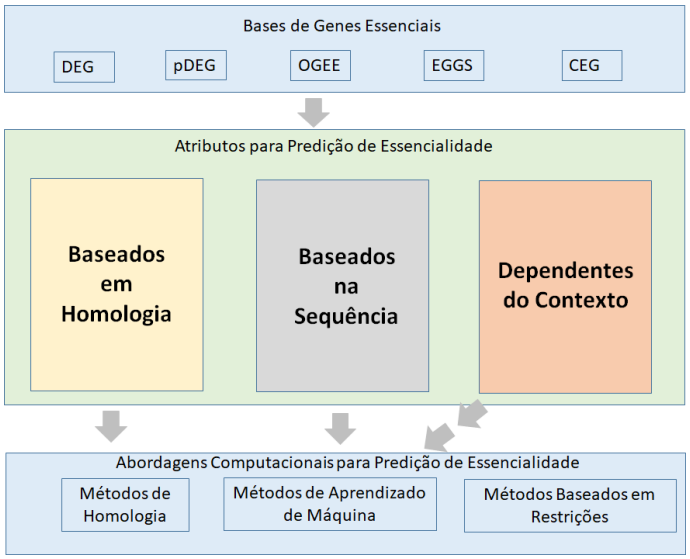

Figura 2 – Métodos computacionais utilizados na identificação e predição de genes essenciais. Adaptado de (Peng et al., 2017).

A drogabilidade é a característica que define a capacidade de um alvo ser modulado por moléculas pequenas que são semelhantes a fármacos, ou seja, se as moléculas semelhantes a fármacos são suscetíveis de interagir com o alvo e gerar um impacto

desejado. A avaliação antecipada de alvos é uma ação importante para o gerenciamento de portfólio de alvos na indústria farmacêutica, direcionando recursos para alvos para fármacos que têm maior probabilidade de fornecer candidatos clínicos (Hopkins and Groom, 2003).

As abordagens de descoberta de fármacos buscam, coletivamente, explorar dois possíveis conjuntos de alvos para fármacos: aqueles que foram validados em outros organismos e doenças, como os alvos de fármacos que já são produzidos e comercializados; e aqueles onde não houve validação, talvez porque são únicos aos patógenos da doença negligenciada, mas que, no entanto, têm potencial como novos sítios de ação (Crowther et al., 2010).

2.3- *Schistosoma mansoni*

O *Schistosoma mansoni* é um dos parasitas causadores da esquistossomose, organismo pertencente ao filo dos Platelmintos (SH, 2018). Durante o ciclo de vida deste parasita, representado na Figura 3, seus ovos deixam o corpo humano na urina (na esquistossomose urinária) ou nas fezes (na esquistossomose intestinal), eclodem na água e liberam larvas (miracídios) que penetram os hospedeiros intermediários, caramujos de água doce. Após várias semanas de crescimento e multiplicação, as cercárias emergem dos caramujos e penetram na pele humana durante o contato com a água contaminada. As cercárias se transformam e subsequentemente migram através dos pulmões para o fígado, onde amadurecem em vermes adultos. Esses vermes adultos se movem para as veias da cavidade abdominal ou do trato urinário. A maioria dos ovos produzidos é aprisionada nos tecidos, mas uma proporção é evacuada pelo intestino ou pela bexiga (WHO, 2018).

A esquistossomose inicialmente é uma doença assintomática, mas pode evoluir e causar graves problemas de saúde crônicos, podendo haver internação ou levar à morte. No Brasil, a esquistossomose é conhecida popularmente como “xistose”, “barriga d’água” ou “doença dos caramujos”. Os vermes, uma vez dentro do organismo da pessoa, vivem nas veias do mesentério e do fígado. A maioria dos ovos do parasita se prende nos tecidos do corpo humano e a reação do organismo a eles pode causar grandes danos à saúde.

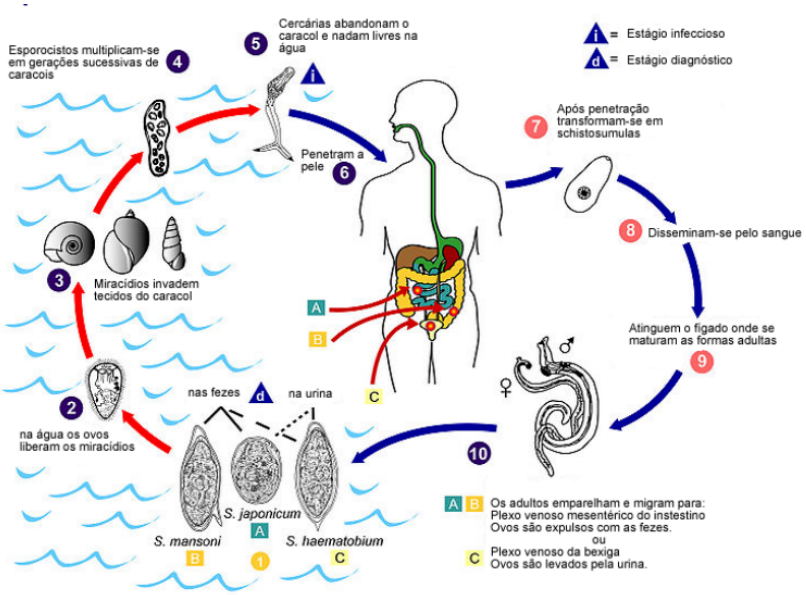

Figura 3 – Esquema resumido do ciclo biológico do parasita causador da esquistossomose. (SH, 2018)

Se não tratada adequadamente, a esquistossomose pode evoluir e provocar algumas complicações, como por exemplo: aumento do fígado, aumento do baço, hemorragia digestiva, hipertensão pulmonar e morte (MS, 2019).

2.4- Homologia

Entre as estratégias existentes relacionadas a descoberta de fármacos para tratar doenças tropicais negligenciadas, podemos citar o método baseado em homologia. Esse método é usado para inferir relações evolutivas entre organismos (Morrison et al., 2015). A homologia é a relação de ancestralidade entre duas ou mais entidades (e.g. genes). Dizer que genes, proteínas, estruturas ou posições de um alinhamento são homólogos significa dizer que os mesmos compartilham um ancestral comum (Koonin, 2005).

Na Figura 4 são apresentadas estruturas homólogas entre diferentes espécies. Pode-se observar que os ossos dos diferentes órgãos mostrados apresentam um plano estrutural semelhante, ou seja, ossos semelhantes e na mesma posição relativa, Mas, também pode-se observar que o grau de desenvolvimento é muito diferente, pois, por exemplo, o úmero do humano é o mais desenvolvido e o do morcego o menos desenvol-

vido. É notável também que os ossos dos dedos do morcego são muito mais compridos que todos os outros, pois os ossos dos dedos do morcego têm de ser revestidos por uma membrana e terão uma função completamente diferente, pois servirão para voar.

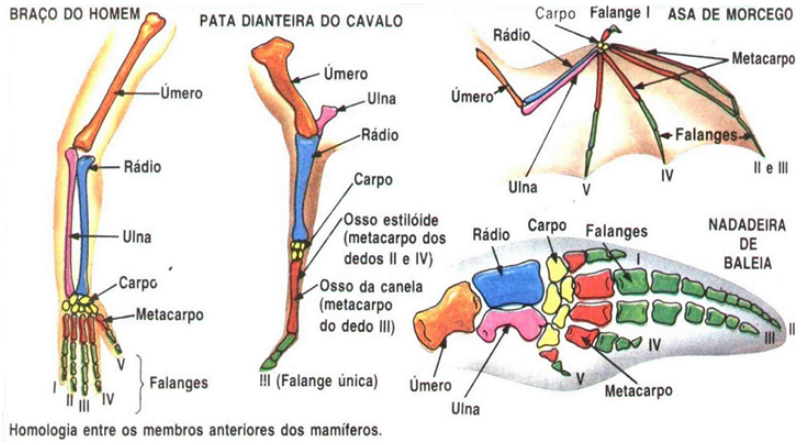

Figura 4 – Estruturas homólogas entre diferentes espécies (BE, 2018).

Dentro do conceito maior de homologia, duas classificações se destacam: ortólogos e parálogos. Ortólogos são genes relacionados via especiação (descendência vertical), enquanto os parálogos são genes relacionados via duplicação. A combinação de eventos de especiação e duplicação, juntamente com perda de genes e rearranjos genéticos, entrelaçam ortólogos e parálogos em complexas teias de relacionamentos (Koonin, 2005).

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 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).

Na figura 5 são apresentados exemplos de genes homólogos, ortólogos e parálogos. A letra S representa um evento de especiação, enquanto a letra D representa um evento de duplicação. Todos os pares de genes formados entre x_1, y_1, x_2, y_2 e z_1 são homólogos entre si. Os pares de genes (x_1, y_1) ou (x_2, z_1) são exemplos de genes ortólogos. Os pares (x_1, x_2) ou (x_1, y_2) são exemplos de genes parálogos.

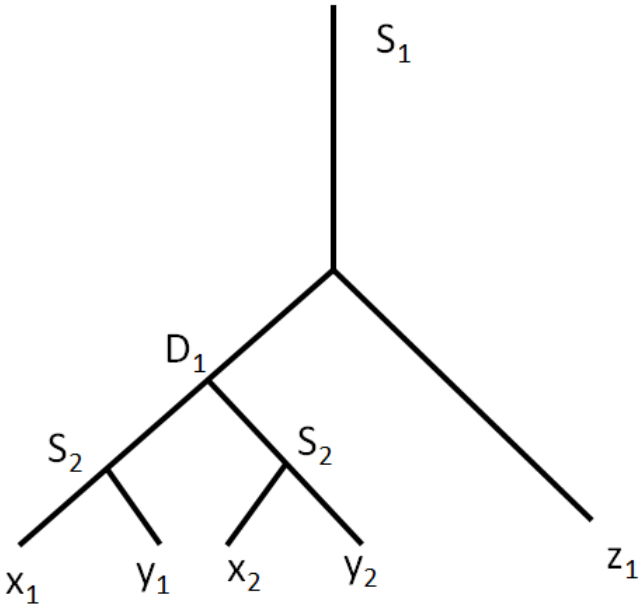

Figura 5 – Exemplos de homologia, ortologia e paralogia. Adaptado de (Altenhoff and Dessimoz, 2012).

2.5- Organismos Modelo

Os organismos modelo surgiram devido à necessidade de se conhecer de modo mais aprofundado as informações genéticas sobre os organismos. Os geneticistas gradualmente passaram a focalizar com atenção especial um pequeno número de organismos, entre eles a mosca da fruta *Drosophila melanogaster*, o camundongo *Mus musculus*, a levedura *Saccharomyces cerevisiae*, o nematódeo *Caenorhabditis elegans* e a bactéria *Escherichia coli*. Esse foco ocorreu basicamente por algumas razões: são organismos fáceis de criar, possuem tempo de geração curto, produzem grande quantidade de prole e sofrem mutações e cruzamentos com facilidade. Aliado a isso, foi constatado que muitas sequências de genes são conservadas desde a levedura até os vertebrados superiores, e esses genes frequentemente desempenham funções semelhantes em grande variedade de organismos (Palladino, 2009).

O conhecimento profundo de alguns organismos, como os organismos modelo, tem um papel vital na busca por informações de organismos pouco conhecidos. A partir da utilização dos conceitos de homologia e ortologia, é possível fazer inferência de função de proteínas de organismos ainda pouco documentados. Neste trabalho as bases

de proteínas de três organismos modelo foram utilizados na execução da metodologia proposta, a saber: *Mus musculus*, *Saccharomyces cerevisiae* e *Caenorhabditis elegans*.

2.6- Comparação de Sequências

Comparação entre sequências biológicas, ácidos nucleicos (nucleotídeos) e proteínas (aminoácidos), é uma das tarefas computacionais mais frequentes entre pesquisadores da área biológica. Sequências de nucleotídeos e aminoácidos contêm padrões e/ou *motifs* que foram preservados através da evolução porque são importantes para a estrutura ou função da molécula. A descoberta de *motifs* em proteínas e sequências de nucleotídeos pode levar a determinação da função e a elucidação das relações de evolução entre sequências, com base no conhecimento acumulado sobre outras sequências já estudadas (Brejová et al., 2000). Se as sequências comparadas são suficientemente similares entre si, infere-se que estas sequências sejam homólogas. E sendo assim, presume-se, no caso da ortologia, que elas tenham a mesma função biológica.

Se denotarmos Σ como o alfabeto de todos os possíveis caracteres que ocorrem nas sequências, teremos $\Sigma = \{$ A, C, G, T$\}$ para sequências de nucleotídeos ou sequência de Ácido Desoxirribonucleico (DNA) e $\Sigma = \{$ A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W e Y$\}$ para o conjunto dos 20 aminoácidos que compõem as sequências de proteína. Neste trabalho as pesquisas foram realizadas no âmbito das proteínas, utilizando portanto as sequências de aminoácidos dos organismos estudados.

As proteínas são organizadas em quatro níveis de estrutura: primário, secundário, terciário e quaternário. Neste trabalho foram utilizadas as informações de proteínas do nível primário. O nível primário consiste em um padrão de letras (20 aminoácidos), representando uma informação de natureza unidimensional (1D), em que a única dimensão descrita é a ordem de aparecimento dos aminoácidos. Na Figura 6 é mostrado um exemplo de estrutura primária de duas proteínas. As informações são apresentadas em um formato de texto denominado FASTA[2].

A estrutura secundária da proteína é o rearranjo espacial de segmentos locais

[2]É um formato baseado em texto para representar tanto sequências de nucleotídeos quanto sequências de peptídeos, no qual os nucleotídeos ou aminoácidos são representados usando códigos de uma única letra.

```
>Smp_105320.1:pep pep chromosome:ASM23792v2:1:1283906:1284385:
transcript_biotype:protein_coding description:60S ribosomal pr
MLTFDFIILAKRTKKVGIVGKYGVRYGASLRKTIKKIEISQHAKYNCNFCGKDSLKRKAA
GIWECKACKKVVAGGAYVCSTTAATTIRAAIRRLRDAHES
>Smp_105330.1:pep pep chromosome:ASM23792v2:1:1288747:1298221:
transcript_biotype:protein_coding description:Putative leucine
MATQYKLSSSQRDRVKKFNAITRSSDKVAIDCLQMSNWRLEQAVDYFYRQNPTPQGPTIN
EAKIDHLFQRYRDSQCPDRILATGMELFLVTDLHIDPESLITLILAWKFSAKTQGEFTRE
EFFRGFRELGCDSISSLRNKLPSLLSDIEDKQNFRSLYLFTFGFANLDKHESKSLVLQYA
IPYWEILLRGRFCHLSLWFKFLQEHHKRPISKDTWDLLLDFVETIYPDMSNYDEEGAWPV
LIDEFVEWAKPQVQMDIPLTNHP
```

Figura 6 – Exemplo de estrutura primária de proteína, exibindo o conjunto de aminoácidos.

de proteínas. Os dois elementos estruturais secundários mais comuns são as hélices alfa e as folhas beta. Elementos de estrutura secundária formam espontaneamente um rearranjo espacial, um estágio intermediário antes que a proteína se dobre em sua estrutura terciária tridimensional (Perticaroli et al., 2013). A estrutura tridimensional é importante devido ao fato de nesse momento ser caracterizada a função da proteína. Algumas proteínas são constituídas por várias cadeias polipeptídicas, também conhecidas como subunidades. Quando essas subunidades se juntam, elas conferem à proteína sua estrutura quaternária (Berman et al., 2002). A Figura 7 apresenta os quatro níveis de organização das proteínas.

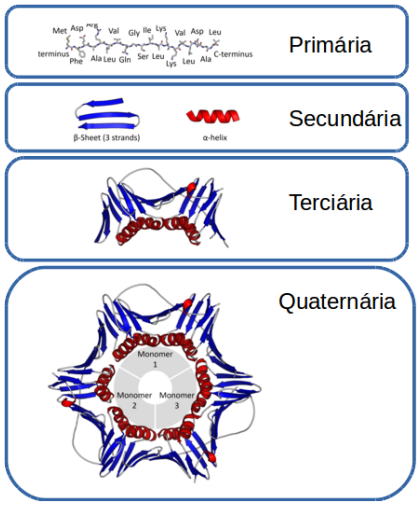

Figura 7 – Exemplos dos quatro níveis de estruturas das proteínas.

2.7- Características Baseadas na Sequência de Proteínas

Vários métodos experimentais podem ser usados para identificação de genes essenciais, tais como: mutação aleatória sistemática, interferência de RNA, nocaute genético único, nocaute condicional e mutagênese de transposão (Nigatu et al., 2017). No entanto, todos estes métodos necessitam de validação experimental em bancada biológica. Contudo, estes processos são demorados e dispendiosos, o que favorece a busca por métodos alternativos para identificação sistemática de genes essenciais. Considerando a importância dos genes essenciais e a limitação de métodos experimentais, realizar a predição de genes essenciais *in silico*, apresenta-se como uma atividade de suma importância na redução de tempo e custo das pesquisas.

Na condução de experimentos *in silico*, as características baseadas nas sequências surgem como aliadas na pesquisa pelo atributo de essencialidade. Tais características são aquelas que podem ser obtidas diretamente do DNA ou das sequências de proteínas, levando em conta unicamente a composição das sequências primárias, ou seja, sequências de nucleotídeos e sequências de proteínas. Características como: uso de aminoácidos (as frequências de 20 aminoácidos em cada proteína, 20 características), uso de códons (as frequências de códons em todos os genes, 64 características), uso de nucleotídeos de 3 posições de códon (as frequências de 4 nucleotídeos em 3 posições de códon de cada gene), uso de di-nucleotídeos (as frequências do uso de nucleotídeos de 2 tuplas de 2 códons em 3 posições de códon de cada gene) (Ning et al., 2014), entre outras, são algumas das características baseadas nas sequências, utilizadas nos experimentos *in silico*.

A medida *AminoAcid Composition* (AAC) é a fração de cada aminoácido dentro de uma sequência de proteínas e tem sido amplamente utilizada para prever diferentes características das proteínas (Kumar et al., 2017). A fração de cada aminoácido é calculada usando a fórmula 2.1.

$$Acomp(i) = [\frac{NumTotaldeumAminoacido(i)}{NumTotaldeAminoacidosnaProteina(p)}] \times 100 \tag{2.1}$$

onde Acomp(i) é a porcentagem do aminoácido i em uma determinada proteína e (i) pode ser qualquer aminoácido de uma proteína específica (p).

Nesta pesquisa, a característica AAC foi utilizada para classificação das proteínas e predição da essencialidade no método baseado em aprendizado de máquina, como o trabalho de (Ning et al., 2014) citado na seção 3.2.

2.8- Integração de Dados

O grande número de bases de dados disponíveis publicamente com informações biológicas, tem beneficiado pesquisadores de todo o mundo, especialmente aqueles de países de baixa renda. No entanto, estas informações encontram-se fragmentadas, dispersas por inúmeras bases de dados, sendo redundantes em muitas circunstâncias. Claramente existe a necessidade de unificação e tratamento dos diversos níveis de informação, utilizando-se de técnicas de manipulação e transformação de dados, como a análise de sequências, a mineração de texto e a mineração de dados, a fim de proporcionar uma visão global e integral dos sistemas biológicos estudados.

O esforço de realizar integração de dados neste contexto de muitas informações disponíveis de forma dispersa vai desde a obtenção dos dados específicos necessários, relacionados aos organismos estudados em bases curadas, avaliação das chaves de identificação de uma base que possam ser relacionadas as outras bases, passando pela análise de informações redundantes, seleção dos atributos desejados até processos de limpeza e definição de critérios de corte para o objetivo a ser atingido. Como exemplo, no presente trabalho foi necessária a utilização de quatro bases de dados distintas, cada uma provendo um tipo de informação específica com atributos específicos. As informações entre as bases de dados foram integradas a partir da existência de identificadores comuns entre as bases.

O desafio nesta área de pesquisa está relacionado à análise da grande quantidade de dados disponíveis de forma dispersa e em diferentes fontes e formatos (Galperin et al., 2017). Neste contexto, as ferramentas e técnicas computacionais para análise e integração de dados, surgem como uma solução para relacionar e integrar as informações disponíveis, de forma a responder as questões existentes. Na presente pesquisa, por exemplo, cuja pergunta se refere a identificação de possíveis alvos para fármacos do *S. mansoni*, foram utilizados dados sobre o proteoma do patógeno, dados

sobre as proteínas essenciais dos organismos modelo, dados sobre proteínas alvos de fármacos e dados sobre a característica de essencialidade das proteínas. Todas as informações foram obtidas em diferentes bases, com diferentes formatos, identificadores e atributos.

A comunidade de pesquisadores bioinformatas tem à disposição um conjunto diverso de bancos de dados com informações sobre patógenos focadas em diferentes abordagens. A seguir são apresentadas as bases de dados utilizadas nesta pesquisa.

- *Ensembl* é um banco de dados para genomas de vertebrados que apoia pesquisas em genômica comparativa, evolução, variação de sequências e regulação transcricional. O banco de dados fornece informações de anotações genéticas, calcula múltiplos alinhamentos, prevê funções regulatórias e coleta dados de doenças (Zerbino et al., 2017). Nesta base foram coletados os proteomas em formato FASTA dos organismos modelo selecionados e do *S. mansoni* (cepa NMRI, de origem de Porto Rico (Protasio et al., 2012)).

- *Database of Essential Gene (DEG)* hospeda registros de entidades genômicas essenciais, como genes codificadores de proteínas e RNAs não codificantes, entre bactérias, arqueias e eucariotas (Zhang et al., 2004). Neste banco de dados foram obtidas as bases em formato FASTA das proteínas essenciais dos organismos modelo selecionados (*Mus musculus*, *Saccharomyces cerevisiae* e *Caenorhabditis elegans*).

- *Online GEne Essentiality (OGEE)* é um banco de dados que armazena genes essenciais e não essenciais testados experimentalmente, além de recursos genéticos associados, como perfis de expressão, status de duplicação, conservação entre espécies, origens evolutivas e envolvimento no desenvolvimento embrionário (Chen et al., 2017). Neste banco de dados foi obtida a lista em formato XLS com a identificação das proteínas essenciais e não-essenciais dos organismos modelo selecionados.

- *DrugBank* é um banco de que contém informações sobre fármacos e alvos para fármacos (Wishart et al., 2017). Neste banco de dados foi obtida a base em formato FASTA de proteínas de fármacos já desenvolvidos e aprovados comercialmente, filtrados pelas categorias *Approved* e *Small*.

2.9- Aprendizado de Máquina

O aprendizado de máquina é um método da ciência de dados que disponibiliza computadores com a capacidade de aprender sem ser programado com regras explícitas. Aprendizado de máquina permite a criação de algoritmos que podem aprender e fazer previsões. Ao contrário dos algoritmos baseados em regras, o aprendizado de máquina se aproveita do aumento da exposição de grandes e novos conjuntos de dados e da capacidade de melhorar e aprender com a experiência (Choy et al., 2018).

As tarefas de aprendizado de máquina são normalmente classificadas em três grandes categorias, dependendo do tipo de tarefa: supervisionada, não supervisionada e aprendizado por reforço, como mostrado na Figura 8 (Deo, 2015).

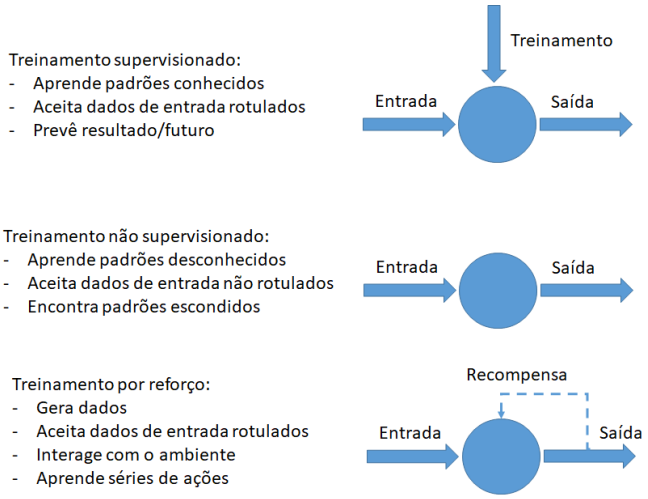

Figura 8 – Diferentes categorias de aprendizado de máquina. Adaptado de (Choy et al., 2018).

No aprendizado supervisionado, rótulos de dados são fornecidos ao algoritmo na fase de treinamento (há supervisão no treinamento). Os resultados esperados são geralmente rotulados por especialistas humanos e servem como base para o algoritmo. O objetivo do algoritmo é geralmente aprender uma regra geral que mapeia entradas para saídas. No aprendizado de máquina, a verdade básica se refere para os dados assumidos como verdadeiros (Choy et al., 2018).

No aprendizado não supervisionado, nenhum rótulo de dados é fornecido ao algoritmo de aprendizado. O objetivo da tarefa de aprendizado de máquina é encontrar a estrutura oculta nos dados e separá-los em agrupamentos (Choy et al., 2018).

No aprendizado por reforço, um programa de computador executa uma certa tarefa em um ambiente dinâmico no qual recebe *feedback* em termos de reforço positivo e negativo (como: Como jogar um jogo contra um oponente) (Kumar et al., 2012). Aprendizado por reforço aprende com as consequências das interações com um ambiente sem ser explicitamente ensinado. Um paradigma de aprendizado de máquina pode usar uma combinação de métodos supervisionados e não supervisionados com um *loop* de *feedback* de reforço (Choy et al., 2018).

Em geral, as técnicas de aprendizado de máquina são desenvolvidas usando um sistema de treino e teste (Choy et al., 2018). Três conjuntos principais de dados para treinamento, testes e validação são idealmente necessários. O conjunto de dados de treinamento é usado para ajustar o modelo. Durante o treinamento, o algoritmo aprende a partir dos exemplos. O conjunto de validação é usado para avaliar se diferentes modelos se encaixam nos dados separados e para ajustar os parâmetros do modelo.

A maioria das abordagens de treinamento tendem a superestimar os dados de treinamento, o que significa que são encontrados relacionamentos que se encaixam bem no conjunto de dados de treinamento, mas não necessariamente suportam os dados no geral. Portanto, iterações sucessivas de treinamento e validação devem ser realizadas para otimizar o algoritmo e evitar o ajuste excessivo. No conjunto de testes, após a classificação inicial realizada no aprendizado de máquina, o ajuste final do modelo pode ser aplicado a um conjunto de dados de teste independente, para avaliar o desempenho, precisão e generalização do algoritmo.

2.9.1 Algoritmos de Aprendizado de Máquina

Algoritmos de aprendizado de máquina são extensivamente usados em tarefas de predição de estrutura e funções de proteínas (Si et al., 2015). Estes algoritmos se apresentam como uma abordagem plausível na construção de modelos preditivos em ambientes onde dados de treinamento validados experimentalmente encontram-se disponíveis. Nesta pesquisa quatro algoritmos de classificação de dados foram usados, a saber, *Random Forest, J48, SMO e Logistic*. Tais algoritmos foram selecionados a partir da utilização da funcionalidade Explorer da ferramenta Weka, que indica os algoritmos

mais adequados para o tipo de dado utilizado.

Random Forest

O algoritmo de *Random Forest* (RF) introduzido por Breiman (2001) é um termo geral para métodos de *ensemble* (combinação dos resultados de vários classificadores) utilizando classificadores do tipo árvore (Breiman, 2001). O RF constrói uma grande quantidade de árvores de decisão a partir de sub-conjuntos de dados do conjunto original de treinamento definido. Tal treinamento é realizado usando *bagging* (um algoritmo para melhorar os modelos de classificação e regressão de acordo com a estabilidade e precisão da classificação). A utilização de *bagging* no treinamento além de reduzir a variância ajuda a evitar o *over-fitting*. Este procedimento extrai casos aleatoriamente a partir de conjuntos de dados de treinamento originais os quais são usados para construir cada uma das árvores de decisão que compõe a RF. Cada árvore classificadora é apontada como um componente preditor. O algoritmo RF constrói sua decisão por meio da contagem dos votos dos componentes preditores em cada classe e, em seguida, seleciona a classe vencedora em termos de número de votos acumulados (Son et al., 2009).

J48

O J48 é um indutor *top-down* de árvores de classificação que reimplementa o algoritmo C4.5, proposto por Quinlan e colaboradores (Kohavi and Quinlan, 2002). A seleção da melhor partição dos nós e o critério de parada são baseados na entropia de Shannon, como é usual em parte da família de indução de árvores de classificação (Mitchell et al., 1997). O J48 também possui uma fase de pós-poda da árvore após a expansão, na qual são convertidas para folhas as sub-árvores que não representam ganhos de informação acima de um limiar especificado (Kohavi and Quinlan, 2002). O algoritmo é capaz de lidar com classes binárias, nominais e valores faltantes de classe.

SMO

O *Sequential Minimal Optimization* (SMO) é um algoritmo para treinamento de máquinas de suporte vetorial, proposto por Platt (1998). O SMO se propõe a resolver o problema de Programação Quadrática (PQ) do *Support Vector Machine* (SVM) sem o armazenamento de matrizes extras e sem o uso de métodos de solução numérica de PQ (Platt, 1998). A solução proposta pelo SMO é a decomposição do problema de PQ global em uma série de sub-problemas e então escolher o menor problema de otimização possível para resolver em cada etapa. Estes pequenos sub-problemas são então solucionados de forma analítica, evitando assim o uso de métodos numéricos de otimização, computacionalmente mais intensivos.

Logistic

Na classificação usando regressão logística o desafio é encontrar parâmetros de melhor ajuste para uma função não-linear chamada sigmóide. São usados métodos de otimização para regredir nos dados e encontrar os parâmetros mais adequados (Harrington, 2012).

2.9.2 Validação Cruzada

A validação cruzada *k-fold* é um procedimento muito usado para estimar o desempenho de um algoritmo de classificação ou para comparação do desempenho entre dois algoritmos de classificação em um conjunto de dados. Este procedimento divide aleatoriamente um conjunto de dados em k partes disjuntas com tamanho aproximadamente igual e cada uma das partes é usada para testar o modelo induzido pela outra k-1 parte por um algoritmo de classificação. O desempenho da classificação feita pelo algoritmo é avaliado pela média das k exatidões resultantes da validação cruzada *K-fold* e, portanto,

o nível da média é assumido (Wong, 2015).

2.9.3 Avaliação de Predição

O sistema de previsão de informações é avaliado usando medidas como, acurácia, precisão e *recall* (Kumar et al., 2009). Essas medições são expressas em termos de *True Positive* (TP), *False Negative* (FN), *True Negative* (TN) e *False Positive* (FP). As medições são apresentadas nas fórmulas 2.2, 2.3 e 2.4.

$$Acurácia = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{2.2}$$

$$Precisão = \frac{TP}{(TP + FP)} \tag{2.3}$$

$$\textit{Recall} = \frac{TP}{(TP + FN)} \tag{2.4}$$

A acurácia é a medida que define o quão frequente o classificador está correto. A precisão é a medida que define, do montante dos itens do grupo classificado como verdade, quantos efetivamente são verdade. Por fim, o *recall* define quando um item realmente é da classe X, o quão frequente o algoritmo o classifica como X de fato.

3- Trabalhos Relacionados

O estudo para identificação de trabalhos relacionados à presente pesquisa, teve como foco buscas por trabalhos de identificação de alvos para fármacos de doenças tropicais negligenciadas, utilizando abordagens como: análises *in silico*, homologia, busca de atributos de essencialidade e drogabilidade, e utilização de algoritmos de aprendizado de máquina na busca pela característica de essencialidade das proteínas. Os trabalhos relacionados são apresentados a seguir considerando o método aplicado, seja baseado em homologia ou baseado em aprendizado de máquina.

3.1- Método Baseado em Homologia

Modelagem baseada em homologia é uma metodologia usada para prever informações da estrutura proteica baseada na observação geral de que proteínas com sequências de aminoácidos semelhantes têm estruturas semelhantes (Cavasotto and Phatak, 2009). Esforços para identificar genes essenciais computacionalmente adotando análises de genômica comparativa são primordiais para as pesquisas pela informação de essencialidade dos genes. A ideia para o método de mapeamento por homologia é simples, genes compartilhados por organismos relacionados de forma distante, são suscetíveis de ser essenciais (Peng et al., 2017).

Todos os trabalhos citados nesta seção utilizam análise *in silico* para identificar alvos para fármacos, o que os distingue são os atributos e as abordagens utilizadas nesta identificação. 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 (Caffrey et al., 2009).

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 apresentado como essa questão pode ser parcialmente vencida utilizando-se do mapeamento de dados de genes homólogos em organismos bem estudados (Crowther et al., 2010).

Florez e colaboradores (2010) previram a rede de interação de proteínas de *Leishmania major* usando três métodos validados: PSIMAP, PEIMAP e iPfam. A partir da combinação dos resultados destes métodos foi calculada uma rede de alta confiança para prever o processo biológico das proteínas e dos agrupamentos detectados. Utilizando a análise topológica desta rede foram detectados possíveis candidatos a alvos para fármacos (Flórez et al., 2010).

Paul e colaboradores (2014) utilizaram a rede metabólica da *Leishmania major* para identificar por meio do processo *Minimization Of Metabolic Adjustment* (MOMA) genes letais. Aliado a isto a pesquisa se utilizou do critério de essencialidade e da filtragem de alvos através da comparação com alvos para fármacos experimentalmente validados, análise sequencial, volume dos bolsos e análise da profundidade dos bolsos (Paul et al., 2014).

Uddin e colaboradores (2018) utilizaram o proteoma não redundante de 68 cepas da bactéria *Pseudomonas aeruginosa*. A cepa PAO1 foi selecionada a partir de *Clusters of Orthologous Proteins* (COPs). Uma lista de proteínas foi pré-selecionada por não serem homólogas ao humano e essenciais para a sobrevivência do patógeno. Em seguida foram identificadas as proteínas hipotéticas conservadas e suas funções foram caracterizadas. As proteínas drogáveis foram selecionadas e os parceiros de interação foram identificadas a partir de investigação da rede de interação proteína-proteína. As proteínas hub[1] foram selecionadas e propostas como potenciais novos alvos de fármacos (Uddin and Jamil, 2018).

A Tabela 1 apresenta um resumo dos trabalhos relacionados à etapa da metodologia de identificação dos atributos de essencialidade e drogabilidade das proteínas a partir

[1] Dentro de uma rede de interação de proteínas, cada proteína representa um nó. Uma proteína hub é aquela com mais de dez interações com outros nós, ou seja, outras proteínas.

do método baseado em homologia.

Tabela 1 – Trabalhos relacionados ao método baseado em homologia.

Trabalho	Organismos estudados	Atributos utilizados	Abordagem utilizada
Caffrey e colaboradores (2009)	*S. mansoni*	Essencialidade	Química comparativa e ortologia entre organismos modelo
Crowther e colaboradores (2010)	*L. major, P. falciparum, M. tuberculosis, T. brucei, T. cruzi, B. malayi, S. mansoni*	Anotação, Estrutura, Filogenia, Essencialidade, Drogabilidade e Testabilidade	Uso das diversas bases de dados disponibilizadas no TDRtargets.org e homologia para identificar drogabilidade dos helmintos
Florez e colaboradores (2010)	*L. major*	Essencialidade	Construção da rede de interação proteica do parasita
Paul e colaboradores (2014)	*L. major*	Essencialidade, Drogabilidade, Testabilidade, Interações epistáticas, Peso molecular, Seqüência, Semelhança estrutural com o proteoma humano e Microbioma	*Minimization Of Metabolic Adjustment* (MOMA) para identificar genes letais
Uddin e Colaboradores (2018)	*P. aeruginosa*	Essencialidade e Drogabilidade	Análise proteômica central de todas as cepas disponíveis, aplicação de genômica subtrativa computacional e mineração de interação proteína-proteína para encontrar proteínas hub como alvos potenciais de fármacos

3.2- Método Baseado em Aprendizado de Máquina

Em relação ao método de busca do atributo de essencialidade a partir de algoritmos de aprendizado de máquina, os trabalhos a seguir foram utilizados como referência. Todos possuem como característica comum, a utilização de atributos baseados nas sequências primárias de genes/proteínas e a utilização de algoritmos de classificação para o aprendizado e predição do atributo de essencialidade dos organismos estudados.

Kumar e colaboradores (2009) utilizaram sequências de proteínas codificadas por 116 atributos, onde os 20 aminoácidos foram caracterizados em 10 grupos funcionais baseados na composição da cadeia do grupo químico, além de atributos baseados em tripeptídeos. Foram montadas matrizes contendo valores quantitativos para as propriedades físico-químicas dos aminoácidos. Para o aprendizado foi utilizado um método de classificação *Random Forest*, denominado DNA-Prot, para identificar proteínas de ligação ao DNA a partir da sequência protéica. Como método comparativo, o algoritmo SVM foi usado com as mesmas bases de dados, obtendo um desempenho inferior ao algoritmo *Random Forest*, 84,37% contra 71,87% (Kumar et al., 2009) .

Na pesquisa de Ning e colaboradores (2014) o foco foi dado à predição de genes essenciais de bactérias, utilizando classificador SVM e demostrando a viabilidade do uso de atributos relacionados às sequências primárias (genes ou proteínas) para a predição de genes essenciais. Entre os atributos utilizados na pesquisa, encontram-se a frequência de aminoácidos em cada gene, a frequência de códons em cada gene, a frequência de nucleotídeos, a frequência de dinucleotídeos e atributos derivados do CodonW[2]. A estratégia de subamostragem foi usada para lidar com o desbalanceamento do número de amostras de dados essenciais e não essenciais, visto que o número de sequências não-essenciais é superior ao número de sequências essenciais (Ning et al., 2014).

Liu e colaboradores (2017) focaram suas pesquisas na predição de genes essenciais e na busca do melhor conjunto de atributos, utilizando um classificador SVM e analisando resultados que maximizassem o desempenho do algoritmo submetido aos dados de 31 espécimes de bactérias. Os dados foram classificados em Genes Essenciais, Genes Não Essenciais e Base completa. Os atributos que ofereceram melhor desempenho do classificador, foram identificados e mantidos nas próximas execuções, enquanto os de menor desempenho foram retirados e os resultados comparados, comprovando melhores resultados no cenário posterior (Liu et al., 2017). Dentre os atributos selecionados com melhor desempenho estão listados a ocorrência dos aminoácidos A,C,G,S,T,F,H,W,Y, devido às suas propriedades físico-químicas (Rice et al., 2000).

Nigatu e colaboradores (2017) utilizaram bases de genes essenciais e não-essenciais de 16 espécimes. Os autores utilizaram atributos relacionados a entropia, informações mútuas, informações mútuas condicionais, divergência de Kullback-Leibler e modelo de Markov, derivados da sequência de DNA. Um algoritmo para o classificador

[2](http://codonw.sourceforge.net/)

Random Forest foi implementado usando linguagem Python. A técnica de subamostragem para balanceamento dos dados foi utilizada e a medida *Receiver Operating Characteristic - Area Under the Curve* (ROC-AUC) usada para avaliação de desempenho com um valor médio de 0,84. Foi realizada uma avaliação dos atributos, sendo selecionados os 20 melhores e comprovando que atributos baseados na sequência primária dos genes podem ser utilizados para predição do atributo de essencialidade (Nigatu et al., 2017).

Campos e colaboradores (2019) avaliaram os organismos: *Saccharomyces cerevisiae, Schizosaccharomyces pombe, Caenorhabditis elegans, Drosophila melanogaster, Mus musculus e Homo sapiens* que possuíam dados curados de essencialidade e não essencialidade das proteínas. Extraíram 9.920 atributos das sequências de proteínas, utilizando linguagem R. Tais atributos eram baseados na composição dos aminoácidos, autocorrelação e propriedades químicas das sequências de proteínas. Os algoritmos *Generalised Linear Model (GLM), Artificial Neural Network (NN), Gradient Boosting Method (GBM), Support Vector Machine (SVM) e Random Forest (RF)* foram usados para predição da essencialidade das proteínas. O desempenho da previsão foi medido a partir das curvas ROC-AUC e *Precision Recall - Area Under the Curve* (PR-AUC), tendo *Random Forest* superado outros métodos na maioria dos cenários (Campos et al., 2019).

3.3- Considerações

O presente trabalho adotou um método baseado em homologia, de maneira similar aos trabalhos apresentados na seção 3.1, utilizando os atributos de essencialidade e drogabilidade para o levantamento de proteínas candidatas a alvo para fármaco de *S. mansoni*. Por meio deste método, utilizam-se 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, que faz uso das proteínas candidatas essenciais do *S. mansoni* (conhecidas a partir das proteínas ortólogas com os organismos modelos) e a partir destas, são encontradas as proteínas com características de drogabilidade, resultando assim em um conjunto de proteínas com características de essencialidade e

drogabilidade ao mesmo tempo.

A pesquisa também utilizou, similarmente aos trabalhos citados na seção 3.2, métodos de aprendizado de máquina, para aprenderem sobre as características das proteínas essenciais e não-essenciais das mesmas bases dos organismos modelos utilizadas no método baseado em homologia. Os modelos de classificação foram criados a partir destas bases e utilizados para posterior predição das proteínas do *S. mansoni* em essenciais e não-essenciais. Uma comparação entre os dois métodos de pesquisa foi viabilizada. A distinção da presente pesquisa em relação aos trabalhos citados na seção na seção 3.2, deve-se à utilização da frequência AAC *(Aminoacid Composition)* para formatação das bases de dados de proteínas e à utilização de diferentes algoritmos de classificação.

4- Metodologia

A metodologia proposta para esta pesquisa é dividida em duas etapas de trabalho. A primeira etapa faz uso de um método baseado em homologia para identificação de proteínas candidatas essenciais e drogáveis do *S. mansoni*. A segunda etapa utiliza um método baseado em aprendizado de máquina para a partir dos exemplos conhecidos de proteínas essenciais e não-essenciais dos organismos modelo, possibilitar a predição deste atributo no universo das proteínas do organismo estudado.

4.1- Método Baseado em Homologia

A primeira etapa da 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. Além disso, três organismos modelo foram selecionados para a condução desta pesquisa, por representarem categorias diferentes de evolução. Assim, a metodologia aplicada considerando suas modificações em relação ao trabalho supracitado é dividida em três atividades, denominadas integração das bases de dados.

Nesta etapa as bases de dados utilizadas no trabalho foram integradas, com o propósito de gerar a base final contendo a lista de proteínas candidatas essenciais e drogáveis. Segue descrição das atividades necessárias para a integração da base de dados.

- Atividade 1: geração da base de proteínas candidatas essenciais, a partir do conceito de ortologia;

- Atividade 2: geração da base de proteínas candidatas essenciais e drogáveis, a partir do conceito de homologia;
- Atividade 3: eliminação de proteínas ortólogas ao *Homo sapiens*;

A Figura 9 apresenta o esquema das atividades 1, 2 e 3, as quais são descritas em detalhes posteriormente.

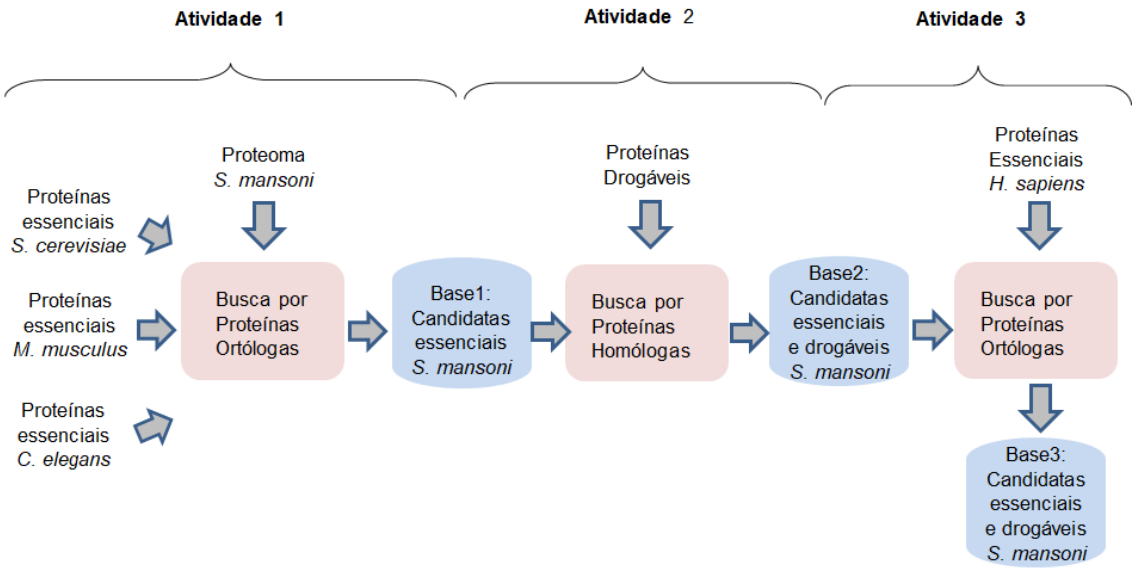

Figura 9 – Atividades 1, 2 e 3: busca por proteínas candidatas essenciais e drogáveis do *S. mansoni*.

Na **atividade 1**, é utilizado o conceito mais específico da homologia, a ortologia, para realizar 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 são submetidas a um critério de corte considerando a ocorrência repetida nos quatro organismos, ou seja, foram selecionadas as proteínas em comum aos quatro organismos. Desta maneira, é obtida a base inicial (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.

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 o 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 grafo de ortogrupos para entrada no *Markov Cluster Algorithm* (MCL) (Enright et al., 2002);
5. Agrupa genes em ortogrupos usando o MCL.

Em seguida, na **atividade 2** é 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, é considerada apenas a homologia entre as sequências, pois duas proteínas homólogas possuem alta similaridade. Logo, se uma proteína que já é considerada 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 são obtidas dos conjuntos de dados das categorias *Approved* e *Small Molecule* do banco de dados *DrugBank*. A ferramenta BLAST é utilizada para identificação das proteínas homólogas e os critérios de corte utilizados na ferramenta são melhor detalhados na seção 5. Esta atividade resultou em uma base de dados (Base 2) composta por proteínas candidatas essenciais e drogáveis do organismo estudado, representadas por sequências primárias.

A **atividade 3** realiza um processo de ortologia, similar ao realizado na atividade 1. Foram utilizadas como entradas a lista de proteínas candidatas essenciais e drogáveis do *S. mansoni* contra a lista de proteínas essenciais do *Homo sapiens*. Como saída foi gerada a Base 3, de modo a garantir que a lista de proteínas candidatas não possua nenhuma proteína essencial ao ser humano.

A Figura 10 resume quais as bases, ferramentas e *scripts* foram utilizados em cada uma das atividades do processo.

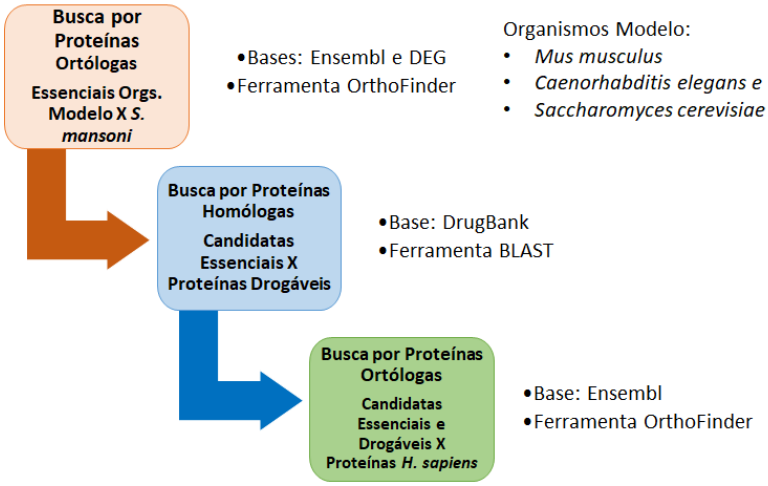

Figura 10 – Parâmetros utilizados no método baseado em homologia.

Os dados da base integrada obtida nas atividades descritas da metodologia representam as proteínas essenciais e drogáveis, candidatas a alvos para fármacos do *S. mansoni*. Importante destacar que para a presente pesquisa houve o direcionamento dos organismos modelo e do organismo de estudo específicos. No entanto, esta abordagem pode ser utilizada para quaisquer que sejam as entradas de organismos.

4.2- Método Baseado em Aprendizado de Máquina

O método baseado em algoritmos de aprendizado de máquina, propõe a identificação de proteínas do *S. mansoni* que possuam a característica de essencialidade. A utilização adicional deste método foi proposta para confrontar/enriquecer os dados inicialmente obtidos com a ortologia. Neste contexto a característica de drogabilidade não foi utilizada, por não se tratar de um atributo capaz de ser deduzido a partir das sequências das

proteínas do organismo estudado. Para tanto, as seguintes atividades são propostas e detalhadas:

- Coleta e integração de dados;
- Transformação de dados;
- Classificação e predição da característica de essencialidade.

A Figura 11 apresenta o esquema das atividades, as quais são descritas em detalhes posteriormente.

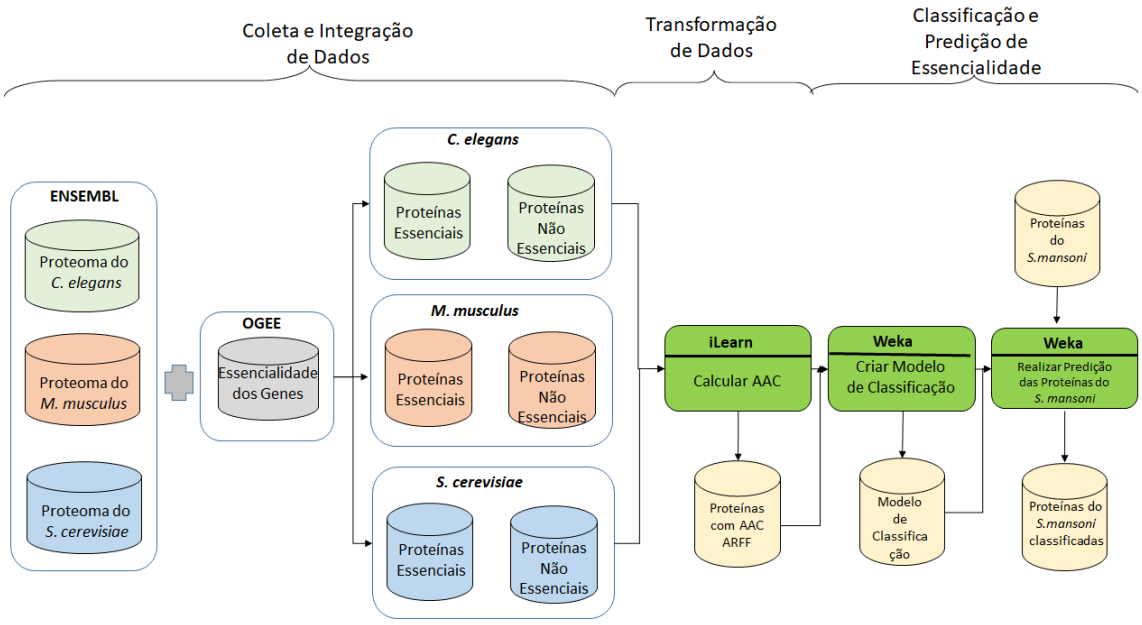

Figura 11 – Atividades que compõem o método baseado em aprendizado de máquina para a busca pela característica de essencialidade do *S. mansoni*.

4.2.1 Coleta e Integração de Dados

O desenvolvimento deste trabalho apoia-se no conhecimento dos três organismos modelo utilizados também no método baseado em homologia: *Mus musculus*, *Caenorhabditis elegans* e *Saccharomyces cerevisiae*. O proteoma de cada organismo em formato FASTA foi coletado a partir da base de dados *Ensembl*, como comentado na seção 2.8. As informações sobre a essencialidade e não-essencialidade de cada proteína desses organismos foram coletadas na base de dados *OGEE* (Chen et al., 2011).

Por meio da execução de um *script* escrito em linguagem Java, desenvolvido para esta finalidade, foi realizada uma limpeza e integração dos dados dos proteomas. As bases dos proteomas foram cruzadas aos identificadores das proteínas existentes na base OGEE, gerando duas bases resultantes para cada organismo, uma com as proteínas essenciais e outra com as proteínas não-essenciais.

4.2.2 Transformação dos Dados

Dado que a base integrada refere-se às sequências primárias de proteínas, foi necessária a transformação dos dados de modo a possibilitar a utilização dos classificadores de aprendizado de máquina. Assim, os dados foram transformados em um modelo discreto de proteínas, no qual foi realizado o cálculo das frequências de ocorrências normalizadas dos 20 aminoácidos nativos em cada proteína, denominado AAC (Nakashima et al., 1986). Para obtenção dos valores de AAC foi utilizada a plataforma iLearn (Chen et al., 2019). Como resultado, foi gerada uma base consolidada com as informações rotuladas de AAC das proteínas dos organismos modelo. Essa base, gerada no formato *Attribute Relation File Format* (ARFF), é organizada em 21 atributos, dos quais, 20 atributos são referentes a cada aminoácido e um atributo é referente à categoria essencial ou não-essencial de cada proteína.

4.2.3 Classificação e Predição

Para a criação do modelo de classificação para os dados da base transformada de acordo com a característica AAC foram aplicados os algoritmos RF, J48, SVO e *Logistic*. Todos os experimentos utilizaram a metodologia de validação cruzada com 10 grupos (10-*fold Cross-Validation*). Em seguida, cada modelo de classificação foi utilizado para a execução da tarefa de predição da característica de essencialidade das proteínas do organismo de estudo. Ao final do processo, foi gerada uma lista contendo todas as proteínas do organismo de estudo rotuladas com a característica de essencialidade ou

não-essencialidade.

A ferramenta Weka (Frank et al., 2016) foi utilizada para esta atividade. A funcionalidade *Experimenter* da ferramenta foi utilizada para a escolha dos algoritmos de classificação que apresentam melhor desempenho em relação ao tipo de dado submetido. Para tanto, a base de dados de treinamento foi carregada e testada de acordo com uma lista de diferentes algoritmos. A medida de acurácia para cada algoritmo foi analisada de forma a permitir a seleção daqueles que apresentam os melhores resultados para essa medida.

A Figura 12 resume as bases, ferramentas e *scripts* utilizados no processo.

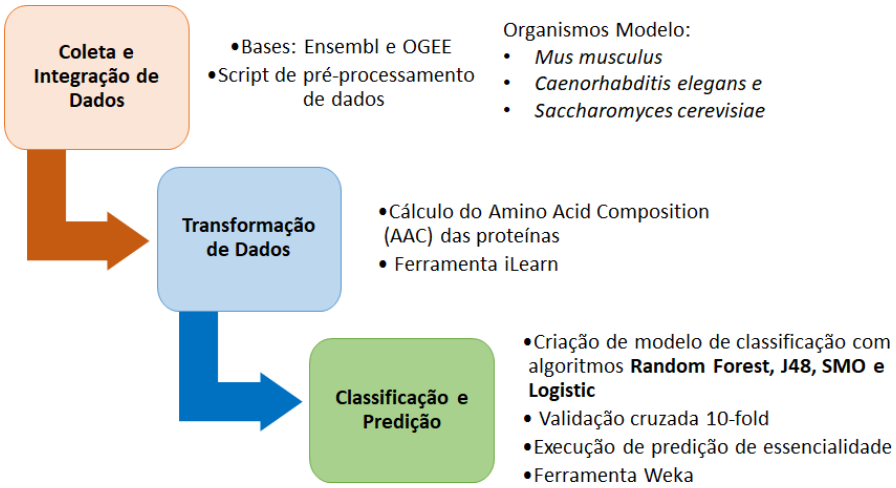

Figura 12 – Parâmetros utilizados no método baseado em aprendizado de máquina.

Semelhante ao método baseado em homologia, podemos destacar que houve o direcionamento dos organismos modelo e do organismo de estudo utilizados nesta pesquisa. No entanto, esta abordagem pode ser utilizada para quaisquer que sejam as entradas de organismos.

5- Resultados

Este capítulo apresenta os resultados desta pesquisa referentes às atividades pertencentes às duas abordagens detalhadas na metodologia nas seções 4.1 e 4.2. As bases de dados criadas nesta pesquisa assim como os resultados obtidos foram disponibilizados para consulta[1].

5.1- Busca por Proteínas Ortólogas - Candidatas Essenciais

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 pelo programa *Orthofinder*. O número de proteínas ortólogas encontradas é mostrado na Figura 13, 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.

De acordo com a Figura 13, 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. A lista com as 138 proteínas candidatas a essenciais do *S. mansoni* está apresentada no apêndice A deste trabalho. Foram exibidas as informações de identificação da proteína e sua respectiva descrição.

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. Logo, foram utilizadas as 138 proteínas (apêndice A) do *S. mansoni* também encontradas na base de proteínas es-

[1] https://github.com/swrg-cefetrj/S.-mansoni-Research.

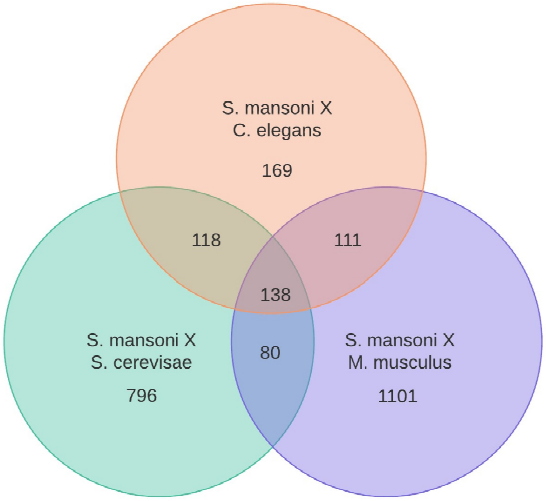

Figura 13 – Análise quantitativa da ortologia realizada entre o proteoma do *S. mansoni* e cada conjunto de proteínas essenciais dos três organismos modelo e suas interseções.

senciais dos três organismos modelo, representando assim uma maior chance de se caracterizarem como essenciais.

5.2- Busca por Proteínas Homólogas - Candidatas Drogáveis

Para a identificação das proteínas homólogas, a base de proteínas drogáveis foi composta de 7.172 sequências extraídas do banco de dados DrugBank, das quais 2.683 sequências são pertencentes à categoria *Approved* e 4.489 sequências são 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 das 7.172 sequências proteínas drogáveis.

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 foruns de discussão sobre o uso do BLAST. Os parâmetros usados realizam as seguintes restrições às combinações 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;
- *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
- *best_hit_overhang* = 0.25

Como resultado, foi obtida uma lista com 91 proteínas candidatas a alvos para fármacos. Uma lista com as 10 proteínas que apresentaram maiores percentuais (%) de identidade é mostrada na Tabela 2, 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*.

Tabela 2 – Lista das 10 proteínas candidatas essenciais e drogáveis do *S. mansoni* que apresentaram 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

Para corroborar os resultados obtidos neste trabalho, foram identificadas na literatura algumas pesquisas direcionadas às proteínas obtidas na Tabela 2. A pesquisa de Morais e colaboradores (2017), aponta a proteína Ubiquitina como um possível alvo

para desenvolvimento de fármacos para *S. mansoni*, devido a sua função importante na regulação de proteínas indesejadas (Morais et al., 2017).

A proteína calmodulina foi estudada por Neves e colaboradores (2015) que tinham como objetivos prever alvos que controlam a função muscular e motilidade do parasita. Os esquistossomas dependem do sistema muscular de motilidade para penetração na pele do hospedeiro (cercárias) e migração (esquistossômulos). Além disso, esquistossomos usam seu sistema muscular para o emparelhamento, acasalamento e reprodução, além de processos digestivos e excretórios. Essas funções são essenciais para a sobrevivência do parasita (Neves et al., 2015).

5.3- Busca por Proteínas Ortólogas - Essenciais ao *Homo sapiens*

Como o objetivo deste trabalho é a identificação de proteínas alvos para fármacos do *S. mansoni*, houve a preocupação em retirar da lista de proteínas candidatas essenciais e drogáveis, aquelas que também fossem essenciais ao *H. sapiens*. Esta retirada é importante, para que o fármaco não atinja uma proteína essencial ao ser humano, causando-lhe eventos adversos.

Para tanto, foi realizado um processo de ortologia entre as proteínas essenciais do *H. sapiens* e a lista de 91 proteínas candidatas essenciais e drogáveis do *S. mansoni*. A ferramenta *Orthofinder* foi utilizada para esta tarefa. Após a retirada das proteínas ortólogas às proteínas essenciais do *H. sapiens*, a lista final resultou em 15 proteínas candidatas a alvos para fármacos do *S. mansoni*, as quais são mostradas na Tabela 3. De forma semelhante à Tabela 2 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*. Da lista de proteínas final obtida e apresentada na Tabela 3, apenas para a proteína Ubiquitina (Morais et al., 2017) (trabalho citado previamente) já há ocorrência na literatura como sendo um possível alvo para desenvolvimento de fármacos para *S. mansoni*.

Tabela 3 – Lista final de 15 proteínas candidatas essenciais e drogáveis do *S. mansoni*. A=Approved, S=Small.

S. mansoni	**Nome da Proteína**	**ID Drugbank**	**Evalue**	**Bit Score**	**Ident%**	**A/ S**
Smp_170820.1	Multidrug resistance protein 1, 2, 3 (P glycoprotein 1, 2, 3) putative	P08183	0.0	687	40.81	A
Smp_055780.1	Smdr2	P08183	0.0	725	40.36	A
Smp_055960.1	Putative ubiquitin-conjugating enzyme E2r	P62837	3E-22	89.4	37.01	S
Smp_139340.1	Putative rab	P51151	4E-26	102	36.19	A
Smp_082820.1	Putative ubiquitin-conjugating enzyme morgue	P62837	3E-31	115	40.97	S
Smp_154320.1	VLG2 DEAD box ATP-dependent	P60842	3E-31	123	29.56	A
Smp_145190.1	RNA helicase, putative Protein kinase	P54646	1E-31	127	28.33	A
Smp_169440.1	Ubiquitin-conjugating enzyme rad6, putative	P62837	7E-37	127	40.56	S
Smp_067980.1	Ubiquitin conjugating enzyme E2, putative	P62837	6E-43	137	83.12	S
Smp_147450.1	Serine/threonine kinase	O00141	4E-49	175	44.69	A
Smp_134890.1	Putative atp-binding cassette, sub-family B,member 6, mitochondrial (Abc6)	P08183	2E-61	208	50.24	A
Smp_204050.1	Putative uncharacterized protein	O00141	6E-71	221	67.10	A
Smp_063000.1	Smdr1	P08183	7E-77	266	35.41	A
Smp_203130.1	Putative uncharacterized protein	P63261	2E-133	380	96.72	A
Smp_077860.1	Putative actin	P63261	9E-161	457	55.16	A

5.4- Treinamento dos Algoritmos de Classificação

Dando continuidade a pesquisa, as atividades que compõem o método baseado em aprendizado de máquina foram realizadas. Inicialmente, foi conduzido um experimento para realizar o treinamento dos quatro algoritmos de classificação selecionados (RF, J48, SVO e *Logistic*) e comparar seus resultados.

Posteriormente, a partir do algoritmo que apresentou o melhor desempenho, o modelo de classificação foi criado e usado para realização da predição da característica de essencialidade sobre o dados não classificados do *S. mansoni*.

Neste experimento, a base de treinamento que originalmente continha as sequências das proteínas essenciais e não-essenciais dos organismos modelo, foi transformada em um modelo discreto de dados, contendo, ao invés de sequências de aminoácidos, as frequências de ocorrências destes aminoácidos em cada proteína, o AAC. Logo, a base foi constituída por 25.599 proteínas, sendo 6.149 proteínas essenciais e 19.450 proteínas

não essenciais. Esta base transformada de acordo com a característica AAC foi submetida à execução nos quatro algoritmos de classificação.

Devido ao desbalanceamento entre o número de proteínas essenciais e não-essenciais da base de treinamento, foi observado que os resultados apresentaram um comportamento enviesado na execução inicial, ou seja, a maioria das proteínas sendo classificadas como não-essenciais. Desta maneira, foi necessária a aplicação de técnicas para realizar o balanceamento dos dados. Nesse estudo, foram aplicadas as técnicas de sobreamostragem e subamostragem e, em seguida, uma execução dos classificadores foi realizada com cada técnica.

Para o cenário de balanceamento por sobreamostragem, foi utilizado o filtro *Synthetic Minority Oversampling TEchnique* (SMOTE), usando 215% no parâmetro de taxa de crescimento. Esta ação ocasionou o crescimento da classe minoritária (proteínas essenciais), que aumentou para 19.369 instâncias, enquanto a classe majoritária possui 19.450 instâncias. Para o cenário de subamostragem, foi utilizado o filtro *SpreadSubSample*, com parâmetro *distributionSpread* igual a 1,0. Esta ação ocasionou a diminuição dos dados majoritários (proteínas não-essenciais), reduzindo-os de forma randômica para ficar com o mesmo tamanho do grupo minoritário, ou seja, 6.149 instâncias. Ambos os filtros estão disponíveis para utilização na ferramenta Weka. Os resultados dessa tarefa para cada algoritmo são apresentados na Tabela 4.

Tabela 4 – Comparação do desempenho de diferentes algoritmos de classificação utilizando as técnicas de sobreamostragem e subamostragem.

		RF	**J48**	**Logistic**	**SMO**
Sobreamostragem	**Acurácia**	0,791	0,708	0,669	0,676
	ROC-AUC	0,873	0,722	0,718	0,676
Subamostragem	**Acurácia**	0,681	0,651	0,664	0,666
	ROC-AUC	0,751	0,672	0,709	0,666

O desempenho dos algoritmos de classificação foi comparado usando as medidas de acurácia e curva ROC-AUC. A acurácia corresponde à proporção de instâncias classificadas corretamente pelos algoritmos. Este valor é a razão entre a soma das diagonais de ocorrência da matriz de confusão (*True Positive* da classe A + *True Positive* da classe B) dividido pelo número total de instâncias (Arora, 2012). A curva *Area Under the Curve* (AUC) *Receiver Operating Characteristic* (ROC) fornece uma comparação entre os valores previstos e esperados definidos como meta em uma classificação (Sahoo and Kumar, 2012). Nesse sentido, o algoritmo *Random Forest* apresentou os melhores resultados,

considerando cada técnica de balanceamento de dados. No entanto, as porcentagens de acurácia e curva ROC são maiores para sobreamostragem.

5.5- Predição das Proteínas do *S. mansoni*

Para a execução da tarefa de predição foi utilizada a base de proteínas não classificadas de *S. mansoni*, constituída de 11.774 proteínas. Inicialmente, esta base foi transformada em um modelo discreto, de acordo com a característica AAC. No arquivo resultante, em formato ARFF, o atributo da classe de essencialidade foi preenchido com o caracter de interrogação (?) para todas as intâncias. Na ferramenta Weka, esta base foi então submetida à tarefa de predição da característica de essencialidade por meio do modelo de classificação gerado pelo algoritmo *Random Forest* com balanceamento por sobreamostragem, descrito na Seção 5.4. A predição classificou 1.412 proteínas como essenciais e 10.362 proteínas como não-essenciais.

5.6- Comparação entre os Métodos baseados em Homologia e em Aprendizado de Máquina

Os resultados obtidos nos dois métodos propostos nesta pesquisa, para a busca da característica de essencialidade nas proteínas do *S. mansoni* foram comparados e apresentados como segue.

No método baseado em homologia, foi obtida uma lista com 138 proteínas candidatas a essenciais do *S. mansoni*, enquanto que no método baseado em aprendizado de máquina 1.412 proteínas do organismo foram classificadas como essenciais. A diferença grande nos volumes identificados nos dois métodos se deve ao fato de, no primeiro, foram consideradas apenas as proteínas ortólogas em comum aos três organismos modelo, e no segundo, termos um modelo estatístico no método baseado em aprendizado de máquina.

No processo de comparação entre os dois métodos foram encontradas 52 proteínas

em comum, sendo que das 10 proteínas mais bem ranqueadas no método por homologia (apresentadas na Tabela 2), seis também foram indicadas como essenciais no método por aprendizado de máquina. Tais proteínas podem ser observadas na Tabela 5. A tabela apresenta o identificador e a descrição da proteína do *S. mansoni*, o percentual de predição da proteína, obtido no método baseado em aprendizado de máquina e o ranque entre as 10 indicadas no método baseado em homologia.

Tabela 5 – Proteínas do *S. mansoni* candidatas a essenciais de acordo com as abordagens baseadas em homologia e aprendizado de máquina.

IdProteína	**Descrição**	**Predição**	**Ranque**
Smp0466001	Actin-1	0.93	5
Smp1837101	Putative actin	0.91	4
Smp1619201	Putative actin	0.82	6
Smp0182401	Cell division control protein 48 aaa family protein	0.63	9
Smp0182402	Cell division control protein 48 aaa family protein	0.66	8
Smp2031301	Putative uncharacterized protein	0.54	3

Em relação a buscas na literatura por informações sobre a característica de essencialidade das proteínas exibidas na Tabela 5, foi identificada que a proteína Actin-1 é essencial para a sobrevivência da maioria das células, sendo responsável pelo suporte mecânico e forças motrizes para movimento (Pollard and Cooper, 2009).

6- Considerações Finais

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 do método baseado em homologia, foi obtida uma lista com 15 proteínas do organismo estudado, consideradas candidatas essenciais e drogáveis para validação experimental em bancada biológica a fim de verificar possíveis novos alvos para fármacos.

Como validação do método por homologia proposto, a pesquisa foi complementada com uma avaliação por predição das proteínas do *S. mansoni*, a partir de algoritmos de aprendizado de máquina, confirmando que 6 das 10 proteínas com melhores índices de similaridade apontadas no método por homologia, também foram apontadas no método por aprendizado de máquina. O algoritmo de classificação Random Forest apresentou o melhor desempenho entre os demais avaliados com 79% de acurácia e classificando 1.412 proteínas do *S. mansoni* como essenciais.

A importância em se identificar a característica de essencialidade de genes e proteínas aliada à grande disponibilização de dados biológicos tem estimulado as pesquisas utilizando abordagens computacionais. Tais pesquisas são importantes para apoiar o processo de desenvolvimento de novos fármacos, e particularmente para doenças negligenciadas, como é o caso da esquistossomose causada pelo *S.mansoni*, devido ao pouco investimento da indústria farmacêutica com essa classe de doenças.

Como sugestão para evolução desta pesquisa, pode ser indicada a identificação de informações sobre a interação proteína-proteína (do inglês, *Protein Protein Interactions* (PPI)) do organismo de estudo, conhecimento útil no desenvolvimento de novos

fármacos, visto que as *Protein Protein Interactions* (PPI)s são vitais para manutenção da homeostase[1] celular . No trabalho de Murakami and Mizuguchi (2014), foi utilizada uma metodologia baseada em homologia para identificação das PPIs. Mais uma sugestão está relacionada à previsão da estrutura secundária de proteínas muito importante para o estudo da estrutura e função das proteínas. No trabalho de (Wang et al., 2016) foi apresentado o método *Deep Convolutional Neural Fields* (DeepCNF) para predição da estrutura secundária das proteínas.

Além disso, também é sugerida a inclusão de outras características relacionadas à sequência primária das proteínas de maneira a enriquecer o modelo de classificação gerado pelo método de aprendizado de máquina. A presente pesquisa utilizou apenas uma característica das proteínas, o AAC. Contudo, a inclusão de novas caraterísticas, aumentando assim o conhecimento sobre as proteínas, pode trazer mais insumos para os algoritmos de aprendizado de máquina, facilitando a tarefa de classificação da característica de essencialidade das proteínas. Exemplos de características englobam a frequência de pares de aminoácidos de um tamanho específico (Chen et al., 2009) e a utilização de descritores que caracterizem os aminoácidos de acordo com suas propriedades físico-químicas (Lee et al., 2011). O desafio nesta inclusão de características é melhorar o aprendizado sem contudo piorar o desempenho relacionado à complexidade dos algoritmos de classificação.

Outro trabalho futuro é a construção de um *workflow* científico (Ocaña et al., 2012) para flexibilizar a execução dos métodos propostos. Dessa forma, seria possível flexibilizar as entradas de dados, com bases de dados de vários organismos modelo, diferentes organismos de estudo, além da variação de parâmetros nas execuções das ferramentas. Com isto seria viabilizada a execução de diversas possibilidades de experimentos, enriquecendo desta forma, os resultados obtidos.

Esta pesquisa foi aprovada para apresentação em formato de pôster no X-Meeting 2017, 13º Conferência Internacional da AB3C, realizado em outubro de 2017; em formato de artigo completo no 12º Brazilian E-Science Workshop (BreSci), um dos eventos que compuseram o 38º Congresso da Sociedade Brasileira de Computação (SBC), realizado em julho de 2018 (Garcia and Belloze, 2018); e em formato de artigo curto no 12º Brazilian Symposium on Bioinformatics (BSB), realizado em outubro de 2019 (Garcia et al., 2019).

O artigo publicado no 12º BreSci intitulado, Integração de Dados na Detecção de

[1] Homeostase designa a tendência dos organismos vivos em manter constante o meio interno. Quando o organismo não consegue manter a homeostase, ocorre a doença

Alvos para Fármacos de *Schistosoma mansoni*, detalhou a metodologia utilizada nesta pesquisa, baseada no método por homologia, com a publicação dos resultados obtidos na pesquisa, referentes à lista de proteínas essenciais e drogáveis do *S. mansoni*. Tal lista pode ser utilizada como insumo para pesquisas por alvos para fármacos, reduzindo desta forma o esforço, o tempo e o custo deduzidos deste processo.

Já o artigo publicado no 12º BSB, intitulado *Identifying Schistosoma mansoni essential protein candidates based on machine learning* focou na metodologia baseada no método de aprendizado de máquina. Neste, a metodologia foi detalhada e os resultados referentes aos algoritmos de classificação utilizados foram divulgados. Este artigo também apresentou uma comparação entre os resultados obtidos pelo método baseado em homologia e pelo método baseado em aprendizado de máquina.

Referências

Altenhoff, A. M. and Dessimoz, C. (2012). Inferring orthology and paralogy. In *Evolutionary genomics*, pages 259–279. Springer.

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.

Arora, R. (2012). Comparative analysis of classification algorithms on different datasets using weka. *International Journal of Computer Applications*, 54(13).

BE (2018). Brasil escola. Disponível em: `https://brasilescola.uol.com.br/biologia/Orgaos-analogos-homologos.htm`. Data do Acesso: 06 de Junho de 2018.

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.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., et al. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6):899–907.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brejová, B., DiMarco, C., Vinar, T., Hidalgo, S. R., Holguin, G., and Patten, C. (2000). Finding patterns in biological sequences. *Unpublished project report for CS798G, University of Waterloo, Fall*.

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.

Campos, T. L., Korhonen, P. K., Gasser, R. B., and Young, N. D. (2019). An evaluation of machine learning approaches for the prediction of essential genes in eukaryotes using protein sequence-derived features. *Computational and Structural Biotechnology Journal*.

Cavasotto, C. N. and Phatak, S. S. (2009). Homology modeling in drug discovery: current trends and applications. *Drug Discovery Today*, 14(13-14):676–683.

Chen, K., Jiang, Y., Du, L., and Kurgan, L. (2009). Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs. *Journal of Computational Chemistry*, 30(1):163–172.

Chen, W.-H., Lu, G., Chen, X., Zhao, X.-M., and Bork, P. (2017). Ogee v2: An update of the online gene essentiality database with special focus on differentially essential genes in human cancer cell lines. *Nucleic Acids Research*, 45(D1):D940–D944.

Chen, W.-H., Minguez, P., Lercher, M. J., and Bork, P. (2011). Ogee: an online gene essentiality database. *Nucleic Acids Research*, 40(D1):D901–D906.

Chen, Z., Zhao, P., Li, F., Marquez-Lago, T. T., Leier, A., Revote, J., Zhu, Y., Powell, D. R., Akutsu, T., Webb, G. I., et al. (2019). ilearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of dna, rna and protein sequence data. *Briefings in Bioinformatics*.

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., Geis, J. R., Pandharipande, P. V., Brink, J. A., and Dreyer, K. J. (2018). Current applications and future impact of machine learning in radiology. *Radiology*, 288(2):318–328.

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.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20):1920–1930.

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.

Flórez, A. F., Park, D., Bhak, J., Kim, B.-C., Kuchinsky, A., Morris, J. H., Espinosa, J., and Muskus, C. (2010). Protein network prediction and topological analysis in leishmania major as a tool for drug target selection. *BMC Bioinformatics*, 11(1):484.

Frank, E., Hall, M. A., and Witten, I. H. (2016). *The WEKA workbench*. Morgan Kaufmann.

Galperin, M. Y., Fernández-Suárez, X. M., and Rigden, D. J. (2017). The 24th annual nucleic acids research database issue: a look back and upcoming changes. *Nucleic Acids Research*, 45(D1):D1–D11.

Garcia, F. P. and Belloze, K. T. (2018). Integração de dados na detecção de alvos para fármacos de schistosoma mansoni. In *Anais do XII Brazilian e-Science Workshop*. SBC.

Garcia, F. P., Guedes, G. P., and Belloze, K. (2019). Identifying schistosoma mansoni essential protein candidates based on machine learning. *Proceedings of the Brazilian Symposium on Bioinformatics 2019*, v. 1.

Gashaw, I., Ellinghaus, P., Sommer, A., and Asadullah, K. (2011). What makes a good drug target? *Drug Discovery Today*, 16(23-24):1037–1043.

Guido, R. V., Andricopulo, A. D., and Oliva, G. (2010). Planejamento de fármacos, biotecnologia e química medicinal: aplicações em doenças infecciosas. *estudos avançados*, 24(70):81–98.

Harrington, P. (2012). *Machine learning in action*. Manning Publications Co.

Hopkins, A. and Groom, C. (2003). Target analysis: a priori assessment of druggability. In *Small Molecule—Protein Interactions*, pages 11–17. Springer.

Katz, N., Dias, E., Araújo, N., and Souza, C. d. (1973). Estudo de uma cepa humana de schistosoma mansoni resistente a agentes esquistossomicidas. *Revista da Sociedade Brasileira de Medicina Tropical*, 7(6):381–387.

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.

Kohavi, R. and Quinlan, J. R. (2002). Data mining tasks and methods: Classification: decision-tree discovery. In *Handbook of data mining and knowledge discovery*, pages 267–276. Oxford University Press, Inc.

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

Kumar, K. K., Pugalenthi, G., and Suganthan, P. N. (2009). Dna-prot: identification of dna binding proteins from protein sequence information using random forest. *Journal of Biomolecular Structure and Dynamics*, 26(6):679–686.

Kumar, R., Kumari, B., and Kumar, M. (2017). Prediction of endoplasmic reticulum resident proteins using fragmented amino acid composition and support vector machine. *PeerJ*, 5:e3561.

Kumar, S., Mohri, M., and Talwalkar, A. (2012). Sampling methods for the nyström method. *Journal of Machine Learning Research*, 13(Apr):981–1006.

Lee, T.-Y., Lin, Z.-Q., Hsieh, S.-J., Bretaña, N. A., and Lu, C.-T. (2011). Exploiting maximal dependence decomposition to identify conserved motifs from a group of aligned signal sequences. *Bioinformatics*, 27(13):1780–1787.

Liu, X., Wang, B.-J., Xu, L., Tang, H.-L., and Xu, G.-Q. (2017). Selection of key sequence-based features for prediction of essential genes in 31 diverse bacterial species. *PloS One*, 12(3):e0174638.

Mitchell, T. M. et al. (1997). Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877.

Morais, E. R., Oliveira, K. C., de Paula, R. G., Ornelas, A. M., Moreira, É. B., Badoco, F. R., Magalhães, L. G., Verjovski-Almeida, S., and Rodrigues, V. (2017). Effects of proteasome inhibitor mg-132 on the parasite schistosoma mansoni. *PloS One*, 12(9):e0184192.

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.

MS (2019). Ministério da saúde. Disponível em: http://http://www.saude.gov.br/saude-de-a-z/esquistossomose#sintomas. Data do Acesso: 07 de Dezembro de 2019.

Murakami, Y. and Mizuguchi, K. (2014). Homology-based prediction of interactions between proteins using averaged one-dependence estimators. *BMC Bioinformatics*, 15(1):213.

Nakashima, H., Nishikawa, K., and Ooi, T. (1986). The folding type of a protein is relevant to the amino acid composition. *The Journal of Biochemistry*, 99(1):153–162.

Neves, B. J., Braga, R. C., Bezerra, J. C., Cravo, P. V., and Andrade, C. H. (2015). In silico repositioning-chemogenomics strategy identifies new drugs with potential activity against multiple life stages of schistosoma mansoni. *PLoS neglected tropical diseases*, 9(1):e3435.

Nigatu, D., Sobetzko, P., Yousef, M., and Henkel, W. (2017). Sequence-based information-theoretic features for gene essentiality prediction. *BMC Bioinformatics*, 18(1):473.

Ning, L., Lin, H., Ding, H., Huang, J., Rao, N., and Guo, F. (2014). Predicting bacterial essential genes using only sequence composition information. *Genet. Mol. Res*, 13:4564–4572.

Ocaña, K. A., de Oliveira, D., Dias, J., Ogasawara, E., and Mattoso, M. (2012). Discovering drug targets for neglected diseases using a pharmacophylogenomic cloud workflow. In *2012 IEEE 8th International Conference on E-Science*, pages 1–8. IEEE.

Palladino, W. (2009). *Conceitos de Genética*. Artmed Editora.

Paul, M. S., Kaur, A., Geete, A., and Sobhia, M. E. (2014). Essential gene identification and drug target prioritization in leishmania species. *Molecular BioSystems*, 10(5):1184–1195.

Peng, C., Lin, Y., Luo, H., and Gao, F. (2017). A comprehensive overview of online resources to identify and predict bacterial essential genes. *Frontiers in Microbiology*, 8:2331.

Perticaroli, S., Nickels, J. D., Ehlers, G., O'Neill, H., Zhang, Q., and Sokolov, A. P. (2013). Secondary structure and rigidity in model proteins. *Soft Matter*, 9(40):9548–9556.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Pollard, T. D. and Cooper, J. A. (2009). Actin, a central player in cell shape and movement. *Science*, 326(5957):1208–1212.

Protasio, A. V., Tsai, I. J., Babbage, A., Nichol, S., Hunt, M., Aslett, M. A., De Silva, N., Velarde, G. S., Anderson, T. J. C., Clark, R. C., Davidson, C., Dillon, G. P., Holroyd, N. E., LoVerde, P. T., Lloyd, C., McQuillan, J., Oliveira, G., Otto, T. D., Parker-Manuel, S. J., Quail, M. A., Wilson, R. A., Zerlotini, A., Dunne, D. W., and Berriman, M. (2012). A systematically improved high quality genome and transcriptome of the human blood fluke schistosoma mansoni. *PLoS Neglected Tropical Diseases*, 6(1):e1455.

Rice, P., Longden, I., and Bleasby, A. (2000). Emboss: the european molecular biology open software suite. *Trends in Genetics*, 16(6):276–277.

Sahoo, G. and Kumar, Y. (2012). Analysis of parametric & non parametric classifiers for classification technique using weka. *International Journal of Information Technology and Computer Science (IJITCS)*, 4(7):43.

SH (2018). Schistosoma mansoni. Disponível em: `https://www.schistosig.com/schisto`. Data do Acesso: 26de Maio de 2018.

Si, J., Cui, J., Cheng, J., and Wu, R. (2015). Computational prediction of rna-binding proteins and binding sites. *International Journal of Molecular Sciences*, 16(11):26303–26317.

Son, J.-D., Niu, G., Yang, B.-S., Hwang, D.-H., and Kang, D.-S. (2009). Development of smart sensors system for machine fault diagnosis. *Expert Systems with Applications*, 36(9):11981–11991.

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.

Uddin, R. and Jamil, F. (2018). Prioritization of potential drug targets against p. aeruginosa by core proteomic analysis using computational subtractive genomics and protein-protein interaction network. *Computational Biology and Chemistry*, 74:115–122.

Wang, S., Peng, J., Ma, J., and Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports*, 6:18962.

WHO (2018). Shistosomiasis. Available 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.

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846.

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., Billis, K., Cummins, C., Gall, A., Girón, C. G., et al. (2017). Ensembl 2018. *Nucleic Acids Research*, 46(D1):D754–D761.

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

A- Lista de Proteínas Candidatas Essenciais do *S. mansoni*

Tabela 6 – Lista de 138 proteínas resultantes do processo de ortologia entre as proteínas essenciais dos organismos modelo selecionados e o proteoma do *S. mansoni*.

S. mansoni ID	**Descrição**
Smp_046360.1	Putative zinc finger protein
Smp_000600.1	Notchless homolog 1
Smp_002550.1	Putative serine/arginine rich splicing factor
Smp_002550.2	Putative serine/arginine rich splicing factor
Smp_007660.1	Putative transcription elongation factor s-II
Smp_010930.1	Putative heat shock protein 70 (Hsp70)-interacting protein
Smp_013790.1	DEAD box ATP-dependent RNA helicase, putative
Smp_018240.1	Cell division control protein 48 aaa family protein
Smp_018240.2	Cell division control protein 48 aaa family protein
Smp_018240.3	Cell division control protein 48 aaa family protein
Smp_020770.1	Defender against cell death, putative
Smp_020920.1	Putative dnaj homolog subfamily B member 4
Smp_020920.2	Putative dnaj homolog subfamily B member 4
Smp_022020.3	Putative cpg binding protein
Smp_022660.1	Putative wd-repeat protein
Smp_026560.1	Putative calmodulin
Smp_026560.2	Putative calmodulin
Smp_026770.1	Small nuclear ribonucleoprotein-associated protein
Smp_027970.1	ATP-dependent RNA helicase, putative
Smp_028100.1	Putative ubiquitin-conjugating enzyme E2 G
Smp_028100.2	Putative ubiquitin-conjugating enzyme E2 G
Smp_032320.6	Putative serine/arginine rich splicing factor
Smp_033260.1	Family C48 unassigned peptidase (C48 family)
Smp_033260.2	Family C48 unassigned peptidase (C48 family)
Smp_033710.1	DEAD box ATP-dependent RNA helicase, putative
Smp_034100.1	DNA-directed RNA polymerase subunit
Smp_035200.1	Putative hsp40, subfamily A, members 1,2,4
Smp_035200.2	Putative hsp40, subfamily A, members 1,2,4
Smp_035200.3	Putative hsp40, subfamily A, members 1,2,4
Smp_035200.4	Putative hsp40, subfamily A, members 1,2,4
Smp_035220.1	Putative kif-3

***S. mansoni* ID**	**Descrição**
Smp_036270.10	Putative arginine/serine-rich splicing factor
Smp_036270.2	Putative arginine/serine-rich splicing factor
Smp_037400.1	Putative cell division cycle
Smp_040800.1	Putative glycyl-tRNA synthetase
Smp_041390.1	Ubiquitin-conjugating enzyme E2 G, putative
Smp_046600.1	Actin-1
Smp_051010.1	Putative ubiquitin-conjugating enzyme E2 C
Smp_053140.1	Tip60, putative
Smp_054340.1	Putative chaperonin containing t-complex protein 1, beta subunit, tcpb
Smp_054340.2	Putative chaperonin containing t-complex protein 1, beta subunit, tcpb
Smp_054340.3	Putative chaperonin containing t-complex protein 1, beta subunit, tcpb
Smp_055780.1	Smdr2
Smp_055920.1	Putative ubiquitin conjugating enzyme E2
Smp_055960.1	Putative ubiquitin conjugating enzyme E2
Smp_063000.1	Smdr1
Smp_064860.1	Putative heat shock protein 70 (Hsp70)-interacting protein
Smp_065980.1	Putative heat shock protein 70 (Hsp70)-interacting protein
Smp_067980.1	Ubiquitin conjugating enzyme E2, putative
Smp_067980.3	Ubiquitin conjugating enzyme E2, putative
Smp_067980.4	Ubiquitin conjugating enzyme E2, putative
Smp_067990.1	Ubiquitin conjugating enzyme E2, putative
Smp_068000.1	Ubiquitin conjugating enzyme E2, putative
Smp_068270.1	Putative heat shock transcription factor
Smp_068270.2	Putative heat shock transcription factor
Smp_068270.3	Putative heat shock transcription factor
Smp_068270.4	Putative heat shock transcription factor
Smp_068270.5	Putative heat shock transcription factor
Smp_068440.1	Putative dead box ATP-dependent RNA helicase
Smp_072560.1	Putative uncharacterized protein
Smp_072870.1	Actin, putative
Smp_072970.1	Putative tetratricopeptide protein
Smp_077860.1	Putative actin
Smp_078640.1	Putative snrnp sm protein
Smp_079730.1	Putative rna recognition motif containing protein
Smp_080860.1	Serine/threonine kinase
Smp_082050.1	Putative ubiquitin-conjugating enzyme E2-25kD
Smp_082820.1	Putative ubiquitin-conjugating enzyme morgue
Smp_083400.1	Ubiquitin conjugating enzyme 13; Ubiquitin conjugating enzyme 13, putative
Smp_085690.1	Serine/threonine kinase
Smp_086600.2	Putative arginine/serine-rich splicing factor
Smp_087010.1	DNA polymerase

***S. mansoni* ID**	**Descrição**
Smp_087270.1	Putative atp-dependent RNA helicase
Smp_087320.1	Putative zinc finger protein
Smp_087930.1	Putative atp-binding cassette, sub-family B, member 7,mitochondrial (Abc7)
Smp_090860.1	Putative uncharacterized protein
Smp_093690.1	Putative dead box ATP-dependent RNA helicase
Smp_093690.2	Putative dead box ATP-dependent RNA helicase
Smp_095350.1	Zinc finger protein, putative
Smp_098890.1	Arginine/serine-rich splicing factor, putative
Smp_098890.2	Arginine/serine-rich splicing factor, putative
Smp_104730.1	Putative dnaj homolog subfamily B member 4
Smp_118220.1	Putative ubiquitin-conjugating enzyme rad6
Smp_123290.1	Serine/threonine kinase
Smp_123640.1	Serine/threonine kinase
Smp_123710.1	Acetyl-CoA carboxylase
Smp_128050.1	Importin-alpha re-exporter (Chromosome segregation 1-like protein)
Smp_132300.1	Putative kif-3
Smp_133120.1	Putative dead box ATP-dependent RNA helicase
Smp_133120.2	Putative dead box ATP-dependent RNA helicase
Smp_134020.1	Putative cell division cycle
Smp_134890.1	Putative atp-binding cassette, sub-family B,member 6,mitochondrial (Abc6)
Smp_136750.1	Putative e3 ubiquitin-protein ligase Bre1
Smp_138680.1	DNAj homolog subfamily C member, putative
Smp_138680.2	DNAj homolog subfamily C member, putative
Smp_139340.1	Putative rab
Smp_141080.1	Putative DNAj (Hsp40) homolog, subfamily B, member
Smp_141630.1	Splicing factor 3b, subunit 1-related
Smp_141630.2	Splicing factor 3b, subunit 1-related
Smp_141630.4	Splicing factor 3b, subunit 1-related
Smp_145190.1	Protein kinase
Smp_147010.1	Putative zinc finger protein
Smp_147450.1	Serine/threonine kinase
Smp_150040.1	Putative cell division protein kinase
Smp_150040.2	Putative cell division protein kinase
Smp_151100.1	Serine/threonine kinase
Smp_151140.1	Serine/threonine kinase
Smp_151350.1	Heat shock protein 70 (Hsp70)-interacting protein, putative
Smp_152330.1	Serine/threonine kinase
Smp_154320.1	VLG2 description:DEAD box ATP-dependent RNA helicase, putative
Smp_154580.1	Putative rab 3 and
Smp_155480.1	Putative heat shock protein 70 (Hsp70)-interacting protein

S. mansoni ID	**Descrição**
Smp_156060.1	Putative uncharacterized protein
Smp_157510.1	Putative atp-dependent RNA helicase
Smp_157680.1	U6 snrna-associated sm-like protein lsm4
Smp_158560.1	Serine/threonine kinase
Smp_158560.2	Serine/threonine kinase
Smp_160140.1	Transcription factor,putative
Smp_160760.1	Serine/threonine kinase
Smp_161920.1	Putative actin
Smp_164380.1	Putative camp-dependent protein kinase catalytic subunit
Smp_165400.1	Putative propionyl-CoA carboxylase alpha subunit
Smp_165540.1	Peptidyl-prolyl cis-trans isomerase-like 4. ppil4
Smp_169440.1	Ubiquitin-conjugating enzyme rad6, putative
Smp_170820.1	Multidrug resistance protein 1, 2, 3 (P glycoprotein 1, 2, 3), putative
Smp_174210.1	Putative uncharacterized protein Smp_174210.1
Smp_174820.1	Serine/threonine kinase
Smp_176360.1	Protein kinase C, putative
Smp_179070.1	Putative ribosomal protein S6 kinase
Smp_180170.1	Ubiquitin-conjugating enzyme e2S, putative
Smp_183710.1	Putative actin
Smp_185230.1	Arginine/serine-rich splicing factor, putative
Smp_194610.1	Serine/threonine kinase
Smp_202970.1	Putative uncharacterized protein Smp_202970
Smp_203130.1	Putative uncharacterized protein Smp_203130
Smp_204050.1	Putative uncharacterized protein Smp_204050
Smp_207000.1	Anamorsin homolog
Smp_210830.1	Family S9 non-peptidase homologue (S09 family)