

RÔMULO LUCIO VALE DE MORAES

**Análise computacional do genoma de
Schistosoma mansoni para identificação de
proteínas potencialmente imunogênicas**

Belo Horizonte

2012



Universidade Federal de Minas Gerais

Departamento de Bioquímica e Imunologia

Programa de Pós-graduação em Bioinformática

RÔMULO LUCIO VALE DE MORAES

**Análise computacional do genoma
de *Schistosoma mansoni* para identificação de
proteínas potencialmente imunogênicas**

Projeto de Tese apresentada ao Curso de Pós-graduação em Bioinformática da Universidade Federal de Minas Gerais em 06 de Julho de 2012 como requisito parcial para a obtenção do grau de Doutor em Bioinformática.

Orientador: **Guilherme Oliveira**
Grupo de Genômica e Biologia Computacional
Instituto René Rachou – CPqRR
Fundação Oswaldo Cruz – FIOCRUZ
Belo Horizonte – MG – Brasil

Belo Horizonte

2012

Esse trabalho foi iniciado no Instituto René Rachou – FIOCRUZ (IRR-FIOCRUZ) em Agosto 2007 como proposta de tese de Doutorado. Parte do trabalho foi desenvolvido sob apoio da estrutura e membros do Centro de Excelência em Bioinformática de Minas Gerais - CEBio e posteriormente concluído também no grupo de Genômica e Biologia Computacional (IRR-FIOCRUZ), sob a orientação do Dr. Guilherme Oliveira. O projeto contou com o suporte financeiro do Fogarty-NIH, FAPEMIG e CNPQ.

*“De tudo ficaram três coisas:
a certeza de que estamos sempre a começar,
a certeza de que é preciso continuar,
e a certeza de que seremos interrompidos antes de terminar.*

*Portanto, devemos:
fazer da interrupção um caminho novo,
da queda, um passo de dança,
do medo, uma escada,
do sonho, uma ponte,
da procura, um encontro.”*

– Fernando Sabino

Agradecimentos

Durante essa jornada houve diversos momentos que achei que jamais escreveria os agradecimentos de uma tese. Em outros momentos eu até ensaiava mentalmente como seria, e por vezes temia que tudo isso ficasse apenas na minha imaginação. Assim, estar escrevendo tal texto agora me é tão estranho, tão surreal, que fico pensando se não mais um dos caprichos de algum sonho distante.

De qualquer forma eu aproveito este momento para deixar registrada a minha gratidão à todos que contribuíram para este fim. E vale lembrar que a ordem dos eventos ou a cronologia dos fatos aqui não tem muita relevância, todos que estão aqui têm o mesmo peso, sigo apenas uma ordem que neste momento me faz algum sentido.

Agradeço, pois:

Aos meus pais, Maggi e Marilda, pela vida, educação, princípios, apoio, orações e, sobretudo pelo amor, sem distância, tempo e limites. **MUITO OBRIGADO!** Essa tese eu dedico a vocês!

Aos meus irmãos, Romero (Dedeim) e Roni (Selva), pela estima, admiração, e torcida sempre. Obrigado por vocês existirem!

Ao meu orientador e amigo, Dr. Guilherme Oliveira, pela oportunidade, formação, paciência, pelo exemplo e inspiração, e por todo incentivo e apoio ofertado desde que cheguei à Belo Horizonte e ingressei na vida acadêmica. Minha eterna gratidão!

Ao meu amigo e co-orientador informal, Adhemar (Neto), presente nos momentos cruciais e pelo companherismo desde o início da minha jornada na Bioinformática. Aprendi muito com você brother!

Aos amigos, Anderson e Flinkas, pelo companherismo, receptividade e por todos os momentos que tornaram a vida longe de casa mais fácil.

À Ronalu, pelo amor, carinho, integridade e força de sua pessoa, pelo sacrifício e por todo apoio que você deu no meu trabalho e na minha vida.

À Dra. Rosiane Pereira e a Fernanda Ludolf pela ajuda e contribuição de vocês com os experimentos de proteômica. Rosiane, sua ajuda foi fundamental.

Ao Dr. Jerônimo Ruiz, pela ajuda no início deste projeto.

À gerente do Lab, Ângela Volpini, pela amizade, atenção constante e disponibilidade em resolver nossos problemas.

À todos os amigos do Grupo de Genômica e Biologia Computacional, Livia, Flávio, Luíza, Larissa, Marina, Kamitami, Laila, Yesid e demais membros, pelos momentos científicos e também de descontração sem os quais a convivência não seria tão bacana. Livia obrigado pelo seu carinho eterno.

Aos amigos do CEBio, Fabiano, Laura, Eric, Sara, Fausto, Eliane e Mariana, pela convivência e também pelo suporte ao meu trabalho.

Aos meus grandes amigos, de jurássicas datas, Roniere Miranda e Edvar Júnior, os quais nem sempre vejo, mas quando encontramos são sempre bons momentos!

Aos amigos e companheiros da graduação em Computação na UFPI e os quais carrego em alto estima pra vida inteira: Lenno, Pardal, Metódio, Ricardo, Faltoso e Iális.

À Dra. Semiramis, que plantou a semente da Bioinformática e da ciência na minha vida.

Aos colegas e aos professores do Programa de Doutorado em Bioinformática da UFMG.

Por fim, a todos os familiares e amigos que de alguma forma contribuíram pra este feito e que infelizmente não foram citados aqui.

Resumo

A esquistossomose continua sendo um dos principais problemas de saúde dos países tropicais, incluindo o Brasil. A doença causada pelo organismo *Schistosoma mansoni*, dentre as parasitoses existentes, é a segunda principal causa de morbidade no mundo depois da malária. No presente trabalho foi desenvolvida uma abordagem para descoberta de alvos de vacinas contra esquistossomose, utilizando recursos *in silico*. O principal objetivo nesta abordagem foi a predição de proteínas antigênicas a partir do genoma do organismo de interesse. Dentre as diversas predições realizadas, estão aquelas que determinam as proteínas que estão potencialmente expostas ao hospedeiro humano (proteínas secretadas e transmembranas) e de posse destas, são identificadas aquelas que têm um melhor perfil imunogênico, através do mapeamento de seus epitopos. Tais análises resultaram em 90 proteínas com potencial imunogênico, que incentivaram algumas validações experimentais que ainda estão em desenvolvimento. Também motivaram a criação de um pipeline que envolve todas as predições realizadas no projeto através de um web-server (EpiFinder: <http://epifinder.cebio.org>). Dessa forma espera-se contribuir para a descoberta de novos alvos de vacinas de forma mais rápida e eficaz, não apenas para esquistossomose, mas também para outras parasitoses.

Abstract

Schistosomiasis continues to be a significant public health problem in tropical countries including Brazil. The disease caused by the parasite *Schistosoma mansoni*, is the second main cause of morbidity in the world among parasitic diseases, right after malaria. In the present research work, we describe the development of an approach aiming discovering vaccine candidates for schistosomiasis using computational resources. The main purpose for this approach was the prediction of antigenic proteins from the genome of the target organism. Among the various predictions made are those that determine which the proteins which are exposed to human host (secreted and transmembrane proteins). The selected proteins were further submitted to epitope prediction. Such analysis resulted in 90 proteins with good immunogenic potential, that encouraged some experimental validations still under development. This work encouraged the creation of a pipeline that involves all the predictions made in the project, provided in the form of a web-server (EpiFinder: <http://epifinder.cebio.org>). In conclusion, we expected to contribute to the discovery of new targets for vaccine more quickly and effectively not only in schistosomiasis, but also other parasites.

Sumário

1. INTRODUÇÃO	1
2. FUNDAMENTAÇÃO TEÓRICA.....	3
2.1. <i>Schistosoma mansoni</i>	8
2.2. Aspectos da terapêutica e da imunologia da esquistossomose	11
2.3. O genoma de <i>Schistosoma mansoni</i>	13
2.4. Predição da localização celular de proteínas.....	15
2.4.1. Predição de proteínas secretadas	16
2.4.2. Predição de proteínas transmembranas.....	18
2.5. Predição de epitopos.....	19
3. OBJETIVOS.....	23
3.1. <i>Gerais</i>	23
3.2. <i>Específicos</i>	23
4. MATERIAIS E MÉTODOS	24
4.1. Obtenção do proteoma predito	24
4.2. Filtragem das proteínas por fase do ciclo de vida.....	27
4.3. Recursos computacionais para as predições e para desenvolvimento do pipeline.....	28
4.4. Predição proteínas secretadas	29
4.5. Predição das proteínas transmembranas	32
4.6. Predição das proteínas antigênicas.....	33
4.7. Criação do banco de perfis de epitopos.....	37
4.8. Validação experimental da metodologia	39
4.9. Implementação do pipeline de busca automática de novos alvos.	42
5. RESULTADOS E DISCUSSÃO	49
5.1. Predições e seleção dos candidatos a validação experimental.	49
5.2. Bancos de perfis de epitopos e Servidor Web para realização das predições.....	51
5.3. Avaliações experimentais.....	54
5.3.1. Resultados de proteômica	60
5.3.2. Resultados in silico	63
1. CONCLUSÕES.....	65
2. LIMITAÇÕES E PERSPECTIVAS.....	66
3. REFERÊNCIAS BIBLIOGRÁFICAS	68
4. ANEXOS	77

Lista de Tabelas

Tabela 1	4
Tabela 2	7
Tabela 3	29
Tabela 4	34
Tabela 5	49
Tabela 6	51
Tabela 7	55

Lista de Figuras

Figura 1 - Descoberta e desenvolvimento de vacinas.....	5
Figura 2 - Ciclo Biológico do <i>S. mansoni</i>	10
Figura 3 - Tipos de Respostas.....	19
Figura 4 – A molécula de MHC.....	20
Figura 5 - SchistoDB.....	25
Figura 6 - Tamanho das proteínas de <i>S. mansoni</i>	26
Figura 7 - Fases do ciclo de vida dos alvos.....	27
Figura 8 - Saída do Signalp (Proteínas secretadas).....	30
Figura 9 - Saída do SherLoc (Proteínas secretadas).....	31
Figura 10 - Saída do SecretomeP (Proteínas secretadas não clássicas).....	32
Figura 11 - Saída do Tmhmm (Proteínas transmembranas).....	33
Figura 12 - Arquitetura do FRED.....	34
Figura 13 – Matriz para predição de ligantes ao modelo alélico HLA-B*1510 do SYFPEITHI.....	36
Figura 14 - Esquema do Banco de Dados.....	38
Figura 15 - Etapas do projeto.....	42
Figura 16 - Pipeline plataforma Galaxy.....	43
Figura 17 - EpiFinder tecnologias empregadas.....	45
Figura 18 - Diagrama de Classes da Arquitetura.....	47
Figura 19 - EpiFinder Webserver.....	48
Figura 20 – Distribuição de Epitopos por Proteínas.....	50
Figura 21 - EpiFinder – Home Page das Proteínas.....	52
Figura 22 - EpiFinder – Resultado da predição de epitopos em proteínas expostas.....	53
Figura 23 - EpiFinder – Mapeamento de epitopos para uma dada proteína.....	53
Figura 24 – Mapeamento de epitopos em uma proteína.....	60
Figura 25 – Perfil eletroforético bidimensional do extrato proteico de vermes adultos de <i>Schistosoma mansoni</i> enriquecido de proteínas de membrana e <i>Western blots</i> bidimensionais utilizando anticorpo anti-Sm29 e pool de soro de indivíduos infectados.....	61

Lista de Abreviaturas

AA: aminoácidos

API: *Application Programming Interface* (Interface de Programação de Aplicativos)

cDNA: DNA Complementar

CEBio: Centro de Excelência em Bioinformática

CG: Complexo de Golgi

CNPq: Conselho Nacional de Desenvolvimento Científico e Tecnológico

DAO: *Data Access Object* (Objeto de Acesso à Dados)

DNA: Ácido Desoxiribonucléico

EST: *Expressed Sequence Tag* (Etiqueta de Sequência Expressa)

FAPESP: Fundação de Amparo à Pesquisa do Estado de São Paulo

GB: *Giga byte* (1.000.000 bytes)

Ghz: *Giga Hertz* (1.000.000 hertz)

GO: *Gene Ontology*

GOLD: **Genome Online Database**

h: hora

HD: *Hard Disk* (Disco rígido)

HMM: *Hidden Markov Model* (Modelo Oculto de Markov)

HLA: *Human Leucocyte Antigen* (Antígeno de Leucócito Humano)

IFN: Interferon

IL: Interleucina

JEE: *Java Enterprise Edition* (Java Edição Empresarial)

JPA: Java Persistence API (API de Persistência em Java)

JSF: *Java Server Faces*

JSP: *Java Server Pages*

kbp: *kilo base pairs* (mil de pares de base)

LT: Linfotoxina

MB: *Mega byte* (1000 bytes)

mbp: *million base pairs* (milhões de pares de base)

MCT: Ministério da Ciência e Tecnologia

MHC: *Major Histocompatibility Complex* (Complexo de Histocompatibilidade Principal)

MS: *Mass Spectrometry* (Espectrometria de massa)

NCBI: *National Center for Biotechnology Information*

NIH: *National Institutes of Health*

NN: *Neural Network* (Rede Neural)

OMS: Organização Mundial de Saúde

ORESTES: *Open Reading Frames ESTs* (ESTs de Janelas Abertas de Leitura)

PCR: *Polymerase chain reaction* (Reação em Cadeia de Polimerase)

POJO: *Plain Old Java Objects* (Singelos Clássicos Objetos Java)

PS: Peptídio Sinal

PZQ: Praziquantel

RAD: *Rapid Application Development* (Desenvolvimento Rápido de Aplicação)

RAM: *Random Access Memory* (Memória de Acesso Randômico)

RE: Retículo Endoplasmático

SQL: *Structured Query Language* (Linguagem de Consulta Estruturada)

SVM: *Support Vector Machine* (Máquina de Vetores de Suporte)

TCR: *T-cell receptor* (Receptores de células T)

TDR: *Tropical Disease Research*

TH0: *T Helper 0*

TH1: *T Helper 1*

TH2: *T Helper 2*

TIGR: *The Institute for Genomic Research*

WGS: *Whole Genome Shotgun* (Sequenciamento de genoma completo por *Shotgun*)

WHO: *World Health Organization* (Organização Mundial de Saúde)

WTSI: Welcome Trust Sanger Institute

XML: *eXtensible Markup Language* (Linguagem de Marcação Extensível)

1. INTRODUÇÃO

A esquistossomose é uma doença parasitária que atinge principalmente indivíduos em áreas rurais de países subdesenvolvidos. Estima-se que existam 240 milhões de casos da doença no mundo, com mais de 700 milhões de pessoas sob o risco de infecção em áreas conhecidamente endêmicas da África, Ásia e América do Sul (Who, 2011). No Brasil a doença é causada pela espécie *Schistosoma mansoni* e estima-se que no país 8 (oito) milhões de pessoas estejam infectadas, com áreas endêmicas principalmente nas regiões Nordeste, Sudeste e Centro-Oeste (Katz e Peixoto, 2000).

Apesar da existência de tratamentos eficazes a situação da doença continua grave, pois nas áreas endêmicas as pessoas se tornam rapidamente reinfetadas (Engels, Chitsulo *et al.*, 2002) e, nesse contexto, o desenvolvimento de vacinas pode fornecer uma inestimável contribuição para o controle dessa parasitose. Assim, desde 1993, a Organização Mundial de Saúde (OMS) induziu o estudo da genômica como uma nova abordagem para o desenvolvimento de novas ferramentas de controle do parasito. O estudo da genômica do parasito gera expectativa para um melhor entendimento da biologia, da fisiologia deste organismo e também na busca de candidatos para diagnóstico, drogas e para produção de novas vacinas (Degrave, Melville *et al.*, 2001).

A possibilidade de desenvolvimento de uma vacina contra a esquistossomose é real e já foi demonstrada em camundongos no modelo de cercarias irradiadas (Aitken, Coulson *et al.*, 1987) e, também, por uso de vacinas atenuadas em babuínos, indicando a possibilidade de proteção de primatas não humanos contra a doença (Kariuki, Farah *et al.*, 2004). Além disso, diversos estudos demonstraram que vacinas de DNA baseadas em proteínas de *S. mansoni* têm a capacidade de reduzir a carga parasitária em modelos animais. Uma vacina de DNA baseada na proteína Sm-p80, por exemplo, exibiu uma redução de 59% da carga de vermes adultos em camundongos (Ahmad, Torben *et al.*, 2009). Tal vacina também foi capaz de diminuir a produção de ovos do parasito em 84% nos animais imunizados.

Porém, até recentemente, o desenvolvimento e a pesquisa de novas vacinas esteve associado a métodos convencionais e, portanto, gerados por meio de abordagens bioquímica, imunológica e microbiológica. Com o advento de novas técnicas de biologia molecular e do

sequenciamento de genomas completos, novas perspectivas têm revolucionado a vacinologia clássica.

Em Abril de 2012, um total de 3173 genomas completos constava como sequenciados (Gold, 2011) e o impacto da disponibilidade dessa informação relacionado ao desenvolvimento de vacinas já pode ser avaliado. Como exemplo real do emprego dessa nova abordagem de estudo, que utiliza metodologias computacionais de predição em associação com dados de proteoma e transcriptoma para o desenvolvimento de vacinas, temos a identificação de candidatos a vacina contra o *meningococo* sorotipo B. Na abordagem de vacinologia reversa descrita por Pizza e colaboradores (Pizza, Scarlato *et al.*, 2000), em apenas 18 meses foram identificados mais candidatos a vacina do que em 40 anos pelo método convencional.

Tomando-se como organismo modelo o genoma de *S. mansoni*, o presente projeto descreve abordagens computacionais direcionadas para identificação de proteínas que seriam novos alvos de estudo para desenvolvimento de uma vacina contra esquistossomose. Os pontos principais são as predições computacionais em escala genômica e a integração destes resultados, que são passíveis de comprovação em laboratório, que aqui iniciados de forma independente.

2. FUNDAMENTAÇÃO TEÓRICA

O controle da esquistossomose deve ser considerado sob dois aspectos, tratamento e transmissão. Para controle de transmissão as medidas disponíveis ainda não são suficientes, por demandar investimentos em longo prazo por parte dos países endêmicos. Por isso o controle da doença depende quase totalmente do tratamento através de uma única droga disponível para o tratamento em massa – o praziquantel (Chitsulo, Engels *et al.*, 2000). O medicamento é utilizado desde a década de 80, apresentando resultados eficientes, com baixa toxicidade e boa tolerância, administrado em dose única e por via oral.

No entanto, apesar dos progressos notáveis obtidos com o tratamento quimioterápico, a doença tem se propagado em novas áreas, pois o tratamento em massa não evita a reinfecção. Além disso, apesar de não existirem evidências claras, foram observados indícios de resistência à droga (Doenhoff e Pica-Mattoccia, 2006). Dessa forma a vacinação se apresenta como componente essencial em complemento à quimioterapia no controle da esquistossomose.

Nos últimos anos diversos trabalhos têm possibilitado uma melhor compreensão da resposta imune à infecção por *S. mansoni* (Correa-Oliveira, Caldas *et al.*, 2000). E os resultados do esforço da OMS (WHO/TDR) na criação de programas que proporcionassem o rápido desenvolvimento de vacinas e outras ferramentas para o controle de doenças parasitárias, sugerem que uma vacina para esquistossomose é possível (Mcmanus e Loukas, 2008). Tal iniciativa permitiu a descoberta de mais de 10 antígenos com forte potencial como candidatos à vacina (Tabela 1).

Tabela 1

Proteínas recombinantes de *S. mansoni* correlacionadas à estudos de resistência em humano e/ou exibem proteção em modelo animal. Fonte adaptada: (Mcmanus e Loukas, 2008)

Proteína ou cDNA	Localização (verme adulto)	Identidade	Proteção da Vacina* (animal)	Referência
SmTSP-2 (tetraspanina D)	Tegumento	Tetraspanina	++ Ovos ++ Vermes	(Smyth, Mcmanus <i>et al.</i> , 2003)
SmTSP-1	Tegumento	Tetraspanina	++ Ovos + Vermes	(Tran, Pearson <i>et al.</i> , 2006)
Sm29	Tegumento	Desconhecida (Transmemb.)	++ Vermes	(Cardoso, Pacifico <i>et al.</i> , 2006)
Sm23	Tegumento	Tetraspanina	+ Vermes	(Wright, Melder <i>et al.</i> , 1991)
Sm-p80	Membrana int. do tegumento	Calpaína	+ Vermes	(Braschi e Wilson, 2006)
Sm14	Corpo inteiro, citosol	FABP	++ Vermes	(Tendler, Vilar <i>et al.</i> , 1995)
Sm28-GST	Corpo inteiro	GST	+ Vermes	(Auriault, Wolowczuk <i>et al.</i> , 1991)
Sm28-TPI	(Tegumento do esquistossômulo)	TPI	+ Vermes	(Al-Sherbiny, Osman <i>et al.</i> , 2003)
Sm27-paramyosin	Musculatura	Paramiosina	+ Vermes	(Al-Sherbiny, Osman <i>et al.</i> , 2003)
CT-SOD	Tegumento e epitélio intestinal	Cu-ZN SOD	++ Vermes	(Mei e Loverde, 1997)

*Proteção da vacina: + Redução de 30-50%, e, ++ Redução de >50% (ovos e vermes no fígado).

Como pode ser visto na Tabela 1, a proteção medida pela diminuição do número de vermes, varia entre os diferentes antígenos mas atingem em média de 40%. Podemos considerar então que todos estes apresentam resultados razoáveis. Mas acredita-se que será necessário combinar vários antígenos para compor uma vacina que seja eficaz contra a doença

(Mcmanus e Loukas, 2008). Por isso, a descoberta de novos possíveis alvos é fundamental para esta tarefa.

Assim, a disponibilidade de seqüências do parasito *S. mansoni*, que recentemente teve seu genoma sequenciado pelo The Institute for Genomic Research - TIGR em associação com o Wellcome Trust Sanger Institute – WTSI (Berriman, Haas *et al.*, 2009), poderá fornecer uma fonte rica de dados, gerando informações que poderão ser usadas para descobertas de diversos produtos inclusive candidatos à vacina.

O impacto do recente desenvolvimento da genômica associado às tecnologias de análise *in silico* de seqüências, procedimento conhecido como *mineração genômica*, tem se mostrado efetivo em diversas áreas e na vacinologia não é exceção. As novas tecnologias de seqüenciamento de genomas completos, e o aumento do número de ferramentas e métodos de análise disponíveis para mineração de informação biológica, têm diminuído consideravelmente o tempo de pesquisa e descoberta de novos alvos para vacinas (Figura 1).

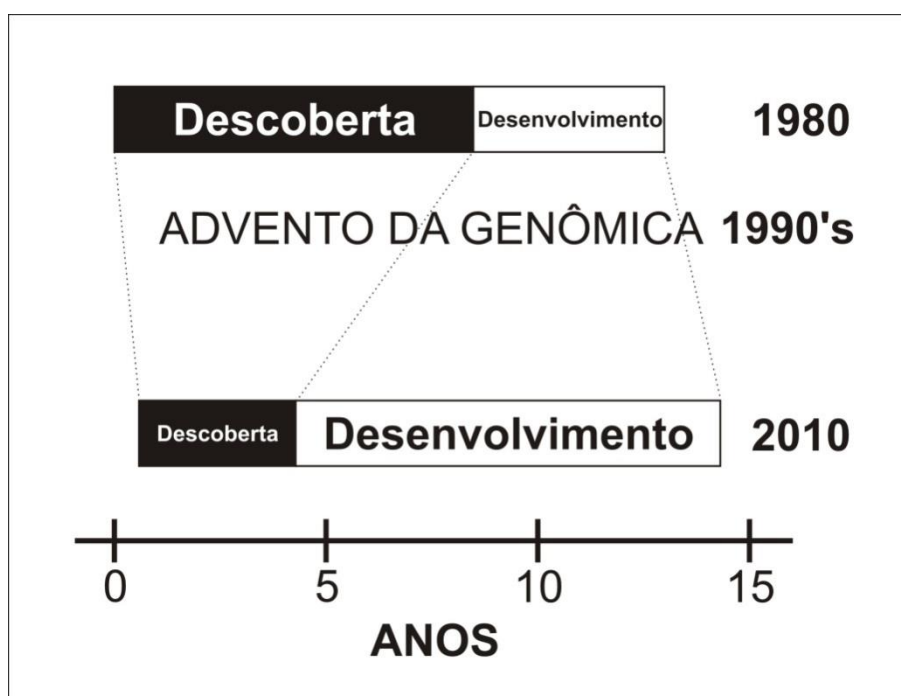


Figura 1 - Descoberta e desenvolvimento de vacinas.

O impacto da genômica na descoberta e desenvolvimento de vacinas. No passado as abordagens convencionais demandavam anos ou décadas para a identificação de antígenos protetores. Com o advento da era genômica na década de 1990, o tempo de identificação de novos alvos caiu drasticamente, no entanto a etapa de Desenvolvimento

tornou-se mais complexa pela disponibilidade de novas ferramentas, bem como a necessidade de vacinas mais seguras e baratas. Fonte adaptada: (Bambini e Rappuoli, 2009)

O termo “vacinologia reversa” vem descrever tais abordagens pós-genômicas para o desenvolvimento de vacinas. A vacinologia reversa usa o genoma do patógeno de interesse em vez de material biológico, como fonte de partida para a identificação de antígenos, cuja atividade deve ser confirmada posteriormente pela biologia experimental (Rappuoli, 2001). Em geral, o objetivo é a identificação de genes que possam estar expostos ao sistema imune do hospedeiro e que codificam fatores de patogenicidade. São necessários algoritmos específicos para a identificação de tais características, e, seguida é avaliada a possibilidade de resposta imune dos peptídeos das proteínas selecionadas.

Esta abordagem tem sido usada como estratégia de desenvolvimento de vacinas para diversos organismos, principalmente vírus e protozoários (Tabela 2), mas recentemente alguns grupos também têm aplicado as metodologias de vacinologia reversa também para eucariotos (Kanoi e Egwang, 2007). Com a disponibilidade de algoritmos e modelos de buscas específicos para estes organismos, a mineração das características biológicas mais interessantes, tem permitido a descoberta de alvos para vacina, em larga escala, em um tempo nunca imaginado para as técnicas convencionais.

Tabela 2

Exemplos de algumas bactérias patogênicas em que foram utilizadas abordagens pós-genômica (vacinologia reversa) para o desenvolvimento de vacinas. Fonte adaptada: (Bambini e Rappuoli, 2009)

Patógeno	Doença	Estágio de Desenvolvimento da Vacina	Referência
<i>Neisseria meningitidis serogroup B</i>	Meningite bacteriana e septicemia.	Fase II/Tratamento clínico	(Pizza, Scarlato <i>et al.</i> , 2000)
<i>Streptococcus pneumoniae</i>	Pneumonia bacteriana, sepse, sinusite, otite média e meningite bacteriana	Descoberta/Pré-clínica	(Wizemann, Heinrichs <i>et al.</i> , 2001)
<i>Bacillus anthracis</i>	Antrax	Descoberta/Pré-clínica	(Ariel, Zvi <i>et al.</i> , 2002)
<i>Porphyromonas gingivalis</i>	Periodontites	Descoberta/Pré-clínica	(Ross, Czajkowski <i>et al.</i> , 2001)
<i>Mycobacterium tuberculosis</i>	Tuberculose	Descoberta/Pré-clínica	(Betts, 2002)
<i>Helicobacter pylori</i>	Úlcera, gastrite atrófica, adenocarcinoma, linfoma	Descoberta/Pré-clínica	(Chakravarti, Fiske <i>et al.</i> , 2000)
<i>Chlamydia pneumoniae</i>	Pneumonia, meningite, outras infecções.	Descoberta/Pré-clínica	(Montigiani, Falugi <i>et al.</i> , 2002)

Assim, o presente projeto descreve uma metodologia computacional combinada, com base nas predições das características mais interessantes para determinação de novos candidatos à vacina contra a esquistossomose em escala genômica. Mas é necessário primeiramente tecer algumas considerações a respeito da biologia do parasito, aspectos da terapêutica e imunologia da doença, e também de seu genoma, para que finalmente as predições sejam descritas. A abordagem descrita aqui não utilizará somente o genoma disponível do organismo *S. mansoni*, mas também os demais estudos de mapeamento de

informações biológicas para este parasito sejam eles: *Transcriptoma* e *Proteoma*. Espera-se a descoberta em curto tempo, de novos candidatos que possam integrar a lista de alvos existentes, a fim de compor uma vacina mais eficaz e segura para a doença.

2.1. *Schistosoma mansoni*

O *S. mansoni* é um parasito *trematódeo*, que vive na corrente sanguínea, cujo ciclo de vida mostra uma alternância de gerações entre o hospedeiro intermediário e os hospedeiros definitivos. Dentre os hospedeiros definitivos, estão alguns vertebrados, inclusive o homem e o hospedeiro intermediário para *S. mansoni* é o caramujo do gênero *Biomphalaria* (Davis, 1984). No caramujo, milhares de cercárias são produzidas por um único esporocisto durante a reprodução assexuada, as quais são liberadas de forma intermitente. A saída das cercárias do hospedeiro intermediário é induzida pela luz, pois ocorre preferencialmente, nas horas mais claras do dia. Quando madura, a cercária sai do caramujo para a água apta a infectar hospedeiros vertebrados. Uma vez que não se alimentam, as cercárias precisam encontrar um hospedeiro vertebrado dentro de um período de 6 a 8 horas (Davis, 1984), e quando o encontram, penetram pela sua pele ou mucosa, liberando a cauda, onde se inicia, então, o processo de transformação em esquistossômulos.

Na pele, após permanecerem por aproximadamente 72h, os esquistossômulos iniciam o processo de migração através do corpo do seu novo hospedeiro, caso não sejam destruídos pelos mecanismos de defesa do mesmo. Por meio da circulação, os esquistossômulos chegam ao coração e em seguida aos pulmões em torno de 5 dias, tornando-se mais longos e delgados, o que facilita a sua migração através da rede vascular pulmonar. Do pulmão, os esquistossômulos voltam ao coração e são enviados pela circulação geral a todas as partes do corpo do hospedeiro. Somente quando alcançam o sistema porta intra-hepático, são capazes de completar seu desenvolvimento (Rollinson e Simpson, 1988).

Quatro semanas após a infecção, a maioria dos vermes encontra-se na forma adulta e prontos para se acasalar. Acasalados, deslocam-se ativamente contra a corrente circulatória do sistema porta e migram para as veias mesentéricas pélvicas. Os vermes adultos têm, aproximadamente, um centímetro em comprimento, são delgados, longos e caracterizados por terem duas ventosas, e se alimentam de plasma e células do sangue venoso do hospedeiro

definitivo (Rollinson e Simpson, 1988). A fêmea é mais fina, cilíndrica e longa que o macho, existindo também diferenças fisiológicas e antigênicas entre macho e fêmea (Aronstein e Strand, 1984). A fêmea depende do contato do macho para completar sua maturação (Loveerde e Chen, 1991). O par permanece em constante acasalamento e a fêmea se localiza no canal ginecóforo do macho. Cerca de 340 ovos de *S. mansoni* são liberados por casal de vermes por dia em camundongos infectados, sendo este número maior em primatas (Cheever, Macedonia *et al.*, 1994).

A produção de ovos começa 30 a 40 dias após a infecção. No momento da oviposição, o desenvolvimento embrionário é ainda incompleto, requerendo mais seis ou sete dias para completar-se. A expectativa de vida dos ovos maduros é de, aproximadamente, 20 dias. Os ovos são eliminados no ambiente por meio das fezes de indivíduos infectados (Rollinson e Simpson, 1988). Os ovos eliminados nas fezes do hospedeiro eclodirão, se encontrarem condições adequadas como água fresca, temperatura morna, baixa hipotonicidade e iluminação adequada, liberando assim o miracídio de dentro da casca do ovo (Rollinson e Simpson, 1988). Uma vez que o ovo é rompido em água fresca, o miracídio emerge e começa a nadar ativamente. Os miracídios morrem caso a expulsão não se complete dentro de três semanas após a oviposição. O miracídio penetra no hospedeiro intermediário específico por movimentos rotatórios e ação lítica (Rollinson e Simpson, 1988).

Por volta do oitavo dia o miracídio apresenta-se como um tubo enovelado, imóvel, repleto de células germinativas em multiplicação, processo pelo qual se transforma em esporocisto. Por volta da segunda semana de existência os esporocistos rompem-se pra liberar esporocistos filhos, em número de 20 a 40. A transformação dos esporocistos em cercárias ocorre apenas quando atingem sua localização permanente nas glândulas digestivas do caramujo. A cercária deixa o hospedeiro e, caindo em água fresca, fecha o ciclo de vida do parasito (Rollinson e Simpson, 1988) (Figura 2)

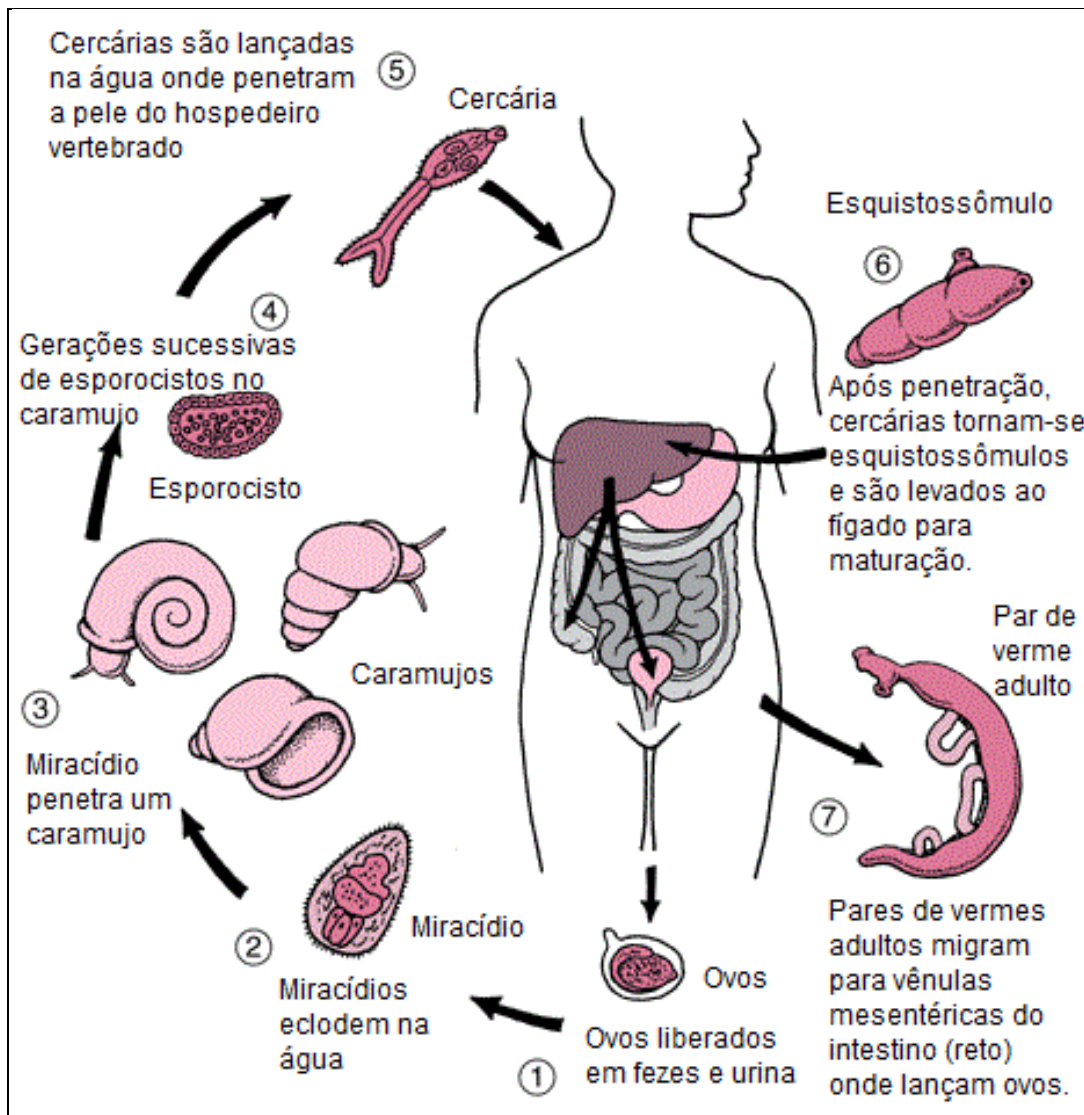


Figura 2 - Ciclo Biológico do *S. mansoni*

(1) Os ovos são eliminados junto com as fezes. (2) Sob condições ótimas, os ovos eclodem e liberam os miracídeos (3) que nadam e penetram no caramujo, hospedeiro intermediário específico. No caramujo, incluem duas gerações de esporocistos e a produção de cercárias (4). Ao abandonar o caramujo, as cercárias infectantes nadam, penetram na pele do hospedeiro humano (5), e perdem sua cauda bifurcada e tornam-se esquistossômulos (6). Os esquistossômulos migram através de diversos tecidos e desenvolvem-se até atingirem a veia porta onde se acasalam. Vermes adultos, nos humanos, residem nas vênulas mesentéricas em várias localizações (7), onde as fêmeas depositam ovos nas pequenas vênulas dos sistemas porta e perivesical. Os ovos são movidos progressivamente para o lúmen do intestino e são eliminados com as fezes (1) (Pearson, 2009).

2.2. Aspectos da terapêutica e da imunologia da esquistossomose

Hoje em dia, o controle da esquistossomose é feito por meio da utilização de drogas e tem sido considerada uma medida de controle eficaz, embora existam altos índices de reinfeção em áreas endêmicas. A importância das drogas não se limita ao tratamento da infecção e controle da morbidade; são também utilizadas para controle da transmissão da doença. A OMS recomenda a utilização do Praziquantel e Oxamniquina para controle da esquistossomose, sendo o PZQ a droga de escolha (Who, 2011). Tal recomendação ao PZQ se deve à sua eficácia contra todas as espécies de *Schistosoma* que infectam o homem, sua alta taxa de cura, baixa toxicidade e baixo custo.

O tratamento com PZQ é feito em dose única por via oral, com dosagem recomendada entre 40 e 60 mg/Kg. O percentual de cura da doença causada por *S. mansoni* é de 60% a 90%, causando redução do número de ovos de 90% a 95%, dependendo do nível de infecção (De Silva, Guyatt *et al.*, 1997). A eficácia do PZQ depende da idade e sexo do parasito, sendo os vermes fêmeas menos sensíveis ao PZQ (Pica-Mattoccia e Cioli, 2004). O tratamento com PZQ pode causar efeitos colaterais transitórios como vômito, cólicas e diarreia. A frequência e a gravidade dos efeitos colaterais produzidos pelo PZQ estão diretamente relacionadas com a intensidade da infecção (Stelma, Talla *et al.*, 1995).

Evidências acerca da resistência do parasito ao PZQ têm levantado questões sobre o uso da droga para o tratamento da esquistossomose, o que poderia torná-la inútil ao longo do tempo. Investigações recentes demonstraram casos de parasitas resistentes ao PZQ no Senegal e Egito, bem como de resistência induzida em condições de laboratório (Doenhoff, Cioli *et al.*, 2008).

Nos últimos anos, diversos trabalhos permitiram um melhor entendimento da resposta imune durante a esquistossomose (Pearce e Macdonald, 2002). Levando diversos grupos a investigarem possibilidades de desenvolvimento de vacinas contra a doença. Na esquistossomose o sistema imune do hospedeiro é exposto a uma série de antígenos derivados do parasito e do ovo que induzem intensa resposta celular e humoral. Na tentativa de se identificar os mecanismos patogênicos e protetores desencadeados pela infecção pelo *S. mansoni*, vários estudos foram desenvolvidos em modelos experimentais. Entre os vários modelos experimentais, o camundongo é o mais utilizado.

A relevância do papel dos linfócitos T na imunidade contra a infecção pelo *S. mansoni* foi descrita inicialmente em estudos utilizando camundongos timectomizados (Pearce e Macdonald, 2002). Esses autores mostraram que a resistência à reinfecção nestes animais foi significativamente reduzida, embora nenhuma diferença na recuperação de vermes da infecção primária tenha sido observada. A deficiência de células T afetou a eliminação dos vermes provenientes da infecção desafio, mas não a da infecção primária. Esses mesmos autores observaram que a diminuição do número de parasitas após a infecção desafio também ocorreu quando animais adultos timectomizados eram tratados com soro anti-timócitos, imediatamente após a infecção inicial. A dependência de células na resistência à infecção foi observada ainda através de estudos de transferência passiva de células de camundongos portadores de infecção crônica para animais. A transferência passiva de células em combinação com soro imune de animais portadores de infecção crônica para animais normais desafiados, também levou a uma redução significativa dos parasitas da infecção.

Vale à pena salientar que a investigação de mecanismos imunes dependentes de células T envolvidos na resistência ou patologia da doença, tornou-se ainda mais evidente (Pearce e Macdonald, 2002), após a descrição das subpopulações de células T auxiliaadoras, dos tipos TH0, TH1 e TH2. Enquanto células TH1 secretam principalmente IFN- γ , IL-2 e linfotóxina (LT) e estão associadas a reações de hipersensibilidade tardia (DTH), células TH2 secretam principalmente IL-4, IL-5 e IL-10, sendo mais efetivas como promotores na estimulação de linfócitos B para secreção de anticorpos (Pearce e Macdonald, 2002).

É importante mencionar que os mecanismos envolvidos na indução das respostas TH1 e TH2 na esquistossomose ainda não estão totalmente esclarecidos. Alguns estudos na esquistossomose experimental revelam a existência de interações complexas entre tipos celulares e demonstram a importância de moléculas acessórias além das citocinas no direcionamento dessas respostas (Pearce e Macdonald, 2002).

Ainda que muito tenha se aprendido sobre a resposta imune contra esquistossomose, métodos tradicionais de identificação de novos candidatos à vacina, não geraram os resultados esperados (como mencionados no início desta seção) e novas abordagens serão necessárias para o desenvolvimento de vacinas contra o *S. mansoni*.

Tendo em vista os novos dados gerados por pesquisas sobre o genoma deste parasito e investigações realizadas à luz de abordagens computacionais, novas medidas de controle poderão ser encontradas.

2.3. O genoma de *Schistosoma mansoni*

Os primeiros registros de sequenciamento de *S. mansoni* iniciaram-se em 1992 por meio de um projeto colaborativo entre instituições brasileiras, para sequenciamento de transcritos, financiado por agências nacionais (Franco, Valadao *et al.*, 2000). Após isso, em 1994 a OMS iniciou o projeto genoma de *S. mansoni*, com a premissa de que a descoberta de genes, desenvolvimento de mapas cromossômicos, sequenciamento genômico completo e análise genômica seriam as abordagens mais promissoras para a identificação de novos alvos para drogas, vacinas e ferramentas diagnósticas. Assim a Rede Genoma de *S. mansoni* recebeu financiamento para descoberta de tais alvos, e durante este período, a comunidade científica mundial produziu aproximadamente 16.000 ESTs (Oliveira e Johnston, 2001).

Posteriormente, dois grandes projetos brasileiros de sequenciamento do transcriptoma de *S. mansoni* foram realizados. O primeiro, que foi financiado pela FAPESP/MCT/CNPq, utilizou uma biblioteca normalizada de verme adulto e minibibliotecas de ORESTES de seis estágios do ciclo de vida do parasito (Verjovski-Almeida, Demarco *et al.*, 2003). O projeto gerou 124.681 ORESTES e ESTs de *S. mansoni*. As sequências foram agrupadas resultando em 30.988 SmAE (*Schistosoma mansoni* assembled ESTs). O segundo projeto, financiado pela FAPEMIG/MCT/CNPq, consistiu de uma rede genômica formada por instituições do Estado de Minas Gerais que caracterizou o transcriptoma de diferentes estágios de desenvolvimento do parasito a partir da geração de, aproximadamente, 42.500 ESTs convencionais (Oliveira, 2007).

A iniciativa internacional para o sequenciamento do genoma completo do organismo foi conduzida pelo instituto *The Institute for Genomic Research* - TIGR em associação com o *Welcome Trust Sanger Institute* – WTSI, por meio de financiamento do *National Institutes of Health* - NIH e da *Welcome Trust*, respectivamente (Loverde, Hirai *et al.*, 2004). As últimas versões do sequenciamento genômico e todas as análises efetuadas estão disponíveis nos bancos de dados online GeneDB (www.genedb.org) e SchistoDB (www.schistodb.net).

O parasito *S. mansoni* tem um genoma haploide de aproximadamente 363 mbp contidos em 7 pares de cromossomos autossômicos e um par de cromossomos sexuais Z e W. Em 2009 um estudo produziu um mapa genético do parasito (Criscione, Valentim *et al.*, 2009), o qual foi utilizado na montagem do genoma, em uma abordagem pioneira em platelmintos.

A sequência do genoma nuclear de *S. mansoni* foi obtida por meio da metodologia WGS (*Whole Genome Shotgun*). Essa técnica consiste na quebra de DNA genômico em pequenos fragmentos que foram selecionados por tamanho em gel de agarose. Os fragmentos foram agrupados em 5745 *scaffolds* com tamanho superior a 2kbp, totalizando 363mbp. Além disso, foram identificados 11812 genes que codificam 13162 transcritos. Apesar de 45% do genoma ser composto de elementos repetitivos, 50% das bases estão presentes em contigs de tamanho maior que 16,3 kbp e em *scaffolds* maiores que 824,5kbp. Ainda, a localização cromossomal de 43% da montagem genômica utilizando-se hibridização *in situ* foi identificada (Berriman, Haas *et al.*, 2009).

E mais recentemente alguns trabalhos produziram experimentos os quais apresentaram uma melhoria sistemática na montagem do genoma original. Um deles, o qual nosso grupo também possui co-autoria (Protasio, Tsai *et al.*, 2012), usando sequenciamento Sanger e a abordagem de nova geração Illumina em vermes clonados, foi capaz de atualizar o fragmentado genoma que antes possuía 5745 *scaffolds* para uma nova versão com apenas 885 *scaffolds*, e com mais de 80% das bases organizadas nos cromossomos. O trabalho também utilizou dados de transcriptoma (RNA-seq) para refinar a predição de genes, que foi reduzida de 11812 para 10852 genes, estes dados também estão disponíveis do GeneDB e no SchistoDB.

De posse de dados gerados por diferentes técnicas de biologia molecular, abordagens computacionais podem ser utilizadas com o intuito de se integrarem diferentes tipos de dados e, assim, possibilitar análises mais aprofundadas.

2.4. Predição da localização celular de proteínas

Em 1999, o Prêmio Nobel em Fisiologia/Medicina foi dado ao biólogo alemão Günter Blobel, por descobrir que as proteínas têm sinais intrínsecos que direcionam seu transporte e sua localização nas células. Como a localização subcelular de uma proteína tem um papel importante na caracterização de sua função, permitindo conhecimento tanto do seu papel biológico como o uso dessa informação na busca de novos alvos para drogas e vacinas, estudos de sublocalização celular tem se mostrado um dos maiores desafios em bioinformática (Emanuelsson, Brunak *et al.*, 2007).

Nos últimos anos, dezenas de métodos foram desenvolvidos para a determinação da localização celular de proteínas a partir de informações de sequência de DNA geradas em larga escala. Tanto métodos experimentais como métodos computacionais têm procurado novas abordagens na execução desta tarefa. Dentre os experimentais se destacam os de imunolocalização (Burns, Grimwade *et al.*, 1994), busca com fluorescência (Hanson e Kohler, 2001) e isótopos (Dunkley, Dupree *et al.*, 2004). Estes métodos embora tenham reconhecida precisão, são lentos e caros. Assim, os métodos computacionais têm ganhado bastante atenção da comunidade científica principalmente na rápida triagem de um grande conjunto de dados, como também nos casos em que proteínas são difíceis de isolar têm sua localização determinada pela sequência genômica.

Em termos gerais, os métodos computacionais para esse tipo de predição podem ser divididos em duas categorias: aqueles que utilizam somente a sequência de aminoácidos como dado de entrada para a predição e, outros que, além da sequência, requerem também outras informações como, por exemplo: dados de expressão (Drawid e Gerstein, 2000), perfis filogenéticos (Marcotte, Xenarios *et al.*, 2000), contexto léxico no banco de dados (Nair e Rost, 2002) ou termos definidos no Gene Ontology (GO-numbers) (Chou e Shen, 2006).

Como grande parte das proteínas preditas para *S. mansoni*, aproximadamente 40%, não tem função predita por similaridade de sequências e são, portanto, classificadas como hipotéticas, optamos pelas estratégias de análise *ab initio*, através das quais apenas as sequências de aminoácidos das sequências proteicas foram utilizadas.

Assim, no presente projeto, obtivemos o conjunto das proteínas do proteoma predito de *S. mansoni*, potencialmente expostas ao hospedeiro humano, utilizando a integração das

seguintes predições: predição das proteínas secretadas clássicas e não-clássicas, e, predição das proteínas transmembranas

2.4.1. Predição de proteínas secretadas

A predição de proteínas secretadas é basicamente determinada pela presença do mais conhecido sinal celular – o peptídeo sinal (PS). O PS é encontrado em todos os três domínios da vida. Ele marca a proteína para secreção através da membrana plasmática em procariotos e através do retículo endoplasmático (RE) em eucariotos (Von Heijne, 1990). O PS possui uma região N-terminal, tipicamente uma sequência de 15-aminoácidos, que é clivada durante a secreção da proteínas através da membrana. Não existe um consenso para o PS, mas eles tipicamente apresentam três zonas composicionais distintas: uma região N-Terminal que contém frequentemente resíduos carregados positivamente, uma região hidrofóbica e pelo menos 6 resíduos de uma região C-Terminal com resíduos polares não carregados.

Assim, grande parte das proteínas exportadas da célula possuem este sinal. Particularmente em eucariotos, as proteínas que contém o PS, são levadas à membrana do RE para serem secretadas por vesículas de secreção através do complexo de Golgi (CG).

Porém, nem toda proteína secretada possui o PS. Alternativamente existe um grupo de proteínas tanto em procariotos como em eucariotos que não seguem a via de secreção chamada de via clássica. São as chamadas proteínas secretadas não-clássicas. Em eucariotos existe um bem conhecido conjunto de proteínas que são exportadas fora da via RE-CG, por exemplo, fatores de crescimento de fibroblastos, interleucinas e galactinas (Bendtsen, Jensen *et al.*, 2004). E apesar deste fenômeno ter sido descoberto há mais de uma década, o mecanismo molecular de secreção continua desconhecido. No entanto, pode ser possível que neste mecanismo as proteínas abandonem as células após uma ruptura ou utilizem uma via de secreção ainda não definida.

Existem diversos métodos e abordagens para a predição de proteínas secretadas. No presente projeto, após uma extensa revisão da literatura em busca dos mais conhecidos e citados métodos foram escolhidos três: SignalP 3.0 (Bendtsen, Nielsen *et al.*, 2004) e SherLoc (Shatkay, Hoglund *et al.*, 2007), para a predição de proteínas secretadas clássicas e o

SecretomeP (Bendtsen, Jensen *et al.*, 2004), para a predição de proteínas secretadas não-clássicas.

O SignalP 3.0 é atualmente o mais popular algoritmo de predição de proteínas secretadas. Ele utiliza duas abordagens distintas em sua predição. Uma baseada em um método computacional de aprendizagem de máquina chamado Redes Neurais (*Neural Networks-NN*) e outra baseada em um modelo estatístico chamado de Modelo Oculto de Markov (*Hidden Markov Model-HMM*). (Bendtsen, Nielsen *et al.*, 2004).

As NNs possuem este nome, pois são inspiradas no processamento que ocorre nos neurônios. Elas não são programas de modo convencional, uma vez que são influenciadas por um conjunto de dados fornecido em um processo denominado de “treinamento”. Elas têm sido bastante utilizadas nas tarefas de classificação, reconhecimento de padrões e aproximação de funções. Já os HMMs têm sua metodologia baseada em um modelo estatístico, que consiste em um conjunto de estados, em que cada estado possui uma distribuição de probabilidade dos resultados possíveis, neste modelo é também atribuído um conjunto de probabilidade de transição para mudança de estados, desta forma o resultado de um evento, depende apenas do estado anterior. É chamado de oculto, pois através deste método é possível se obter um conjunto de eventos artificiais, no qual não é possível determinar de que estado teve origem, pois tal informação é ocultada pelo modelo (Lund, 2005).

O SherLoc diferente de outros métodos, que se concentram em localizações específicas, tenta cobrir uma ampla gama de localizações celulares (Shatkay, Hoglund *et al.*, 2007). São 09 localizações celulares possíveis para eucariotos na versão 2 da ferramenta. O SherLoc integra vários preditores baseados em sequências e traz uma novidade que é um preditor baseado em *Text-Mining* (mineração de textos). O sistema integra seus preditores utilizando SVM (*Support Vector Machine* – Máquina de vetores de suporte), que é uma abordagem de aprendizado de máquina, que analisam os dados e reconhecem padrões, usado para classificação e análise de regressão. O SVM padrão toma como entrada um conjunto de dados e prediz, para cada entrada dada, qual de duas possíveis classes a entrada faz parte, o que faz do SVM um classificador linear binário não probabilístico. Dados um conjunto de exemplos de treinamento, cada um marcado como pertencente a uma de duas categorias, um algoritmo de treinamento do SVM constrói um modelo que atribui novos exemplos a uma categoria ou outra. Um modelo SVM é uma representação de exemplos como pontos no espaço, mapeados de maneira que os exemplos de cada categoria sejam divididos por um

espaço claro que seja tão amplo quanto possível. Os novos exemplos são então mapeados no mesmo espaço e preditos como pertencentes a uma categoria baseados em qual o lado do espaço eles são colocados (Lund, 2005).

O último algoritmo utilizado na determinação das proteínas secretadas foi o SecretomeP, ele concentra seu esforço na predição de proteínas secretadas não-clássicas. Apenas um grupo limitado de proteínas têm demonstrado experimentalmente não entrar na via de secreção clássica (Bendtsen, Jensen *et al.*, 2004), não possuem nenhum sinal característico e nem motivos em comum. No entanto, os autores deste método descobriam algumas características independentes que são compartilhadas por proteínas exportadas, por exemplo: a composição de aminoácidos, a estrutura secundária e as regiões de desordem, e são neste conjunto de características que se baseia este algoritmo.

2.4.2. Predição de proteínas transmembranas

A predição de proteínas transmembranas é realizada através da busca de domínios de membrana conhecidos em proteínas. Podemos destacar dois métodos de predição dessas características: o TMHMM (Sonnhammer, Von Heijne *et al.*, 1998) e o SherLoc (já citado).

O TMHMM é um método para prever a localização e orientação das alfa-hélice em proteínas de membrana. É baseado em HMM com uma arquitetura que corresponde praticamente ao sistema biológico. O modelo é tão próximo da realidade que nos permite inferir quais as partes da arquitetura do modelo são importantes para capturar as informações que codifica a topologia da membrana, e para obter uma melhor compreensão dos mecanismos e restrições envolvidas (Sonnhammer, Von Heijne *et al.*, 1998).

2.5. Predição de epitopos

O sistema imune de vertebrados pode ser dividido, de modo simplificado, em dois ramos: Inato (ou Natural) e Adaptativo (ou Adquirido) (Figura 3). Na imunidade inata, envolve componentes humorais pré-formatados para a proteção contra infecções e doenças. Por outro lado, a imunidade adquirida é aquela que o organismo desenvolve após ter contatos com agentes patológicos externos. Existe também uma grande diferença quanto ao tempo de ativação da resposta inata e adaptativa, na primeira a resposta ocorre no período de horas, enquanto na segunda se dá em dias (Abbas, Lichtman *et al.*, 2007).

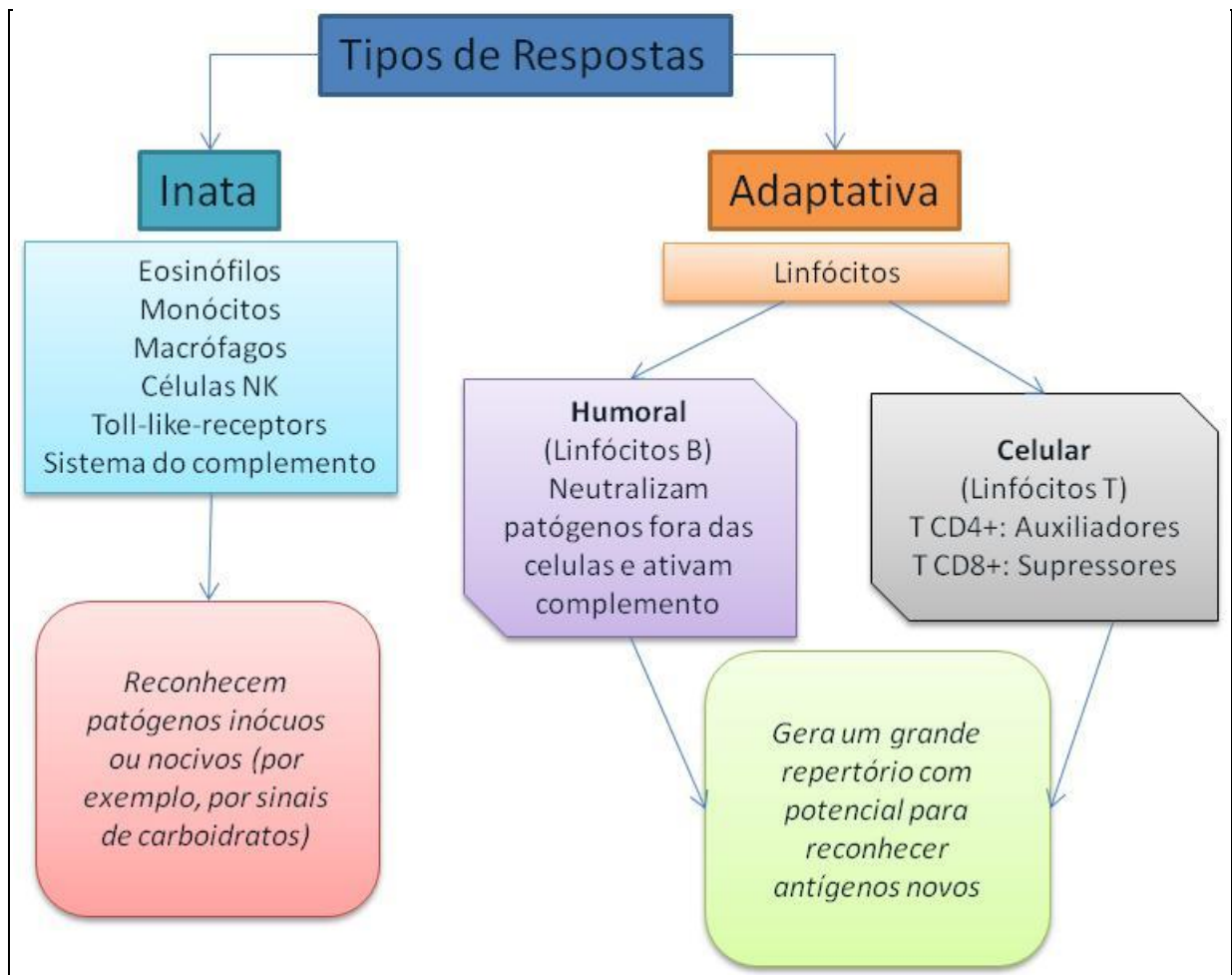


Figura 3 - Tipos de Respostas

Representação da divisão do tipo de respostas do sistema imune humano, apresentando seus componentes e função (Lund, 2005).

O sistema imune adaptativo apresenta, de forma didática, dois tipos de resposta: a humoral e a celular. A resposta humoral é mediada por linfócitos B e anticorpos, que neutralizam patógenos fora das células humanas. Enquanto na resposta celular é mediada por linfócitos T citotóxicos, CD8+, que eliminam células infectadas. Focando apenas a resposta celular, sabemos que o sistema imune humano reconhece via TCR (receptores de células T) e apresenta via MHC (Complexo de Histocompatibilidade Principal), antígenos produzidos por vírus, bactérias, parasitos ou mesmo proteínas super-expressas ou mutadas em suas células pela via do MHC (Figura 4).

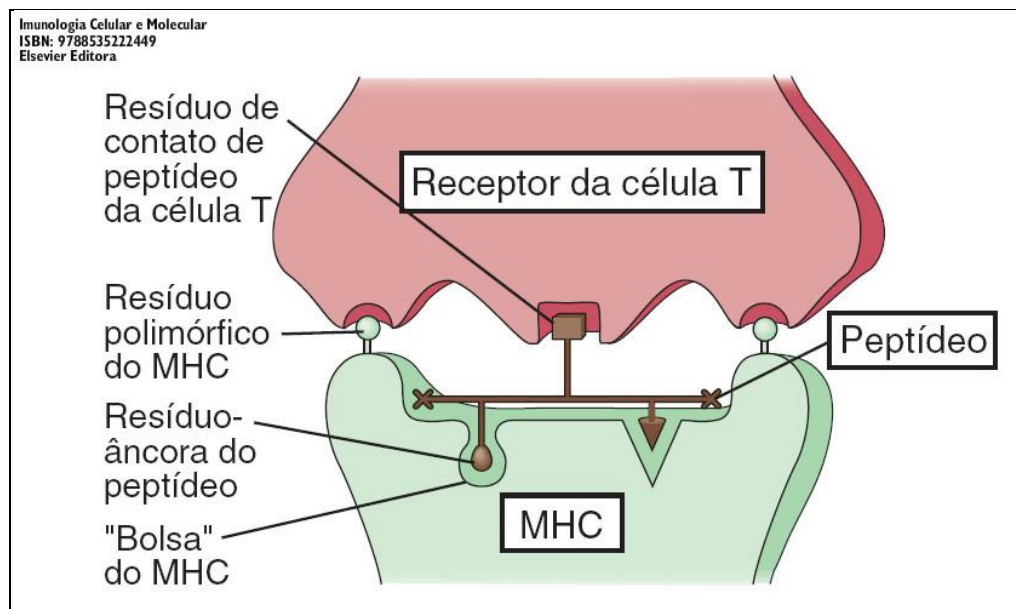


Figura 4 – A molécula de MHC

O MHC - Complexo de Histocompatibilidade Principal. Reconhecimento de um peptídeo pela molécula de MHC. Fonte: (Abbas, Lichtman *et al.*, 2007)

O *locus* MHC é uma região genômica e seus produtos cujo papel principal é a estimulação da resposta imune através dos linfócitos T. Em humanos o MHC é também chamado de HLA (*Human Leucocyte Antigen*) e possui centenas de alelos, o que o torna o gene extremamente polimórfico. Existem duas classes de MHC: O MHC de Classe I e o MHC de Classe II.

O MHC de Classe I detecta e apresenta aos linfócitos T CD8+, peptídeos de origem intracelular gerados através da degradação na via do proteassoma. Estes peptídeos ligam-se a

uma molécula chamada TAP que o transloca até o RE e lá são ligados ao MHC Classe I que os apresenta na superfície celular.

Já o MHC Classe II têm seus peptídeos de origem extracelular, onde proteínas são endocitadas, processados internamente e degradados nos endossomos e nos lisossomos para geração dos peptídeos, cuja a cadeia invariante será ligada à molécula MHC Classe II produzida no retículo endoplasmático, que o transporta até a superfície celular para apresentação as células T CD4+.

Os linfócitos T CD4+, são células especializadas em ativar outras células do sistema imune, como os macrófagos, que são ativados para matar os patógenos intravesiculares, e as células B, para secretarem imunoglobulinas.

O reconhecimento do complexo MHC-peptídeo pelo receptor de célula T (TCR) que acontece na superfície dos linfócitos T citotóxicos (CTLs) deflagra a citólise da célula apresentando o peptídeo.

Uma forma de estimular a resposta imune é, portanto, a administração de peptídeos derivados de antígenos que são reconhecidos pelo MHC. Esses peptídeos, chamados de epitopos, são frequentemente resíduos de 8-25 aminoácidos.

Pelo fato de a via de desenvolvimento de vacinas implicar em gastos orçamentários elevados e ser bastante laboriosa e lenta, existe um grande incentivo no desenvolvimento de tecnologias computacionais que selecionem candidatos potenciais antes do desenvolvimento experimental.

Apesar de cada etapa da via de processamento de antígeno adicionar uma especificidade à seleção do antígeno, a incorporação de peptídeos do complexo MHC parece ser um ponto essencial e discriminante no processo (Lauemoller, Kesmir *et al.*, 2000). Pois, a identificação desses epitopos imunodominantes é extramamente importante, não só porque torna a imunização possível, excluindo o risco de uma reatividade cruzada ou o desenvolvimento de uma auto-imunidade ou alergia, mas também porque permite a construção de vacinas com peptídeos relevantes de antígenos candidatos diferentes, melhorando a chance de alcançar altos níveis de proteção (Fonseca, Cunha-Neto *et al.*, 2005). Dessa forma, os métodos computacionais têm se concentrado na predição da afinidade de ligação de peptídeos candidatos ao complexo MHC.

Porém, vale ressaltar que a escolha de um algoritmo para predição de epitopos é uma tarefa crítica e não trivial. Existem vários métodos em que cada método está associado a um grupo de alelos do MHC específico sob no qual foi treinado, e as diferentes os resultados fornecidos, tornam comparações de métodos uma tarefa complexa.

3. OBJETIVOS

3.1. Gerais

1. Identificar, no genoma de *S. mansoni*, proteínas potencialmente expostas ao sistema imune do hospedeiro vertebrado;
2. Mapear epitopos em moléculas classificadas como candidatas potenciais ao desenvolvimento de vacinas.

3.2. Específicos

1. Avaliar os diferentes algoritmos existentes utilizados para predição da localização subcelular de proteínas.
2. Estabelecer uma metodologia em que os resultados das predições serão combinados e a consistência das predições serão avaliadas;
3. Predizer e mapear epitopos ligantes ao Complexo de Histocompatibilidade Principal (MHC) Classe II.
4. Construir um banco de dados com epitopos que possam ser minerados para a busca de alvos para o desenvolvimento de vacinas;
5. Integrar as predições *in silico* relacionadas à identificação de potenciais alvos para o desenvolvimento de vacinas com os dados de proteoma e transcriptoma gerados no grupo de pesquisa em um banco de dados;
6. Validar experimentalmente os alvos selecionados.
7. Implementar um *pipeline* de análise computacional integrando as tarefas descritas acima.

4. MATERIAIS E MÉTODOS

4.1. Obtenção do proteoma predito

Como etapa inicial deste trabalho a obtenção do proteoma predito que será alvo de todas as análises do projeto, foi feita através do SchistoDB (Zerlotini, Heiges *et al.*, 2009). O SchistoDB (Figura 5) é um banco de dados genômico desenvolvido por nosso grupo e disponível na internet (www.schistodb.net) para toda comunidade científica. Tal repositório foi criado para armazenar e permitir a integração e análise de diversos dados disponíveis para o gênero *Schistosoma*.

O banco armazena a última versão do genoma do parasito de *S. mansoni* (5.0), a primeira versão do genoma de *Schistosoma haematobium* (contigs) e a versão 3.0 do genoma de *Schistosoma japonicum*. O SchistoDB também armazena o genoma mitocondrial, cepa NMRI16, obtida do NCBI e alguns experimentos de RNA-Seq de *S. mansoni* e *S. haematobium*. Assim, o banco provê acesso a um total de 13.273 genes preditos e anotados automaticamente de *S. mansoni*, que foram baixados através da ferramenta Sequence Retrieval (também disponível no banco), em formato de sequência de aminoácidos, em um arquivo FASTA, para compor o *dataset* inicial deste estudo (Figura 6).

Figura 5 - SchistoDB

Página inicial do SchistoDB. Aproximadamente 30 buscas estão disponíveis para possibilitar buscas nos dados de *S. mansoni*, além de acesso à diversas ferramentas. As buscas podem ser combinadas através do histórico de queries ampliando consideravelmente a capacidade de filtragem dos dados.

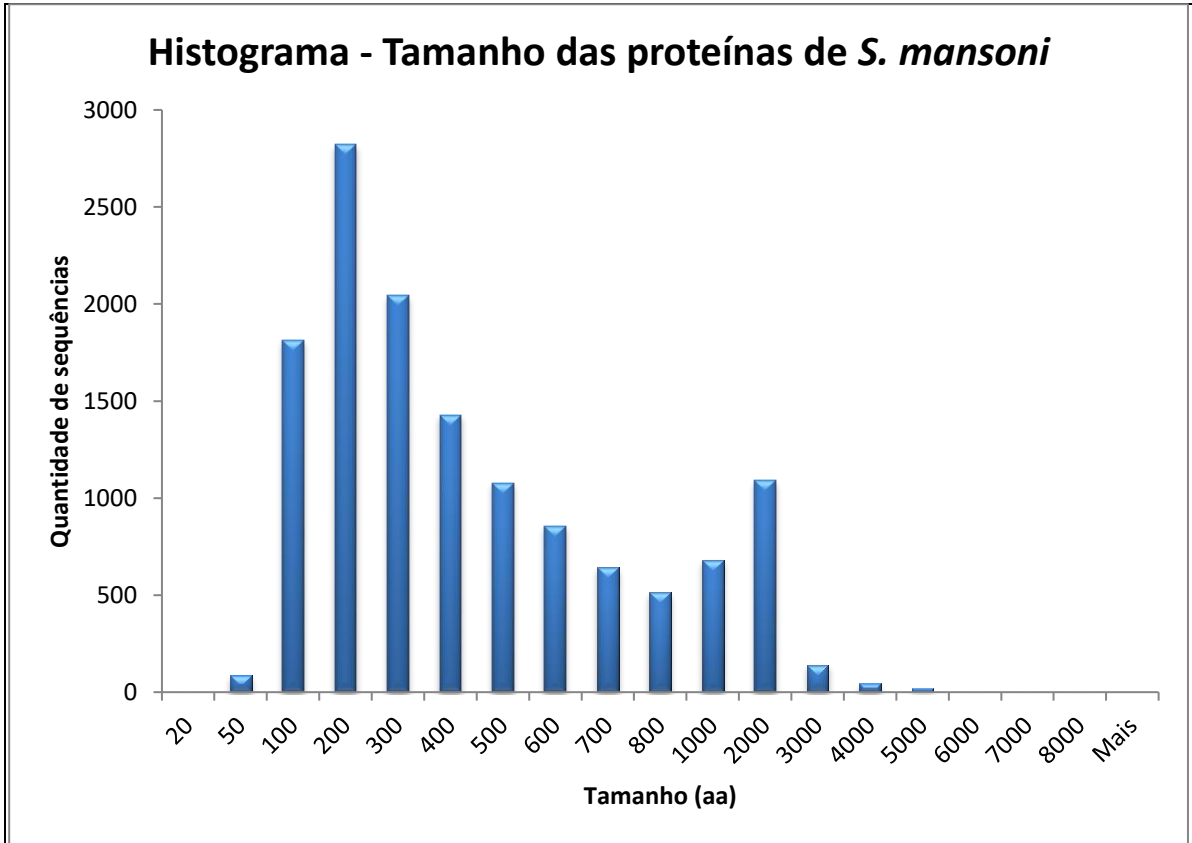


Figura 6 - Tamanho das proteínas de *S. mansoni*

Histograma do tamanho das proteínas de *S. mansoni* obtidas através do SchistoDB. Podemos observar que grande parte das seqüências possui de 100 a 500 aminoácidos (70% das 13.273 proteínas)

4.2. Filtragem das proteínas por fase do ciclo de vida

Devido o interesse de se trabalhar apenas com as proteínas que possuem evidência de expressão nas fases do hospedeiro humano, e a fase de morbidade da doença, (Figura 7), houve a necessidade de realizar uma filtragem de proteínas por fase específica.

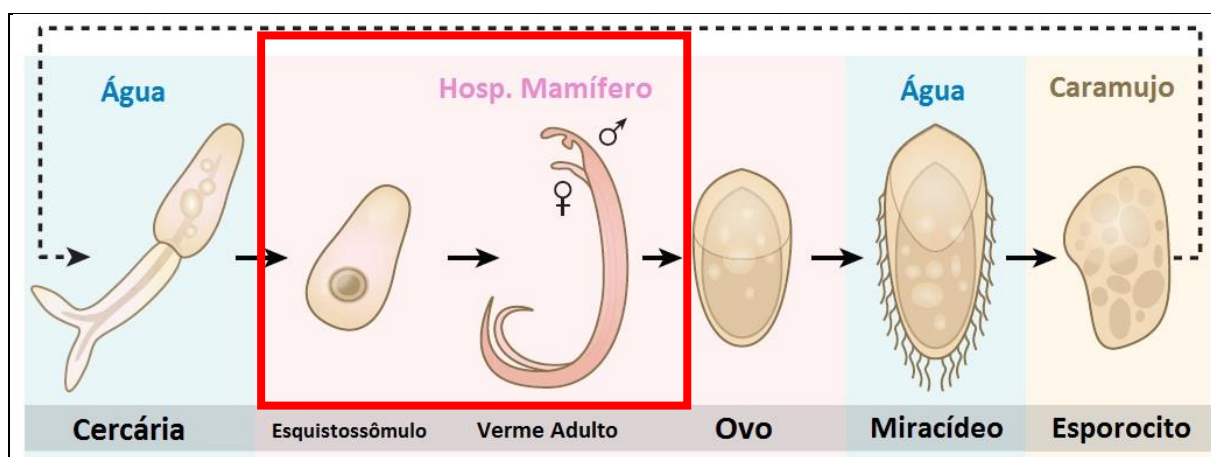


Figura 7 - Fases do ciclo de vida dos alvos.

Estágios de desenvolvimento no ciclo de vida do parasito *S. mansoni*, destacando as fases alvo em hospedeiro mamífero, que serão utilizadas no projeto. Fonte adaptada: (Han, Brindley et al., 2009)

Tal filtragem também foi realizada com ajuda do SchistoDB, o qual provê uma busca por genes utilizando dados de ESTs mapeadas no genoma (Find Genes by EST Evidence by Stage). A busca permite obter todos os genes com evidência de ESTs nos 6 (seis) estágios de desenvolvimento do verme. O mapeamento dessas ESTs foi realizado através da ferramenta BLAT de 205.892 ESTs obtidas no NCBI-DbEST (Ncbi-Dbest, 2008). O BLAT, em comparação a outros métodos de busca por similaridade, permite alinhar rapidamente sequências de 25 ou mais nucleotídeo, com 95% de similaridade (Kent, 2002). De posse desse mapeamento do SchistoDB, realizados umas consulta SQL diretamente na base de dados do site para obter apenas as fases do ciclo de vida que nos interessam. Tal consulta resultou numa tabela SQL (Anexos – Tabela 1) a qual foi normalizada para obtenção das fases que tínhamos interesse.

Foram considerados apenas aqueles genes que tinham evidência de ESTs das fases de *Esquistossômulo* e *Verme Adulto*. Também foram considerados os genes que não possuíam *Nenhuma Evidência de Estágio* de expressão, afim de não perder nenhum gene com potencial imunogênico, embora não tivesse evidência de qual estágio é expresso no parasito. Os genes com evidência de expressão exclusivamente na fase de *Ovo* foram descartados, pois embora façam parte do ciclo no hospedeiro mamífero, não se deseja uma vacina para a fase crônica da doença, a qual é caracterizada pela presença destes ovos nos tecidos, que resulta na formação de granulomas, os quais já são resultantes da resposta do sistema imune do hospedeiro (Katz, Faria *et al.*, 1986). O órgão mais atingido neste processo é o fígado, sendo o baço também um órgão bastante acometido.

4.3. Recursos computacionais para as predições e para desenvolvimento do pipeline.

Todas as predições foram realizadas em um servidor do CEBio (Dell Poweredge 2950III) com as seguintes características: Processador: Intel 2X Xeon X5460 Quad-Core de 3.16 Ghz, 6MB de memória cache, 24 GB RAM, 1,5TB HD. Os programas usados foram obtidos junto aos autores, instalados e executados no supracitado servidor. A Tabela 3 apresenta um resumo de todos os programas e onde podem ser obtidos.

Tabela 3**Lista de programas e pacotes usados para as predições computacionais.**

Algoritmo	Versão	Autores	Disponível em:
SignalP	3.0	(Bendtsen, Nielsen <i>et al.</i> , 2004)	http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?signalp
SecretomeP	1.0	(Bendtsen, Jensen <i>et al.</i> , 2004)	http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?secretomep
SherLoc	2.0	(Shatkay, Hoglund <i>et al.</i> , 2007)	http://abi.inf.uni-tuebingen.de/Services/SherLoc2/sherloc2_download
TMHMM	2.0c	(Sonnhammer, Von Heijne <i>et al.</i> , 1998)	http://www.cbs.dtu.dk/cgi-bin/nph-sw_request?tmhmm
FRED	1.0	(Feldhahn, Donnes <i>et al.</i> , 2009)	http://abi.inf.uni-tuebingen.de/Software/FRED

Para a execução remota das predições, bem como desenvolvimento do pipeline e web-server associado, foi usada estação de trabalho do CEBio (Dell Opitplex 745) com as seguintes características: Processador: Intel Core 2 Duo de 2.66 Ghz, 4MB de memória cache L2, 4GB RAM e 500GB HD.

4.4. Predição proteínas secretadas

A obtenção das proteínas secretadas foi realizada em um processo utilizando 3 algoritmos distintos já discutidos na Seção 2. Na primeira etapa usamos o SignalP para identificação das proteínas secretadas clássicas. A predição foi realizada submetendo o arquivo (FASTA) de entrada com todas as proteínas das fases de interesse. Os parâmetros utilizados foram: -t euk (o organismo a ser usado é um eucarioto), -short (a saída deve ser um arquivo tabular – facilitando a extração por scripts), -method hmm+nn (usaremos os algoritmos de HMM e NN) e -trunc 70 (serão considerados os primeiros 70 resíduos de aa para a predição – de acordo com dados da literatura (Bendtsen, Nielsen *et al.*, 2004)). Em seguida usamos o SherLoc com os parâmetros padrões, no mesmo arquivo de entrada. As

Figuras 8 e 9, trazem exemplos da saída destes programas, os quais extraímos os resultados positivos de ambos os métodos através de scripts desenvolvidos em Perl (Perl, 2010).

#	name	Cmax	pos	?	Ymax	pos	?	Smax	pos	?	Smean	?	D	?
Smp_000020		0.336	22	Y	0.106	1054	N	0.199	1049	N	0.004	N	0.055	N
Smp_000030.1		0.144	803	N	0.051	886	N	0.343	874	N	0.007	N	0.029	N
Smp_000030.2		0.146	34	N	0.109	34	N	0.348	836	N	0.055	N	0.082	N
Smp_000030.3		0.144	542	N	0.051	625	N	0.347	1	N	0.013	N	0.032	N
Smp_000040		0.454	98	Y	0.046	38	N	0.429	3	N	0.081	N	0.064	N
Smp_000050		0.204	857	N	0.220	857	N	0.760	759	N	0.031	N	0.125	N
Smp_000070		0.085	22	N	0.169	22	N	0.988	6	Y	0.895	Y	0.532	Y
Smp_000080		0.562	70	Y	0.123	98	N	0.885	94	Y	0.083	N	0.103	N
Smp_000090		0.171	24	N	0.105	24	N	0.293	2	N	0.123	N	0.114	N
Smp_000100		0.165	407	N	0.125	148	N	0.525	135	N	0.019	N	0.072	N
Smp_000130.1		0.108	22	N	0.037	22	N	0.120	553	N	0.041	N	0.039	N
Smp_000130.2		0.092	238	N	0.036	43	N	0.122	529	N	0.031	N	0.034	N
Smp_000130.3		0.092	212	N	0.126	17	N	0.513	6	N	0.309	N	0.217	N
Smp_000130.4		0.092	238	N	0.036	43	N	0.122	529	N	0.031	N	0.034	N
Smp_000140.1		0.088	22	N	0.162	22	N	0.838	12	N	0.372	N	0.267	N
Smp_000140.2		0.085	30	N	0.146	18	N	0.574	11	N	0.302	N	0.224	N
Smp_000150		0.112	37	N	0.039	37	N	0.088	1	N	0.018	N	0.028	N
Smp_000170		0.065	37	N	0.031	37	N	0.045	3	N	0.016	N	0.024	N
Smp_000190		0.290	35	N	0.410	35	Y	0.824	32	N	0.452	N	0.431	Y
Smp_000210		0.146	16	N	0.104	16	N	0.321	1	N	0.104	N	0.104	N
Smp_000220		0.029	167	N	0.023	9	N	0.083	1	N	0.031	N	0.027	N
Smp_000230		0.300	23	N	0.172	23	N	0.562	1	N	0.161	N	0.166	N
Smp_000240.1		0.386	17	Y	0.144	17	N	0.648	301	N	0.074	N	0.109	N
Smp_000240		0.386	17	Y	0.142	17	N	0.220	12	N	0.073	N	0.107	N

Figura 8 - Saída do Signalp (Proteínas secretadas)

O Resultado do Signalp compreende 3 diferentes *scores*: C, Y, S e mais dois *scores* adicionais: S-mean e D-Score, os quais são apenas numéricos. O S-score indica a predição de peptídeo sinal, é reportado para cada posição de um único aminoácido na sequência apresentada, com altos valores indicando que o aminoácido correspondente faz parte de um peptídeo sinal, e baixos valores, indicando que o aminoácido é parte de uma proteína madura. O C-score é a pontuação reportada ao sítio de clivagem. Para cada posição na sequência apresentada, um C-score é relatado, que só deve ser significativamente alto no local de clivagem. Ambos (C-score e S-score) são computados pelo algoritmo baseado em uma Rede Neural multicamadas, que apresenta valores de probabilidade entre 0 e 1 para todos os aminoácidos da sequência de entrada. Y-max é derivado da combinação do C-score com o S-score, resultando em uma melhor predição do sítio clivagem no lugar do C-score sozinho. O S-mean é a média do S-score, que vão desde o aminoácido N-terminal ao aminoácido atribuído com a mais alta pontuação Y-max, assim, a pontuação S-mean é calculada para indicar o comprimento do peptídeo sinal predito. O D-score, no nosso caso o mais importante, que faz a melhor discriminação de proteínas secretada, daquela que não é secretada e é calculado como a média da pontuação média entre o S-score e o máximo Y-score.

```

romulo@acara:~/doutorado2010/data
SherLoc2 Prediction Result

origin = animal

Smp_078570      cytoplasmic: 0.67      nuclear: 0.26      mitochondrial: 0.03
Smp_063300      cytoplasmic: 0.9       nuclear: 0.08      Golgi apparatus: 0.01
Smp_145060      peroxisomal: 0.52     extracellular: 0.19 cytoplasmic: 0.1
Smp_045160      nuclear: 1.0          cytoplasmic: 0.0   Golgi apparatus: 0.0
Smp_024390.1    Golgi apparatus: 0.51 plasma membrane: 0.29 ER: 0.1 lysosoma
Smp_007900.2    cytoplasmic: 0.94     mitochondrial: 0.03 nuclear: 0.02
Smp_063040.1    cytoplasmic: 0.97     nuclear: 0.03      mitochondrial: 0.0
Smp_126600      cytoplasmic: 0.55     nuclear: 0.21      plasma membrane: 0.1
Smp_132740      cytoplasmic: 0.53     mitochondrial: 0.32 peroxisomal: 0.1
Smp_055760      mitochondrial: 0.93    cytoplasmic: 0.03 peroxisomal: 0.0
Smp_012010      nuclear: 0.99          cytoplasmic: 0.01 mitochondrial: 0.0
Smp_181360      extracellular: 0.75    peroxisomal: 0.08 cytoplasmic: 0.0
Smp_010260      nuclear: 0.97          cytoplasmic: 0.03 mitochondrial: 0.0
Smp_045420      cytoplasmic: 0.5       peroxisomal: 0.43 Golgi apparatus:
Smp_007180      nuclear: 0.65          cytoplasmic: 0.3   mitochondrial: 0.02
Smp_105680.1    ER: 0.89              Golgi apparatus: 0.09 lysosomal: 0.0 extracel
Smp_135480      cytoplasmic: 0.85     nuclear: 0.07      peroxisomal: 0.05
Smp_102690.2    nuclear: 0.74          cytoplasmic: 0.22 mitochondrial: 0.03
Smp_176470      plasma membrane: 0.6   cytoplasmic: 0.23 nuclear: 0.09
Smp_085240.2    plasma membrane: 0.31 extracellular: 0.24 ER: 0.2 Golgi ap
Smp_166420      cytoplasmic: 0.75     nuclear: 0.23      peroxisomal: 0.01
Smp_155270      cytoplasmic: 0.78     nuclear: 0.13      mitochondrial: 0.06
Smp_002160      nuclear: 0.99          cytoplasmic: 0.01 mitochondrial: 0.0
Smp_069170      Golgi apparatus: 0.38 plasma membrane: 0.37 ER: 0.19
Smp_075110      nuclear: 0.76          cytoplasmic: 0.2   mitochondrial: 0.03
Smp_179800      cytoplasmic: 0.98     nuclear: 0.01      Golgi apparatus: 0.0
Smp_129430      nuclear: 1.0           cytoplasmic: 0.0   mitochondrial: 0.0

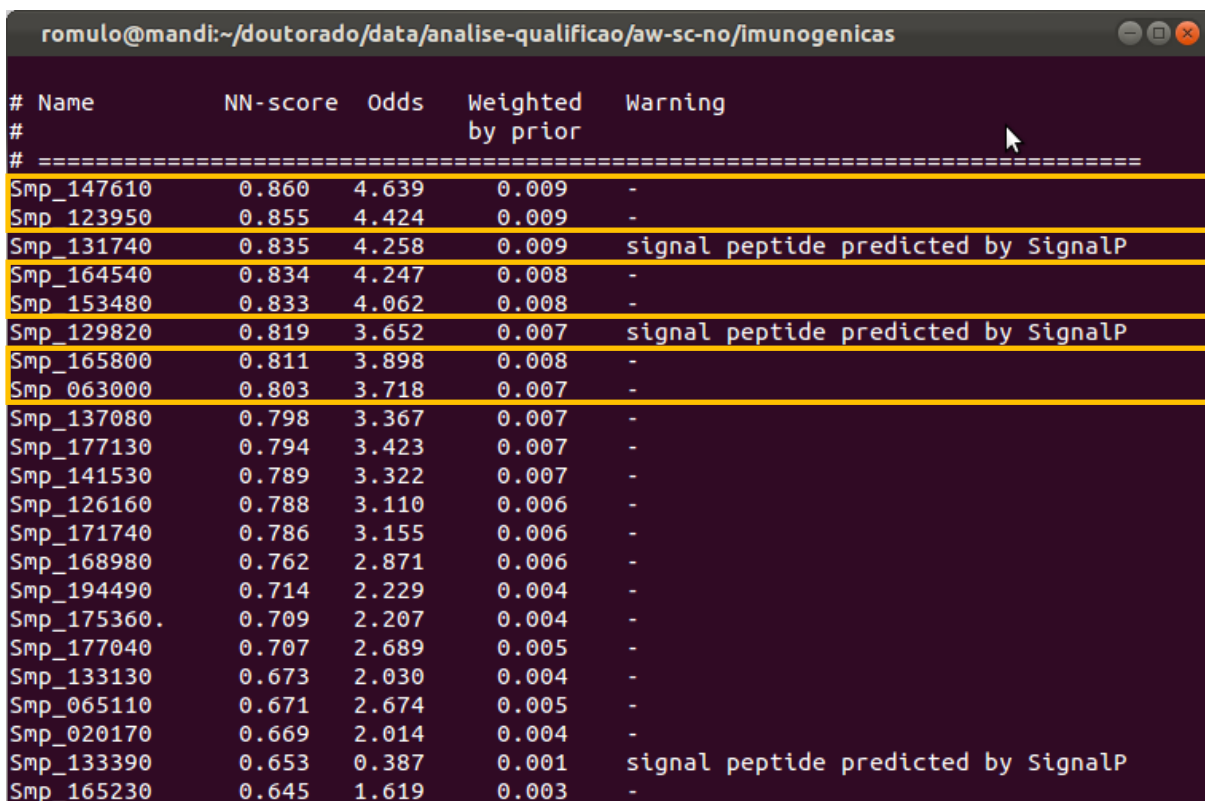
```

Figura 9 - Saída do SherLoc (Proteínas secretadas)

O Resultado do SherLoc, traz o a probabilidade de localização para cada proteínas em 9 possibilidades. Neste exemplo uma proteína predita como secretada. O método do SherLoc obtém tais valores através da integração de 4 preditores do MultiLoc (Hoglund, Donnes *et al.*, 2006), 3 destes baseados em SVM (já discutidos na seção 2.4.1), os quais apresentam o resultado da predição através de um vetor de probabilidades, e 1 preditor binário de *motifs*, que indica presença ou ausência de motivos conhecidos na sequência apresentada. O método ainda traz outro preditor desenvolvido também sob SVM, mas baseado em *Text-Mining* que também gera um vetor de probabilidades como resultado, mas é obtido através de termos da literatura que referenciam a sequência apresentada. Um termo não necessariamente inclui o nome da organela que ele representa. Mas tende a ocorrer em documentos que a referenciam. O SherLoc possui um classificador SVM final, o qual integra os vetores que resultam em cada preditor para a apresentação do resultado final.

Na segunda etapa utilizamos o algoritmo SecretomeP para identificação daquelas proteínas ditas secretadas não-clássicas. O programa funciona de forma simples, o único parâmetro utilizado foi a opção `-s`, a qual indica o nome do arquivo fasta de proteínas. Os

resultados positivos também foram extraídos por scripts e o exemplo da saída desse método pode ser visto da Figura 10.



# Name	NN-score	Odds	Weighted by prior	Warning
# =====				
Smp_147610	0.860	4.639	0.009	-
Smp_123950	0.855	4.424	0.009	-
Smp_131740	0.835	4.258	0.009	signal peptide predicted by SignalP
Smp_164540	0.834	4.247	0.008	-
Smp_153480	0.833	4.062	0.008	-
Smp_129820	0.819	3.652	0.007	signal peptide predicted by SignalP
Smp_165800	0.811	3.898	0.008	-
Smp_063000	0.803	3.718	0.007	-
Smp_137080	0.798	3.367	0.007	-
Smp_177130	0.794	3.423	0.007	-
Smp_141530	0.789	3.322	0.007	-
Smp_126160	0.788	3.110	0.006	-
Smp_171740	0.786	3.155	0.006	-
Smp_168980	0.762	2.871	0.006	-
Smp_194490	0.714	2.229	0.004	-
Smp_175360	0.709	2.207	0.004	-
Smp_177040	0.707	2.689	0.005	-
Smp_133130	0.673	2.030	0.004	-
Smp_065110	0.671	2.674	0.005	-
Smp_020170	0.669	2.014	0.004	-
Smp_133390	0.653	0.387	0.001	signal peptide predicted by SignalP
Smp_165230	0.645	1.619	0.003	-

Figura 10 - Saída do SecretomeP (Proteínas secretadas não clássicas)

O Resultado do SecretomeP, indicando a probabilidade de uma determinada proteína ser secretada (consideramos o *score* igual ou superior 0.8 para proteínas secretadas). Tais valores são obtidos através de uma rede neural construída a partir de padrões identificados em proteínas secretadas não-clássicas.

4.5. Predição das proteínas transmembranas

Para a determinação das proteínas com domínios transmembranas utilizados o TMHMM, no arquivo fasta com as proteínas também filtradas por fase. O parâmetro -short foi utilizado para garantir o arquivo de saída em formato tabular. A Figura 11, traz o exemplo de saída deste método. O SherLoc também foi usado na predição de transmembranas e os resultados de ambos os algoritmos foram extraídos por scripts.


```
romulo@mandi:~/doutorado/data/analise-qualificao/aw-sc-no
```

Smp_078570	len=357	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
Smp_158080	len=76	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
Smp_063300	len=102	ExpAA=0.00	First60=0.00	PredHel=0	Topology=i
Smp_145060	len=1196	ExpAA=35.36	First60=0.00	PredHel=2	Topology=o898
Smp_045160	len=706	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
Smp_024390.1	len=158	ExpAA=44.38	First60=38.02	PredHel=2	Topology=o15-37i44-60
Smp_007900.2	len=116	ExpAA=0.10	First60=0.09	PredHel=0	Topology=o
Smp_135070	len=86	ExpAA=0.44	First60=0.44	PredHel=0	Topology=o
Smp_122370.2	len=1603	ExpAA=0.18	First60=0.00	PredHel=0	Topology=o
Smp_063040.1	len=409	ExpAA=0.14	First60=0.00	PredHel=0	Topology=o
Smp_049930	len=628	ExpAA=28.16	First60=18.94	PredHel=1	Topology=i30-49o
Smp_126600	len=2016	ExpAA=0.01	First60=0.00	PredHel=0	Topology=o
Smp_132740	len=490	ExpAA=68.73	First60=0.08	PredHel=3	Topology=o164-186i302
Smp_055760	len=602	ExpAA=20.15	First60=0.01	PredHel=1	Topology=i119-138o
Smp_139350	len=1062	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
Smp_141410	len=1832	ExpAA=23.29	First60=0.00	PredHel=1	Topology=o154i
Smp_175210	len=102	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
Smp_169250	len=230	ExpAA=0.04	First60=0.00	PredHel=0	Topology=o
Smp_129000	len=321	ExpAA=21.58	First60=0.00	PredHel=1	Topology=o168-190i
Smp_128010	len=274	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
Smp_079640	len=540	ExpAA=257.58	First60=21.19	PredHel=12	Topology=i36-58o78-10
7-219i287-304o314-336i343-361o371-393i406-428o438-460i					
Smp_161540	len=198	ExpAA=45.55	First60=20.05	PredHel=2	Topology=i40-62o90-11
Smp_012010	len=641	ExpAA=0.00	First60=0.00	PredHel=0	Topology=o
Smp_181360	len=308	ExpAA=1.14	First60=1.14	PredHel=0	Topology=o

Figura 11 - Saída do Tmhmm (Proteínas transmembranas)

O Resultado do Tmhmm, indicando para um grupo de proteínas, quais tem domínios transmembranas (PredHel > 0, número de hélices preditas). O algoritmo do Tmhmm é baseado em HMM (já discutido) identifica domínios transmembranas nas sequências submetidas através de seu modelo de estados muito próximo do modelo biológico.

4.6. Predição das proteínas antigênicas

Após testes em vários algoritmos (Tabela 4), foi escolhida a plataforma FRED para a predição dos epitopos em *S. mansoni*. O FRED é um framework para detecção de epitopos de células T (Classe I e Classe II), que permite o uso e integração de vários preditores. Tal escolha se deu pela possibilidade de usar em uma só ferramenta vários métodos em uma mesma análise, bem como compará-los de modo sistematizado. Desenvolvido na linguagem Python, o FRED é modular, extensível, e provê diversas métricas para comparação de preditores (Figura 12) (Feldhahn, Donnes *et al.*, 2009).

Tabela 4

Algoritmos testados para a predição de epitopos de MHC classe II em *S. mansoni*

Algoritmo	Referência
MHC-BPS	(Cui, Han <i>et al.</i> , 2006)
MHCPred	(Guan, Hattotuwigama <i>et al.</i> , 2006)
Multipred1	(Zhang, Khan <i>et al.</i> , 2005)
NetMHCI	(Nielsen, Lundegaard <i>et al.</i> , 2007)
NetMHCIpan	(Nielsen, Lundegaard <i>et al.</i> , 2008)
ProPred	(Singh e Raghava, 2001)
Rankpep	(Reche, Glutting <i>et al.</i> , 2002)
SVMHC	(Donnes e Kohlbacher, 2006)
SVRMHC	(Wan, Liu <i>et al.</i> , 2006)
SYFPEITHI	(Rammensee, Bachmann <i>et al.</i> , 1999)
FRED	(Feldhahn, Donnes <i>et al.</i> , 2009)

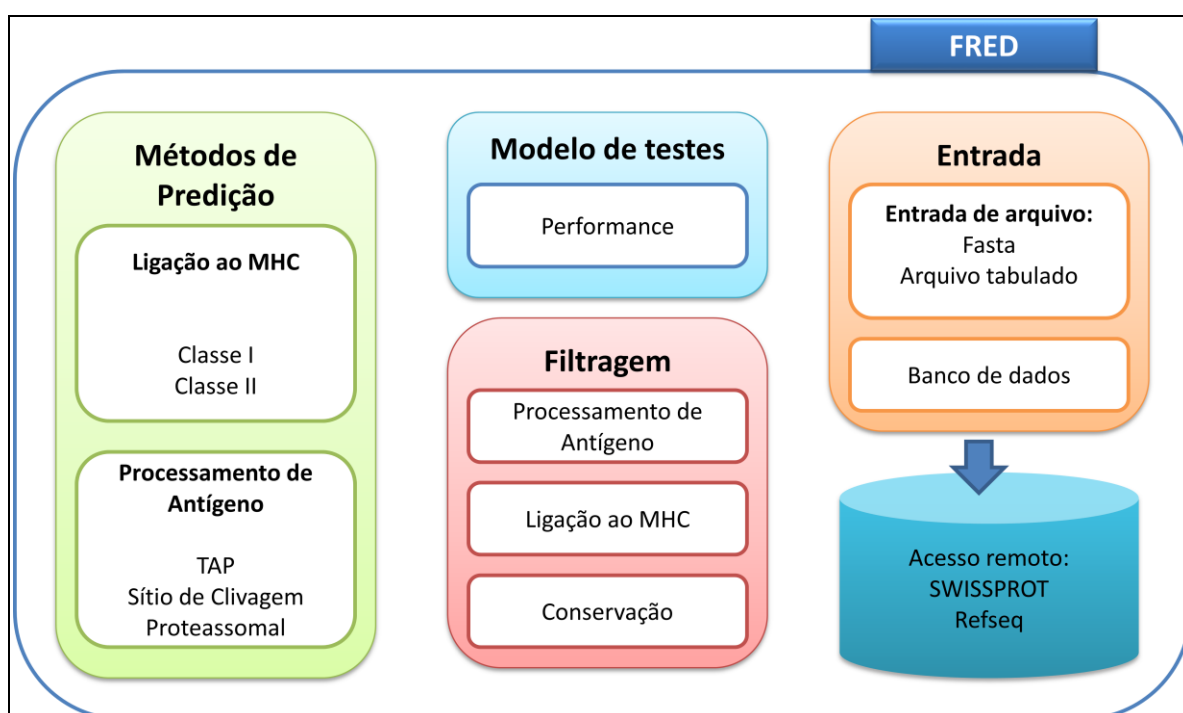


Figura 12 - Arquitetura do FRED

Arquitetura do framework FRED. O framework provê diversos módulos para: carregamento de dados, filtrar resultados e avaliar performance de diferentes preditores
Fonte: (Feldhahn, Donnes *et al.*, 2009).

Os métodos utilizados nesta tarefa foram aqueles que buscam ligantes da via do MHC Classe II (peptídeos de origem extracelular). Os algoritmos integrados para a tarefa foram: Hammer (Hammer, 1995) e SYFPEITHI (Rammensee, Bachmann *et al.*, 1999), que são métodos nativos do framework. Foi realizada a integração de um método externo que também foi usado na predição – NetMHCII (Nielsen e Lund, 2009).

O Hammer é um método baseado em matrizes, as quais foram construídas a partir da combinação de alinhamentos múltiplos de sequências de HLA-DR (MHC Classe II) que tinham a estrutura terciária conhecida, e, um banco de perfis de *pockets* de HLA-DR, que nada mais são que as alças onde os peptídeos se ligam à molécula de HLA. Assim, as matrizes são montadas com base na demonstração, de que os *pockets* que compartilham os mesmos resíduos polimórficos tem perfil de ligação semelhante (Sturniolo, Bono *et al.*, 1999). Os ligantes então preditos por este método têm então seus *scores* computados a partir destas matrizes.

O SYFPEITHI é um banco de *motifs* de peptídeos e ligantes ao MHC, e possui um algoritmo para predição de epitopo também baseado em matrizes. Tais matrizes de dados bidimensionais são construídas a partir dos *motifs* para cada modelo alélico do MHC. Onde as linhas representam os aminoácidos, e as colunas as posições que tal aminoácido pode se ligar ao *pocket* do MHC (Figura 13).

AA	1	2	3	4	5	6	7	8	9
A	0	0	1	0	0	0	0	1	0
C	0	0	0	0	0	0	0	0	0
D	0	0	0	1	0	0	0	0	0
E	1	0	1	1	0	0	0	1	0
F	0	0	0	0	0	0	0	0	6
G	1	0	0	1	1	0	0	0	0
H	0	10	0	0	0	0	0	0	0
I	2	0	0	0	0	1	0	0	0
K	0	0	0	1	0	1	0	0	0
L	0	0	0	0	0	0	0	0	10
M	0	0	0	0	0	1	0	0	6
N	0	0	0	0	1	0	0	0	0
P	0	0	0	2	1	1	1	0	0
Q	0	0	0	1	0	0	0	0	0
R	0	0	0	0	0	1	2	2	0
S	0	0	1	0	0	0	0	0	0
T	1	0	0	0	0	0	0	1	0
V	0	0	0	1	0	1	2	2	0
W	0	0	0	0	0	0	0	0	0
X	0	0	0	0	0	0	0	0	0
Y	1	0	0	0	0	0	0	0	0

Figura 13 – Matriz para predição de ligantes ao modelo alélico HLA-B*1510 do SYFPEITHI.

O cálculo dos *scores* podem ser obtidos diretamente através de um par de índices (aminoácido, posição de ligação no *pocket*). Começando no primeiro aminoácido a sequência é dividida em octâmeros ou nonâmeros e para cada oligômero a soma das pontuações dos aminoácidos é então computada. O processo é repetido até que se chegue ao fim da sequência. Na matriz os aminoácidos que frequentemente se ligam as posições de ancoragem, recebem o valor 10, o valor 8 é dado para o aminoácido presente em um número significativo de ligantes e 6 para resíduos que raramente ocorrem. Aminoácidos auxiliares às posições de ancoragem, também recebem o valor o valor 6, e quando menos frequente, no mesmo conjunto recebem 4. Os valores de 1-4 recebem aqueles aminoácidos que ocorrem em sequências individuais, de acordo com o sinal de sequenciamento. Os aminoácidos que são considerados desfavoráveis para a ligação, tem os valores -1 a -3. Todos estes valores são levados em conta no algoritmo (Rammensee, Bachmann *et al.*, 1999)

O NetMHCII (*NN-Align*) é um algoritmo baseado em redes neurais e foi escolhido para esta análise por apresentar o peptídeo predito bem como a afinidade da ligação. O

método cobre 14 modelos alélicos do MHC II de humano e segundo os autores foi treinado de uma maneira que permite a correção de erros durante o treinamento (Nielsen e Lund, 2009).

Juntos os três métodos fornecem a cobertura de 84 modelos alélicos do MHC II de humana. Não foi possível verificar a frequência destes alelos para a população brasileira, uma vez que não há registros de estudos da frequência dos mesmos no *Allele Frequency Database* (Middleton, Menchaca *et al.*, 2003) disponíveis no Brasil. Para realizar a predição, desenvolvemos um script, em linguagem python, utilizando os métodos do FRED para: carregamento das sequências, divisão dos peptídeos, seleção de métodos e modelos alélicos, filtragem dos resultados e extração dos peptídeos (Anexos – Script 1). Também foram considerados ligantes que em consenso foram preditos por no mínimo 5 modelos alélicos, o que aumenta a confiabilidade de nossa análise, uma vez que na literatura têm se considerado consenso de 3 modelos para projetos em larga escala (Feldhahn, Donnes *et al.*, 2009).

4.7.Criação do banco de perfis de epitopos

Utilizamos o sistema de gerenciamento de bancos de dados MySQL (Mysql, 2010), para a criação de uma base de dados onde foram armazenados os resultados de todas as predições (Figura 14). O banco foi modelado usando a ferramenta SQLyog (Sqllyog, 2010) e povoado com os arquivos de resultados dos preditores usando scripts desenvolvidos em Perl.

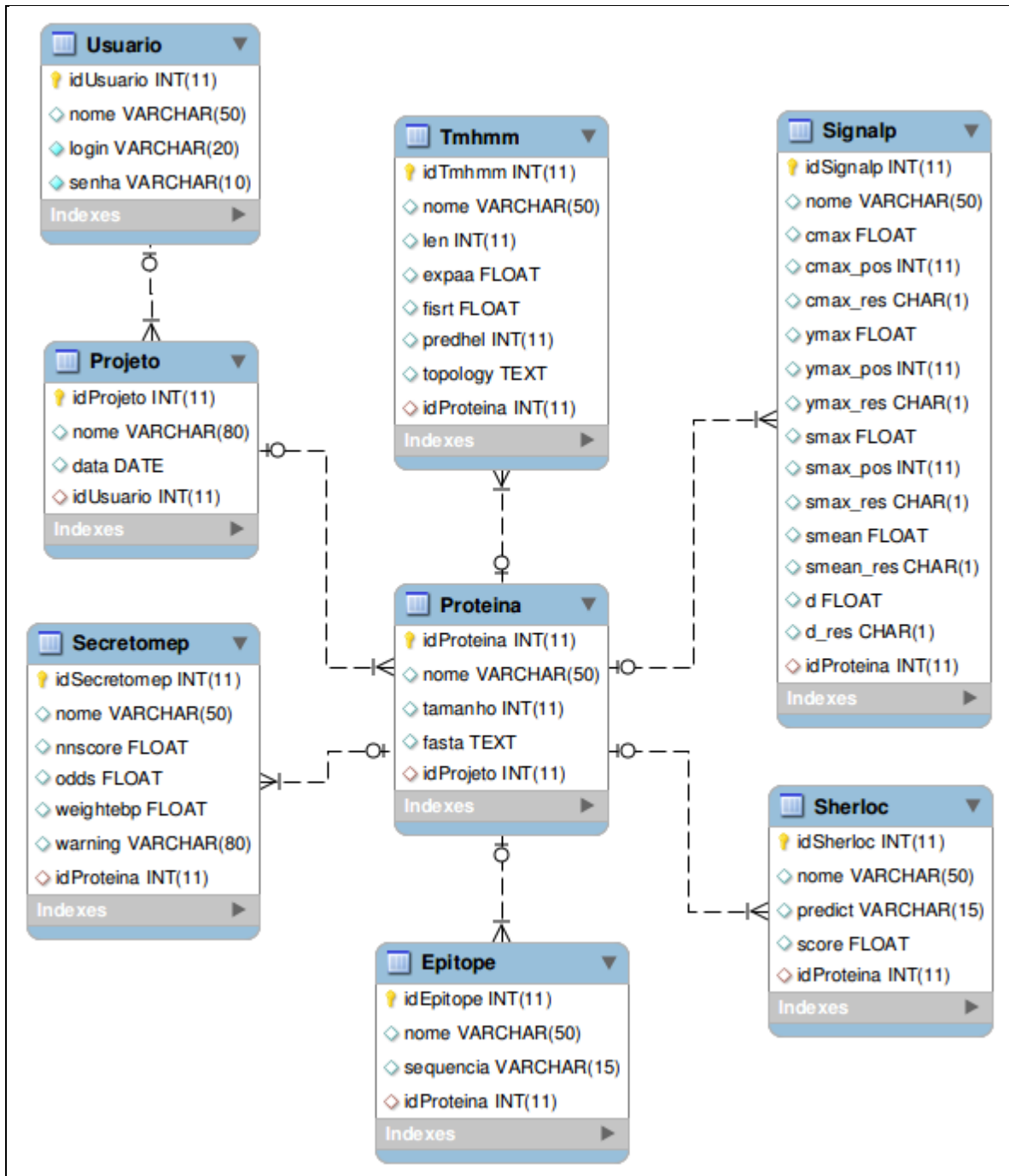


Figura 14 - Esquema do Banco de Dados

Esquema do banco de dados relacional criado para armazenar a saída das predições e para comparações estes resultados.

4.8. Validação experimental da metodologia

Na tentativa de validar experimentalmente nossas análises, um experimento preliminar de proteômica foi conduzido em colaboração com a Dra. Rosiane Pereira, do Laboratório de Parasitologia do Instituto René Rachou.

Extrato protéico de vermes adultos de *S. mansoni* enriquecido de proteínas de membrana foi obtido utilizando o protocolo para extração de proteínas baseado no seu grau de hidrofobicidade, como descrito no kit para fracionamento de amostras para eletroforese bidimensional 2-D Fractionation Kit (GE Healthcare). Resumidamente, proteínas solubilizadas a partir de um pellet de material insolúvel, obtido após a lise mecânica de vermes adultos de *S. mansoni*, foram submetidas à separação por eletroforese bidimensional (2-DE) em fitas e IPG de 7 cm, pH 3-10 (BioRad).

Para isso, 100 µg de proteínas foram solubilizadas em tampão IEF de re-hidratação [8M Ureia, 2M Tiourea, 4% CHAPS, 65mM DTT, 0,0025% Azul de bromofenol e 1% anfólito BioLyte 3-10 buffer (100X) (BioRad)] para um volume final de 125 µl e aplicadas sobre as fitas de IPG. As mesmas foram submetidas à re-hidratação e focalização isoeletrica no equipamento Protean IEF Cell (BioRad) a 50 µA/fita e 20°C. Re-hidratação passiva foi conduzida por 4 hr, seguida de re-hidratação ativa a 50 V por 12 hr. A focalização isoeletrica foi conduzida a 500 V por 30 min, seguida por 1000 V por 30 min, 4000 V por 1 hr e 4000 V até 16.000 V/h.

Logo antes da separação das proteínas pela segunda dimensão, as fitas de IPG foram submetidas à etapa de equilíbrio (redução e alquilação das proteínas). As fitas foram incubadas por 10 min, a temperatura ambiente, sob agitação constante, em tampão de equilíbrio [6M Uréia, 30% glicerol, 2% SDS, 50mM solução de Tris-HCl (pH 8,8) e 0,001% azul de bromofenol] contendo 130mM DTT e, em seguida, por mais 10 min no mesmo tampão contendo 135mM iodoacetamida. Na segunda dimensão, as proteínas foram separadas por peso molecular em SDS-PAGE 12% no sistema Mini-Protean III (BioRad). A eletroforese foi realizada a 50 V por aproximadamente 10 min e a 100 V até o corante atingir a porção inferior do gel.

Para a realização destes experimentos, três 2D-PAGEs (eletroforese em gel de poliacrilamida bidimensional) foram realizados simultaneamente, dois para serem usados nos

experimentos de Western-blotting com pool de soro de indivíduos infectados de área endêmica para esquistossomose e com anticorpo anti-Sm29, gentilmente cedido pelo Dr. Sérgio Costa Oliveira (UFMG), e outro 2D-PAGE para ser corado por Azul de Coomassie Colloidal G-250 para excisão dos spots e identificação das proteínas por espectrometria de massas.

Para realização dos experimentos de Western blotting bidimensional as proteínas previamente separadas por 2-DE foram transferidas para membranas de PVDF. Estas foram bloqueadas por 16 hr em TBS (20mM Tris-HCl, 500mM NaCl, pH 7,5) contendo 0,1% Tween-20 e 3% BSA (TBS-T/BSA-3%). As membranas foram incubadas separadamente com o pool de soro de indivíduos infectados de área endêmica para esquistossomose e com anticorpo anti-Sm29, ambos diluídos 1:500 em TBS-T/BSA-1%. Após serem lavadas, as membranas foram incubadas com o anticorpo secundário, sendo o anticorpo polivalente de cabra anti-Ig's total humana (Invitrogen) diluído 1:100.000 em TBS-T/BSA-1%, e o anticorpo anti-IgG de camundongo (GE Healthcare) diluído 1:50.000 em TBS-T/BSA-1%, respectivamente. Ambos anticorpos secundários são conjugados a HRP. Após as lavagens, as proteínas foram reveladas por quimioluminescência utilizando ECL plus Western-blotting Detection System (GE Healthcare) e as membranas foram expostas a filmes de raio-X.

A próxima etapa foi preparar os spots para identificação das proteínas na Plataforma de Espectrometria de Massas (RPT02H) do Programa de Desenvolvimento Tecnológico em Insumos para Saúde (PDTIS) da Fundação Oswaldo Cruz no Rio de Janeiro (IOC). Os spots imunorreativos ao soro de indivíduos infectados foram localizados no 2D-PAGE corado pelo Azul de Coomassie Colloidal, excisados, descorados e submetidos a digestão in gel por tripsina. Os peptídeos trípticos foram extraídos e purificados em microcolunas de fase reversa Zip Tip C18 (Eppendorf) de acordo com as instruções do fabricante. A identificação das proteínas por espectrometria de massas do tipo MALDI-ToF-ToF (Matrix Assisted Laser Desorption Ionization Time-of-Flight mass spectrometry) foi realizada em um espectrômetro de massas modelo 4700 Proteomics Analyzer (Applied Biosystems), a qual é baseada na ionização dos peptídeos por desorção a laser auxiliada por matriz (MALDI) e na análise do tempo de voo in tandem dos íons (ToF-ToF).

O espectro de massas dos peptídeos gerados pela digestão com tripsina (MS) e os dados de fragmentação dos peptídeos (MS/MS) foram analisados utilizando o programa MASCOT (<http://www.matrixscience.com/>) e sua busca em banco de dados MS/MS íon

search. Como usamos o programa online, escolhemos o banco de dados NCBIInr (banco de proteínas não redundante do NCBI), e aplicamos o filtro de taxonomia para metazoários. Os parâmetros usados foram: tolerância de peptídeo MS de ± 0.6 Da, tolerância de fragmento de massa MS/MS de ± 0.2 Da, duas falhas de clivagem permitidas, carga do peptídeo +1 e modificações variáveis em cisteína (carbamidometilação e adição de propionamida) e em metionina (oxidação). A fim de evitar identificações aleatórias, apenas íons com valor de score individual acima do indicado pelo MASCOT para identidade e homologia extensiva ($p < 0,05$) foram aceitos.

4.9.Implementação do pipeline de busca automática de novos alvos.

Para conclusão do projeto, trabalhamos no desenvolvimento de um pipeline computacional que integra todas as etapas aqui já descritas (Figura 15). Este produto permite a execução e re-execução de cada uma das predições de forma automatizada e simplificada, e, também poderá ser facilmente gerenciado para atender possíveis necessidades na busca de candidatos à vacinas para outros organismos.

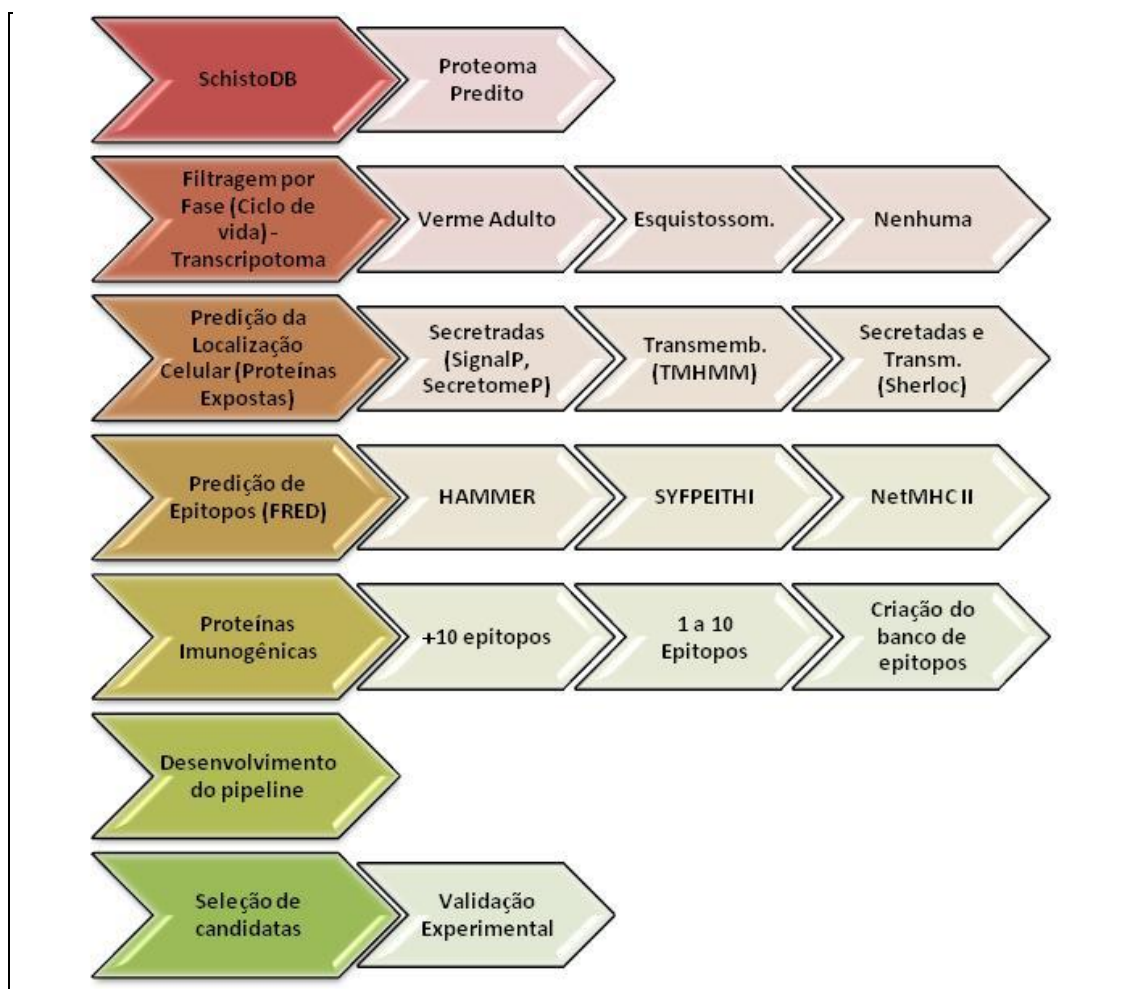


Figura 15 - Etapas do projeto

Metodologia do projeto, destacando todas as etapas e algoritmos usados em cada uma delas.

Inicialmente a ferramenta foi criada na plataforma Galaxy (Galaxy, 2010). O Galaxy permite criar interfaces através de uma página web para a execução de programas em fluxo. Uma vez instalados os programas no servidor, é feita a configuração do pipeline através de arquivos XML. Alguns scripts foram escritos para manipulação de entrada e saída de programas, para evitar problemas de compatibilidade durante o desenvolvimento do pipeline

O pipeline para a plataforma Galaxy tem seu funcionamento incompleto, pois ainda não estão integrados todos os algoritmos. No entanto já tem sido usada por outros membros do grupo que necessitam da rápida triagem de proteomas na identificação de proteínas expostas (Figura 16).

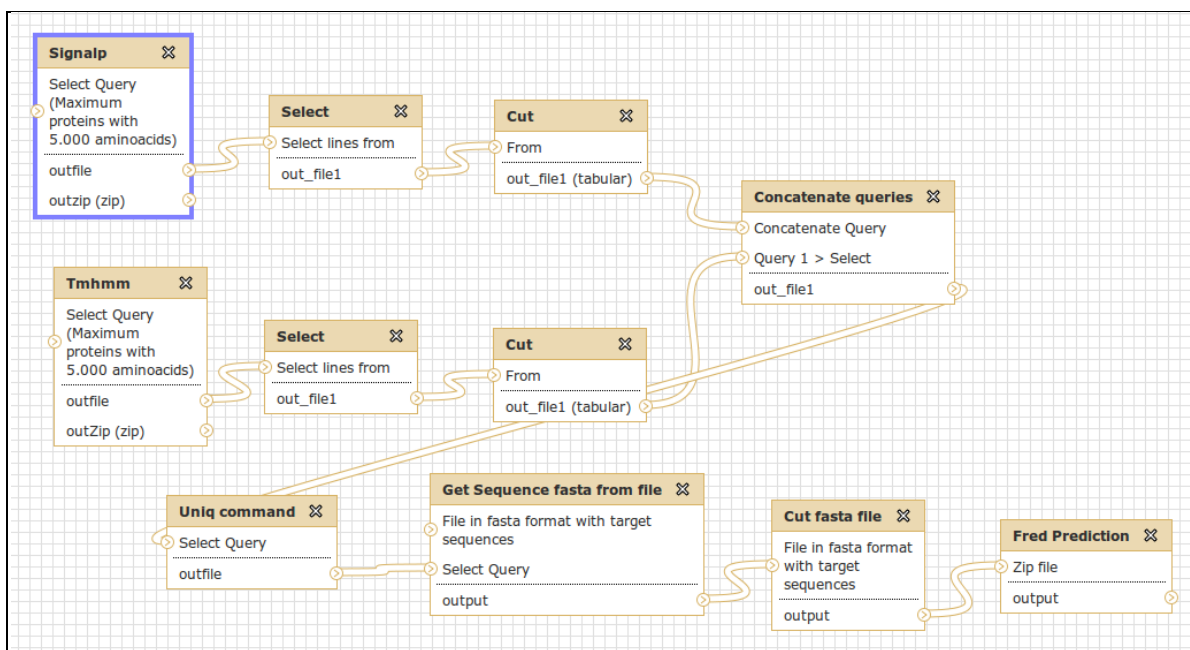


Figura 16 - Pipeline plataforma Galaxy

Snapshot do pipeline com predições para busca de candidatos a vacina.

Para disponibilização do pipeline de forma mais completa e mais funcional, associado ao banco de perfis de epitopos criado, desenvolvemos um *webserver* (EpiFinder – <http://epifinder.cebio.org>), utilizando a metodologia de Padrões de Projeto de Software em Java. Tal metodologia se baseia no uso arquiteturas testadas para construir softwares orientados a objetos flexíveis e sustentáveis. Utilizar padrões de projeto reduz substancialmente a complexidade do processo de *design* de um software (Deitel, 2005).

A arquitetura escolhida para o desenvolvimento do servidor web foi a *Spring Framework*, o qual é um *framework open source* para a plataforma Java criado por Rod Johnson e descrito em seu livro "*Expert One-on-One: JEE Design e Development*" (Johnson, 2003). Trata-se de um framework centrado nos conceitos de desenvolvimento leve e ágil. O *Spring* permite diversas facilidades para desenvolvimento e, sobretudo a integração com diversas soluções do mundo Java como, por exemplo, JPA/Hibernate e JSF (Java Server Faces), também utilizados no projeto do servidor (Gomes, 2008).

A JPA (Java Persistence API) é uma recém criada especificação para permitir persistência objeto-relacional com bancos de dados relacionais (Biswas, 2006). O termo persistência de dados é bastante comum em computação e se refere à ação de armazenamento não-volátil de dados, ou simplesmente guardar uma informação em um dispositivo físico ou virtual. Ao invés de salvar dados em tabelas, o código do sistema que utiliza esta especificação solicita a tarefa de salvar os valores desejados, assim, as tecnologias que utilizam JPA, como o Hibernate (<http://www.hibernate.org/>), transformam automaticamente as requisições de salvar e consultar via classes do sistema, em comandos SQL que são enviados ao banco de dados, sem a necessidade que o desenvolvedor escreva tais comandos. O JSF é a tecnologia padrão J2EE para criar aplicações web. Este padrão foi criado a partir das tecnologias JSP e *Servlets* e estende seus conceitos com um conjunto de componentes e recursos mais sofisticados e focados no desenvolvimento RAD para Web (Gomes, 2008). O ambiente de desenvolvimento utilizado foi o Netbeans (Netbeans, 2011), ambiente completo e extensível, e, compatível com todas as tecnologias citadas. A Figura 17 apresenta um esquema das tecnologias envolvidas para a construção do servidor.

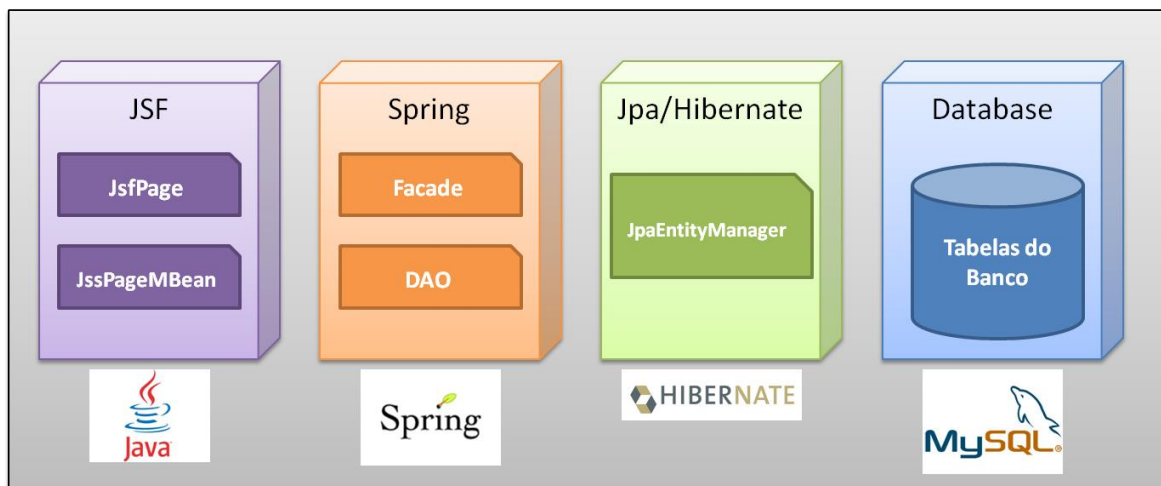


Figura 17 - EpiFinder tecnologias empregadas

O EpiFinder foi construindo aplicando as mais novas tecnologias para desenvolvimento corporativo e de padrões de projeto.

O Diagrama de Classes da Arquitetura, com as tecnologias empregadas pode ser visto na Figura 18, e segue uma breve descrição dos componentes representados na figura:

A *JSPPage* (Página JSF), é criada visualmente utilizando o *Visual Web Pack* do Netbeans. Trata-se de uma paleta de componentes prontos, desenvolvidos utilizando JSF 1.2 da Sun. São dezenas de componentes visuais e não visuais para o desenvolvimento RAD de aplicações Web com Java, JSP e JSF (Gomes, 2008).

O *JSPPageMBean*, para cada página JSF, deve haver um *Managed Bean (MBean)* que possui o código Java de servidor que cuida de responder as requisições do usuário. É a lógica de aplicação da pagina web que o usuário está utilizando. Ao clicar em um botão Ok, por exemplo, e assim submeter valores ao *MBean* serão executados os procedimentos necessários e devolvidos resultados desejados ao usuário da aplicação (Gomes, 2008).

O *Facade* é um Padrão de Projeto definido pelo no livro *Design Patterns: Elements of Reusable Object-Oriented Software* (Gamma, 1994) como uma boa prática para implementar, encapsular e reutilizar código de negócio. Ele é o ponto de chamada para utilizar as funcionalidades de negócio da aplicação. Por exemplo se na aplicação houver uma funcionalidade que calcula o peso molecular de um epitopo, haverá um método em um *facade* apropriado para implementar este cálculo. A vantagem de se utilizar *facades* é encapsular e

isolar o código de negócios das classes que cuidam da interação com o usuário ou com outros sistemas. Facilitando a reutilização de uma mesma funcionalidade de negócio por duas ou mais telas, ou por um ou mais sistemas ou componentes deste (Gomes, 2008).

O DAO (Data Access Object), é uma boa prática de desenvolvimento catalogada no *Sun Blueprints Design Patterns* (<http://java.sun.com/blueprints/patterns/>) como uma forma para implementar código de acesso a recursos de banco de dados que fornecem e consomem informação. Ele é ponto de chamada para utilizar recursos do banco de dados, tanto para consultas (*select*), quanto para manipulação de dados (*insert*, *update*, *delete*). A vantagem de se utilizar DAOs, é permitir a separação inteligente de código de negócio do código de manipulação de banco dados. Isto possibilita ter independência de banco de dados tal forma que a mudança do mesmo não gera manutenção das classes de negócio. Ou seja, o negócio não deve se preocupar com tecnologia de armazenamento de dados e detalhes técnicos, apenas negócios (Gomes, 2008).

A *JPAEntityManager* (Tabela do banco de dados), é o gerenciador de persistência de banco de dados definido na especificação JPA. É um componente que permite acesso objeto-relacional aos recursos do banco. Ao invés de salvar dados de uma classe, criando um comando “*inser into*”, é solicitado ao *JpaEntityManager* que grave a classe em banco de dados diretamente, sem que haja necessidades de escrever código SQL para inserir na tabela de banco de dados (Gomes, 2008).

O POJO, trata-se de um objeto Java utilizado para implementar o modelo de domínio da aplicação (Gomes, 2008).

Assim, o servidor foi desenvolvido utilizando navegação em abas *master-detail*, permitindo realizar e combinar todas as predições aqui indicadas em duas análises possíveis: completa e simples (Figura 19).

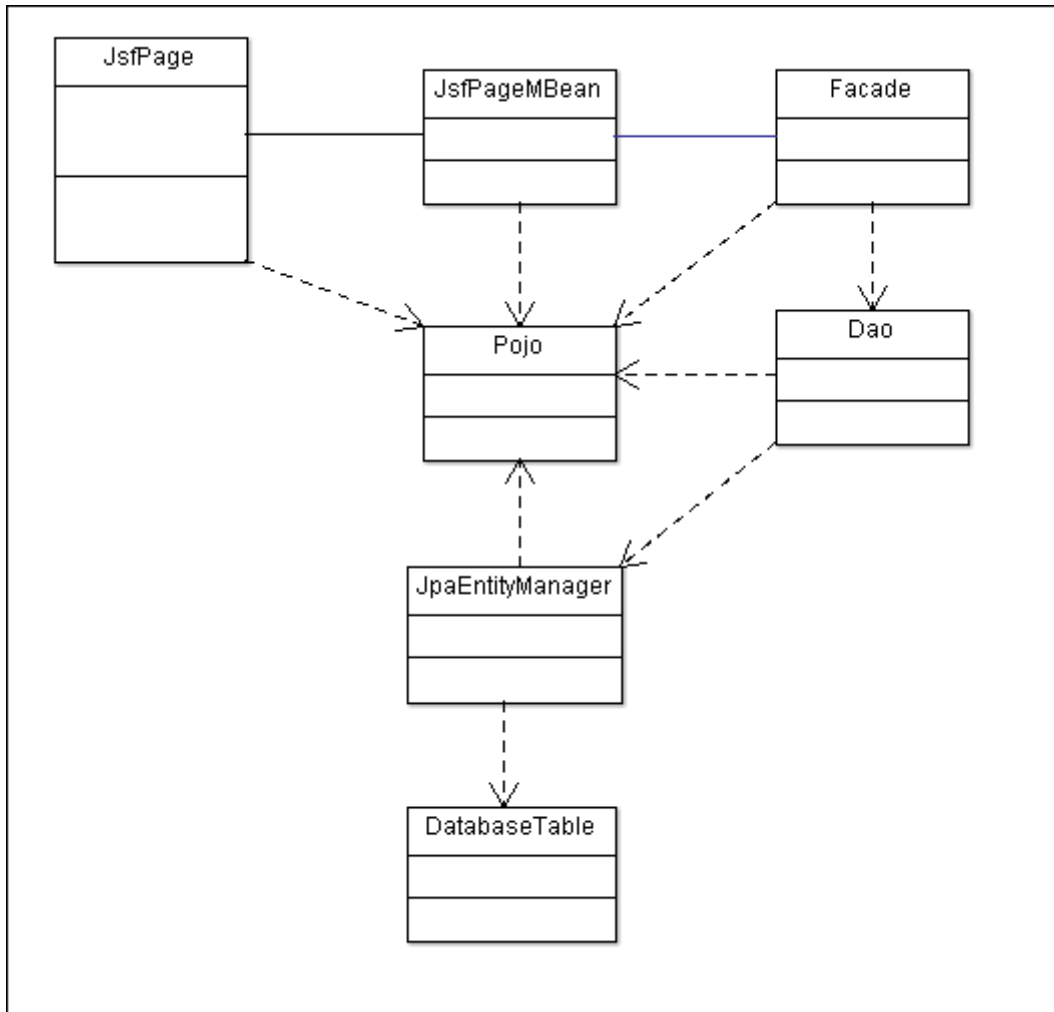


Figura 18 - Diagrama de Classes da Arquitetura

O diagrama de classes da arquitetura Spring e tecnologias integradas (JSF e JPA).

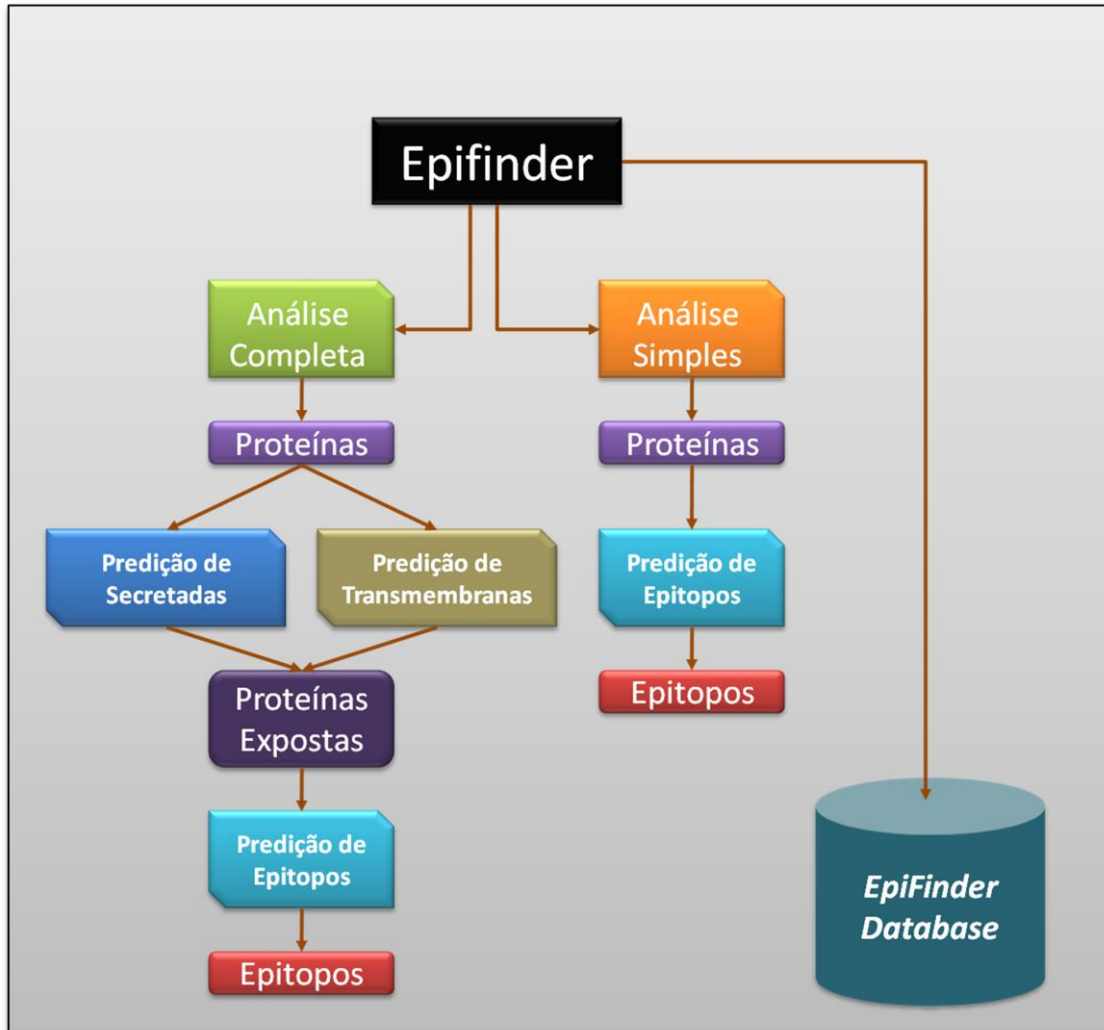


Figura 19 - EpiFinder Webserver

Análises disponíveis no servidor. Completa e Simples. Na análise completa é possível realizar inicialmente a busca de proteínas expostas (secretadas e de membrana), em seguida o mapeamento dos epitopos. Na simples, é possível realizar o mapeamento de epitopos diretamente em um grupo de proteínas.

5. RESULTADOS E DISCUSSÃO

5.1. Predições e seleção dos candidatos a validação experimental.

O foco principal do projeto foram as análises computacionais para obtenção de possíveis candidatos à vacina contra esquistossomose. Ao longo do trabalho alguns resultados merecem ser destacados.

A Tabela 5 apresenta um resumo do resultado das predições integradas (em consenso), para a identificação das proteínas expostas ao hospedeiro, sejam elas secretadas ou associadas à regiões de transmembrana, por abordagens distintas, bem como destaca a quantidade proteínas potencialmente imunogênicas, e a média de epitopos em cada uma, obtidas através da plataforma FRED.

Tabela 5

Resumo das predição por localização e por epitopos. Podemos observar que que todas as fases apresentam perfil semelhantes nas predições, embora o grupo com Nenhuma evidência de expressão apresentem maior queda tanto no número de proteínas com epitopos como média de epitopos por proteínas.

Grupo	Total	Secretadas (S)	Transmemb. (T)	S+T (expostas) ^①	Proteínas com epitopos	Média epitopos ^②
Verme Adulto	7093	703	1235	1531	1245	3,93
Esquistossômulo	6032	562	952	1205	982	4,00
Nenhuma evi	3307	338	541	702	450	2,84
VA-ES-NE^③	11601	1151	1964	2469*	1883	3,64

* Foram agrupadas para cada fase as suas respectivas proteínas transmembranas e secretadas, então removidas as devidas redundâncias.

① Este número representa o total de proteínas potencialmente expostas no parasito.

② Média de epitopo por proteína.

③ VA (Verme Adulto), ES (Esquistossômulo) e NE (Nenhuma Evidência)

Podemos observar na Figura 20 uma distribuição de epitopos por proteína, que podemos concluir que o número de epitopos em uma proteína está bastante ligado a seu tamanho, quanto maior a proteína mais chances de apresentar um maior número de epitopos. No entanto também temos exemplos que contrariam esse comportamento: a Smp_126600, que possui 2016 aminoácidos, tem mapeado em sua sequência 4 epitopos, enquanto a Smp_131740, que possui 137 aa foram preditos 10 epitopos (peptídeos 9mer).

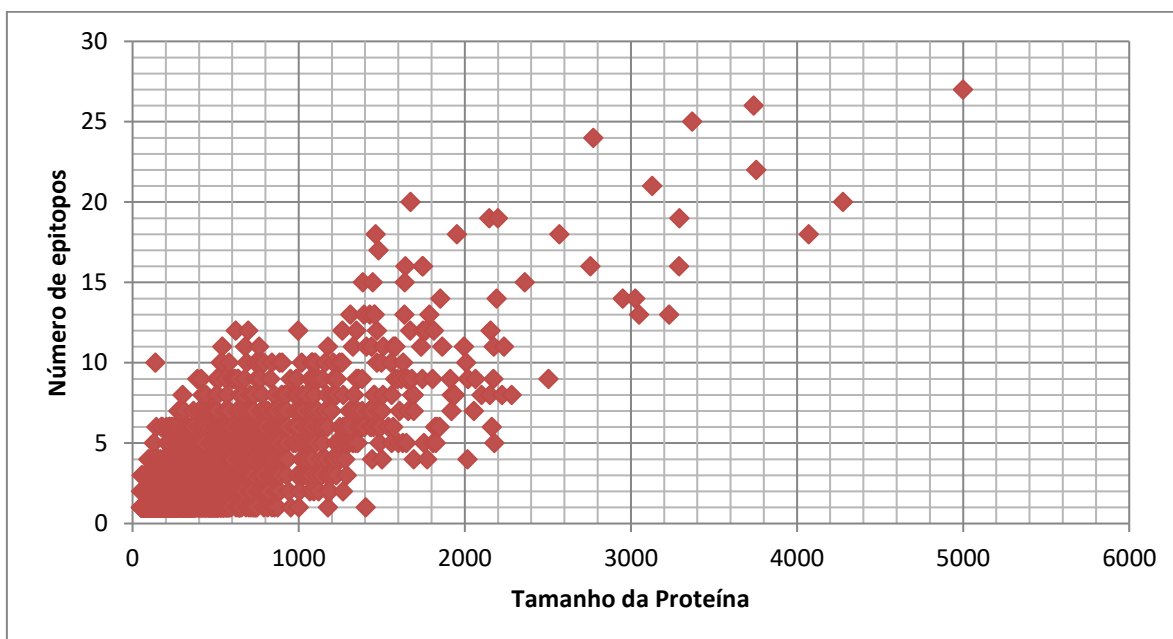


Figura 20 – Distribuição de Epitopos por Proteínas

A quantidade de epitopos preditos em uma proteína está relacionada ao seu tamanho.

Para a seleção dos possíveis candidatos à validação experimental, foram identificadas em cada grupo as proteínas aquelas que se ligam a um maior grupo de alelos e que apresentam um maior número de epitopos, por entendermos que estas têm um maior potencial imunogênico. E também por recomendação dos autores do *framework* de predição de ligantes, que fazem uma investigação semelhante no proteoma humano, para estudos de câncer (dados ainda não publicados). A Tabela 6 apresenta o resultado desta análise.

Tabela 6**Seleção de proteínas para validação experimental.**

Grupo	S+T (expostas)	Proteínas com epitopos	Média de epitopo	Potenciais Alvos
Verme Adulto	1531	1245	3,93	72
Esquist.	1205	982	4,00	62
Nenhuma evid.	702	450	2,84	10
VA-ES-NE	2469	1883	3,64	90*

* Este número o total de potenciais alvos, proteínas com o maior número de ligantes ao MHC Classe II

5.2. Bancos de perfis de epitopos e Servidor Web para realização das predições

Também criamos um banco de dados relacional que armazena o resultado de cada uma das análises e nos permitiu facilmente integrar e comparar as predições através de consultas SQL. Tal banco de dados foi modelado e em seguida desenvolvido no sistema de gerenciamento de banco de dados MySQL.

Além disso, foram criados scripts para leitura da saída de cada um dos algoritmos usados, e também para carregamento desta informação na base relacional. Ainda foi possível a criação de um pipeline para a Plataforma Galaxy, que já tem sido usado por nosso grupo, afim de rapidamente classificar um proteoma.

Como um dos principais produtos do trabalho, apresentamos um servidor web que permite realizar a abordagem aqui proposta, de forma completa e também de forma simples. Existem vários algoritmos e métodos que se propõem a realizar a abordagem aqui descrita, porem funcionam de forma isolada e apresentam resultados nos mais diversos formatos, não havendo uma integração de metodologias diretamente, deixando a cargo do usuário tal integração. No presente projeto estabelecemos um servidor, gratuito, onde é possível combinar as mais interessantes predições para identificação de candidatos à vacinas para organismo extracelulares.

Cada usuário cadastrado no sistema tem acesso banco de dados modelado para tais análises, o qual é possível gerenciar diversos projetos e proteínas associadas a estes (Figura 21). Realizar e armazenar o resultado para todas as predições (Figura 22) e no final obter o mapeamento dos epitopos associados àquela proteína imunogênica de interesse (Figura 23).

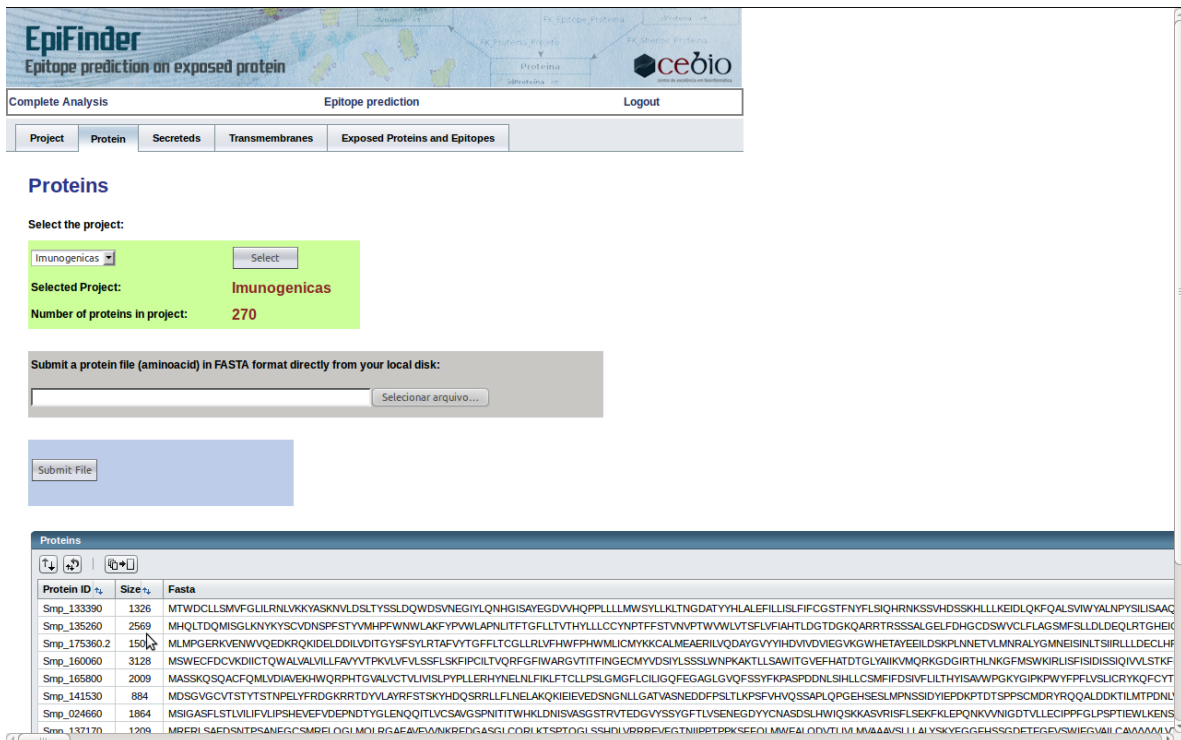


Figura 21 - EpiFinder – Home Page das Proteínas.

No servidor é possível gerenciar diversos projetos de predição, e submeter conjuntos de seqüências de proteínas para cada projeto.

EpiFinder
Epitope prediction on exposed protein

Complete Analysis Epitope prediction Logout

Project Protein Secreteds Transmembranes **Exposed Proteins and Epitopes**

Epitopes

Select the project:
 Teste [Select]
 Selected Project: **Teste**
 Number of exposed proteins in project: **90**

Get Epitopes:
 Run Analysis!

Exposed proteins:

Name	Size
Smp_133390	1326
Smp_135260	2569
Smp_175360.2	1509
Smp_160060	3128
Smp_165800	2009
Smp_141530	884
Smp_024660	1864
Smp_137170	1209
Smp_147250.1	1385
Smp_163820	1311

Page: 1 of 9

Minimum number of Epitopes: [input] Show

Tabela

nome	tamanho
Smp_133390	1326
Smp_135260	2569
Smp_175360.2	1509

Figura 22 - EpiFinder – Resultado da predição de epitopos em proteínas expostas

Detalhe da aba onde é realizada a predição de epitopos, é possível seleccionar um grupo de proteínas informando o número mínimo de epitopos que deseja.

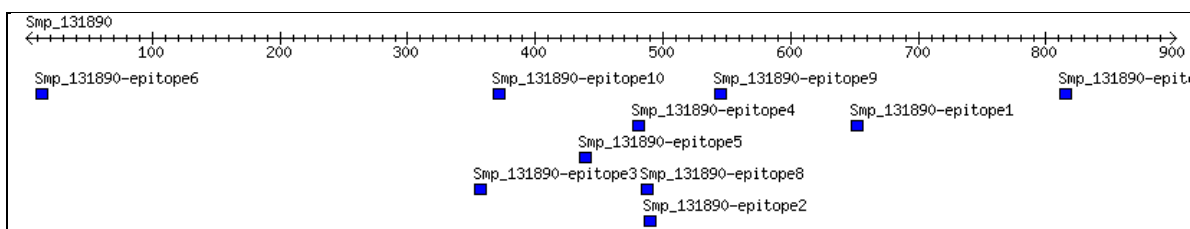


Figura 23 - EpiFinder – Mapeamento de epitopos para uma data proteína.

O servidor mapeia todos os epitopos de uma proteínas que resulta da análise.

5.3. Avaliações experimentais

A Tabela 7 apresenta as 90 proteínas resultantes das predições e que irão ser submetidas à análise experimental tanto pra confirmação da metodologia, como para tentar apontar novas possibilidades compor a lista de alvos já conhecida pela comunidade científica. Destas, 50 foram anotadas manualmente observando alguns critérios para restrições à melhores dados para o experimento de expressão destas proteínas (Anexos – Tabela 2), como por exemplo: eliminação de redundância, maior número de lisinas, menor peso molecular, maior número de loops, e epitopos mapeados nestes loops (Figura 24), no caso de uma proteína transmembrna. As proteínas serão então enviadas para Universidade de Nottingham onde ocorrerá a expressão dos antígenos em array. Os *primers* já foram desenhados para todas e serão preparadas para expressão, maiores detalhes deste experimento serão discutidos na seção 7.

Tabela 7

Lista de possíveis candidatas, obtidas através da integração de predições para busca de novos candidatos à vacina.

ID Proteína	Tam (aa)	Peso Molecular (Kda)	Lisinas (Nr.)	Peptído Signal	Sítio Clivagem (Pos)	Domínios Transmb. (Nr.)	Loops Externos	Epitopos (Nr.)	Produto	Estag. Desenvolvi
Smp_009990	1736	195,61112	82	Y	26	0	0	11	Tyrosine Kinase Receptor_TK Group_InsR Family	Adult worm,Schistosomula
Smp_020170	2199	253,90696	114	N	-	19	9	19	high voltage-activated calcium channel Cav2A	NE
Smp_020270	1582	180,73561	88	N	-	18	9	11	high voltage-activated calcium channel Cav1	Adult worm,Schistosomula
Smp_024660	1864	205,63045	72	Y	24	2	1	11	cell adhesion molecule, putative	Schistosomula
Smp_028690	761	84,15856	29	N	-	12	6	11	sodium/chloride dependent transporter, putative	NE
Smp_034080	3754	430,14066	236	Y	18	0	0	22	erythrocyte membrane protein, putative	Adult worm
Smp_063000	687	75,42125	31	N	-	3	1	10	smdr1	Adult worm
Smp_065110	790	90,04662	67	N	-	1	0	10	leucine zipper-ef-hand containing transmembrane protein, putative	Adult worm,Schistosomula
Smp_076950.2	696	77,55708	31	N	-	11	5	12	solute carrier family 33 (acetyl-CoA transporter, putative	Adult worm,Schistosomula
Smp_123950	838	95,92882	35	N	-	11	5	10	expressed protein	Adult worm
Smp_123960	4276	491,9147	282	N	-	1	1	20	expressed protein	Adult worm,Schistosomula
Smp_124200	1576	180,60121	73	Y	27	1	0	11	scavenger receptor-related	NE
Smp_125510	1498	170,15665	65	Y	27	2	1	10	cadherin, putative	NE
Smp_126160	1077	121,92989	42	N	-	10	5	10	poly(p)/ATP NAD kinase, putative	Adult worm,Schistosomula
Smp_126350	1260	142,39901	76	Y	19	7	4	10	glutamate receptor, ionotropic, n-methyl d- aspartate, putative	Adult worm,Schistosomula
Smp_127000.1	1455	165,95968	85	N	-	14	7	13	cationic amino acid transporter, putative	Adult worm,Schistosomula
Smp_127000.2	1456	166,05882	85	N	-	14	7	13	cationic amino acid transporter, putative	Adult worm,Schistosomula
Smp_127680	4070	450,7932	172	N	-	2	1	18	rabconnectin-related	Adult

Smp_129390	1852	209,84581	84	Y	20	0	0	14	vacuolar protein sorting-associated protein (vps13b), putative	worm,Schistosomula Adult worm,Schistosomula
Smp_129820	1428	161,29737	53	Y	34	12	6	13	multidrug resistance protein 1 (ATP-binding cassette C1), putative	Adult worm
Smp_131740	137	16,32337	1	Y	33	4	2	10	conserved hypothetical protein	NE
Smp_131890	901	100,87094	46	N	-	12	6	10	sodium/chloride dependent transporter, putative	Adult worm,Schistosomula
Smp_133130	2146	247,97636	112	N	-	19	9	19	high voltage-activated calcium channel Cav2A, putative	NE
Smp_133180.2	1786	204,38188	96	N	-	9	5	13	zinc finger protein, putative	Adult worm,Schistosomula
Smp_133190	1748	200,05931	89	N	-	21	10	12	voltage-gated cation channel, putative	Adult worm,Schistosomula
Smp_133390	1326	147,67429	42	Y	25	8	4	11	restin-like	Adult worm,Schistosomula
Smp_133700	1673	189,18484	93	Y	21	0	0	20	udp-glucose glycoprotein:glucosyltransferase, putative	Adult worm
Smp_134180	1569	175,89314	56	N	-	1	0	11	receptor tyrosine phosphatase type r2a, putative	Adult worm,Schistosomula
Smp_135260	2569	289,61389	87	N	-	2	1	18	ethanolaminephosphotransferase, putative	Adult worm,Schistosomula
Smp_135610	3738	410,78851	124	N	-	2	1	26	expressed protein	Adult worm,Schistosomula
Smp_136030	763	85,74045	35	N	-	7	4	10	anion exchange protein, putative	NE
Smp_136310	539	61,15924	25	N	-	8	4	11	sodium-bile acid cotransporter related	NE
Smp_136560	1995	227,8244	87	N	-	15	7	11	expressed protein	Adult worm,Schistosomula
Smp_136960	1480	167,72416	78	N	-	1	1	17	sideroflexin 1,2,3, putative	Adult worm,Schistosomula
Smp_137080	1263	141,22165	60	N	-	11	5	12	multidrug resistance protein 1, 2, 3 (p glycoprotein 1, 2, 3), putative	Adult worm,Schistosomula
Smp_137170	1209	133,56858	59	N	-	8	4	10	plasma membrane calcium-	Adult

Smp_137200	2236	256,05964	126	N	-	1	1	11	transporting atpase, putative expressed protein	worm,Schistosomula Adult worm
Smp_137710	540	60,69695	13	N	-	10	5	10	drug transporter, putative	Adult worm,Schistosomula
Smp_138620	3049	349,58406	204	Y	17	0	0	13	expressed protein	Adult worm,Schistosomula
Smp_139020	1471	165,67901	71	N	-	9	5	10	Hybrid Protein Kinase_Other Group_TBCK Family	Adult worm,Schistosomula
Smp_139320	1159	126,61933	35	N	-	1	1	10	expressed protein	Schistosomula
Smp_139540.1	1245	141,34684	62	N	-	1	1	10	expressed protein	Adult worm
Smp_141530	884	101,73806	44	N	-	7	3	10	expressed protein	Adult worm,Schistosomula
Smp_141710	1463	168,05073	77	Y	23	0	0	18	Hypothetical protein, putative	Adult worm
Smp_142550	2192	244,35615	70	Y	16	0	0	14	nuclear pore membrane glycoprotein gp210- related	Adult worm,Schistosomula
Smp_143730	1393	156,9267	73	Y	38	1	0	13	carboxypeptidase regulatory region-containing, putative	Adult worm,Schistosomula
Smp_145420	2773	315,21427	140	Y	27	2	1	24	plexin, putative	Adult worm,Schistosomula
Smp_146170	3291	372,34125	145	N	-	8	4	19	DNA polymerase epsilon, catalytic subunit, putative	Adult worm,Schistosomula
Smp_147250.1	1385	152,85368	47	N	-	11	6	15	multidrug resistance-associated protein, putative	Adult worm
Smp_147250.2	1445	159,55142	49	N	-	11	6	15	multidrug resistance-associated protein, putative	Adult worm
Smp_147610	741	82,74779	26	N	-	7	3	10	conserved hypothetical protein	Adult worm
Smp_148650	3026	339,77524	125	Y	21	0	0	14	expressed protein	Adult worm,Schistosomula
Smp_148740	1348	155,43553	71	N	-	4	2	12	expressed protein	NE
Smp_149390	3229	372,40191	160	Y	24	0	0	13	fras1 related extracellular matrix protein	Adult worm,Schistosomula
Smp_150770	2950	334,20021	119	N	-	1	0	14	teneurin and n- acetylglucosamine-1- phosphodiester alpha-n- acetylglucosaminidase, putative	Adult worm,Schistosomula
Smp_152020	1018	115,55589	42	Y	22	10	5	10	expressed protein	Adult worm,Schistosomula
Smp_152680	1766	201,04338	114	N	-	2	1	12	Tyrosine Kinase Receptor_TK	NE

Smp_153480	996	111,66296	27	N	-	11	5	12	Group_EGFR Family cation chloride cotransporter, putative	Adult worm
Smp_153500	1556	177,03938	71	N	-	2	1	10	Tyrosine Kinase Receptor _TK Group_VKR Family	Schistosomula
Smp_153590	1081	121,47715	49	N	-	1	0	10	beta-tubulin cofactor d, putative	Adult worm,Schistosomula
Smp_154760	1814	202,63235	63	Y	27	0	0	12	egf-like domain protein	Adult worm,Schistosomula
Smp_154790	1748	198,88099	96	N	-	3	2	16	homeobox protein, putative	NE
Smp_156040	1671	192,15965	99	N	-	9	5	12	sugar transporter, putative	Adult worm,Schistosomula
Smp_157090	1471	168,15963	108	Y	17	0	0	12	cathepsin l, putative	Adult worm,Schistosomula
Smp_157490	1083	124,15844	61	N	-	6	3	10	voltage-gated potassium channel, putative	Adult worm,Schistosomula
Smp_159780	2755	305,68979	125	N	-	4	2	16	expressed protein	Adult worm
Smp_160060	3128	357,51881	150	Y	34	1	0	21	expressed protein	Adult worm,Schistosomula
Smp_163160	1098	126,36366	62	N	-	6	3	10	transient receptor potential channel, putative	NE
Smp_163570	4998	559,03865	236	N	-	6	3	27	ryanodine receptor related	Adult worm,Schistosomula
Smp_163820	1311	148,51328	67	N	-	8	4	13	phospholipid-transporting atpase-related ((aminophospholipid flippase))	Adult worm
Smp_164130	2360	265,39254	97	N	-	2	1	15	dopey-related	Adult worm,Schistosomula
Smp_164270	1628	183,21604	52	N	-	1	0	10	cell adhesion molecule, putative	Adult worm
Smp_164540	1176	131,60947	62	N	-	8	4	11	cation-transporting atpase 13a1 (G-box binding protein), putative	Adult worm
Smp_165230	3369	379,08481	148	N	-	13	7	25	polycystin 1-related	Adult worm
Smp_165800	2009	227,93042	91	N	-	11	6	10	ATP-binding cassette transporter, putative	Adult worm,Schistosomula
Smp_167690	3289	364,98586	83	Y	24	0	0	16	ubiquitin-protein ligase, putative	Adult worm,Schistosomula
Smp_168980	677	74,88575	19	N	-	12	6	11	Hypothetical protein, putative	NE
Smp_169750	2174	248,04958	117	Y	23	0	0	11	dicer-1, putative	Adult

										worm,Schistosomula
Smp_169770	620	70,87995	35	N	-	5	3	12	conserved hypothetical protein	Adult worm
Smp_171740	1745	197,26743	78	N	-	15	7	16	multidrug resistance protein 1 (ATP-binding cassette C1), putative	Adult worm
Smp_171820.1	1638	183,83268	84	N	-	1	0	15	neuropathy target esterase/swiss cheese -related protein	Adult worm,Schistosomula
Smp_171820.2	1642	184,55677	85	N	-	1	0	16	neuropathy target esterase/swiss cheese -related protein	Adult worm,Schistosomula
Smp_172040	2154	244,66581	119	Y	24	0	0	12	unc-13 (munc13), putative	Adult worm,Schistosomula
Smp_175360.1	1437	162,15217	58	N	-	11	5	11	cation-transporting atpase worm, putative	Adult worm,Schistosomula
Smp_175360.2	1509	170,44476	61	N	-	11	5	11	cation-transporting atpase worm, putative	Adult worm,Schistosomula
Smp_175510	1953	220,14293	114	N	-	2	1	18	hypothetical protein	NE
Smp_176930	581	64,49285	24	N	-	11	6	10	cationic amino acid transporter, putative	NE
Smp_177040	531	60,97008	16	N	-	10	5	10	expressed protein	Adult worm,Schistosomula
Smp_177130	1637	187,1236	60	N	-	2	1	13	dep domain containing protein, putative	Adult worm,Schistosomula
Smp_194490	1406	157,69549	72	N	-	12	6	11	vesicular amine transporter, putative	Adult worm

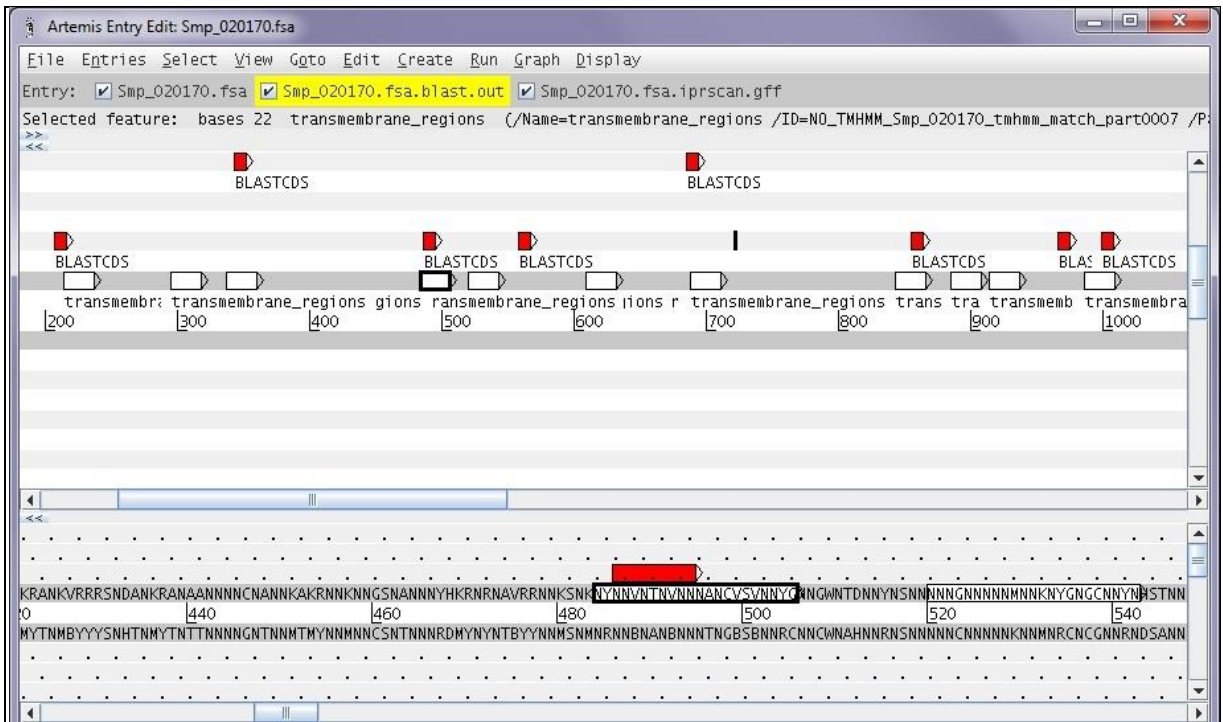


Figura 24 – Mapeamento de epitopos em uma proteína

Um dos critérios utilizados para escolha para validação experimental, são aquelas proteínas que apresentam maior número de epitopos (caixas vermelhas) na região dos loops de membrana (caixas brancas).

5.3.1. Resultados de proteômica

Extrato protéico de vermes adultos de *S. mansoni* enriquecido de proteínas de membrana foi separado por eletroforese bidimensional utilizando fitas de IPG de 7 cm pH 3-10. Na figura 25A é possível visualizar o padrão de separação do extrato proteico utilizado através da coloração do 2D-PAGE por Azul de Coomassie Coloidal, sendo obtida uma boa resolução dos spots e poucos arrastes. Os spots proteicos foram resolvidos em uma ampla faixa de pH e de peso molecular.

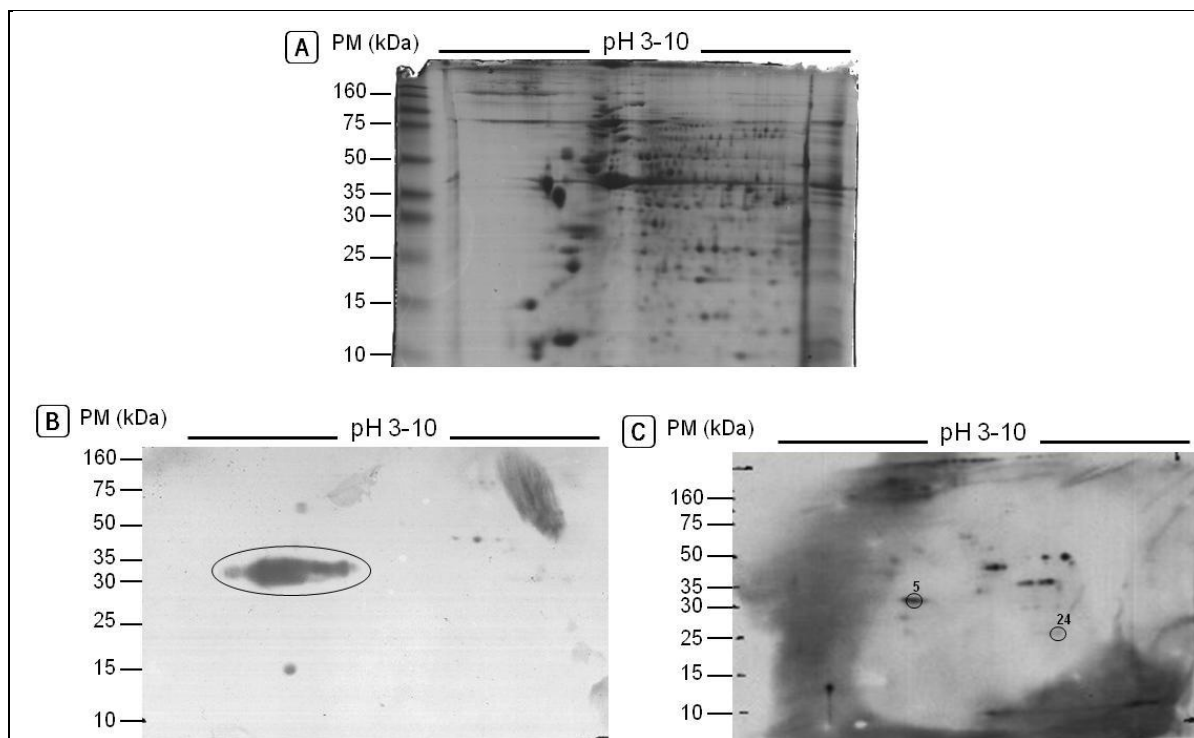


Figura 25 – Perfil eletroforético bidimensional do extrato proteico de vermes adultos de *Schistosoma mansoni* enriquecido de proteínas de membrana e Western blots bidimensionais utilizando anticorpo anti-Sm29 e pool de soro de indivíduos infectados.

100 µg de proteínas foram separadas por eletroforese bidimensional utilizando fitas de IPG de 7 cm, pH 3-10 e SDS-PAGE 12% corado por Azul de Coomassie Coloidal (A). Western blots bidimensionais correspondentes utilizando anticorpo anti-Sm29 (B) e pool de soro de indivíduos infectados (C). As membranas foram posteriormente incubadas com anticorpo secundário anti-IgG de camundongo ou anti-Ig's total humana, respectivamente, ambos conjugados à HRP. Para revelação foi utilizado o kit ECL Plus Western Blotting Detection (GE Healthcare). Padrão de peso molecular: Full range Rainbow (GE Healthcare).

Uma vez otimizado o protocolo de obtenção e separação de proteínas do extrato protéico de vermes adultos de *S. mansoni* enriquecido de proteínas de membrana por eletroforese bidimensional, partimos para a realização dos experimentos de Western-blotting bidimensional. Para confirmar a presença de proteínas de membrana de células do parasito neste extrato protéico, foi utilizado um anticorpo anti-Sm29 recombinante, gentilmente cedido pelo Dr. Sérgio Costa Oliveira (UFMG). A proteína Sm29 está localizada na superfície de esquistossômulos de fase pulmonar e de vermes adultos (Cardoso, Macedo *et al.*, 2008) e foi também identificada na fração de proteínas de membrana em uma análise do proteoma do tegumento de *S. mansoni* (Braschi e Wilson, 2006). Um conjunto de spots de

aproximadamente 29 kDa, presente na porção do gel onde estão localizadas proteínas ácidas, foi fortemente reconhecido pelo anticorpo anti-Sm29 (Figura 25B), demonstrando, portanto, que o protocolo utilizado para obtenção do extrato protéico foi capaz de extrair proteínas de membranas do tegumento do parasito. Além disso, estas proteínas foram separadas por eletroforese bidimensional, obtendo-se um perfil satisfatório de separação de spots protéicos, superando uma das dificuldades encontradas na análise de proteínas de membrana por eletroforese bidimensional.

O mesmo extrato protéico foi também utilizado em um experimento de Western-blotting bidimensional utilizando pool de soro de indivíduos infectados de área endêmica para esquistossomose. Como observado na Figura 25C, alguns spots protéicos reagiram ao pool de soro utilizado. Estes spots foram localizados e excisados do 2D-PAGE correspondente corado por Azul de Coomassie Colloidal (Figura 25A) e submetidos à análises por espectrometria de massas para identificação das proteínas.

Dentre os spots imunorreativos que foram identificados por espectrometria de massas, dois correspondem a proteínas que foram selecionadas anteriormente pela abordagem computacional, **Smp_171780** (Spot 5, Figura 25C) e **Smp_131910** (Spot 24, Figura 25C).

A Smp_171780 corresponde a uma proteína hipotética de *S. mansoni*. De acordo com os nossos dados obtidos pela predição computacional, esta proteína possui peptídeo sinal, têm evidência de expressão em cercária e verme adulto e ainda possui dois domínios funcionais: superfamília de inibidores de serina protease e EF-hand, um domínio estrutural comum em proteínas carreadoras de cálcio.

A Smp_131910, anotada como uma proteína de superfície de *S. mansoni* também possui peptídeo sinal e forte evidência de expressão em verme adulto e esquistossômulo. Seus domínios correspondem à superfamília de ligantes à galatose e superfamília de imunoglobulinas.

Embora não estejam na lista das 90 proteínas que foram previamente selecionadas para validação experimental por microarranjo de proteínas, outros experimentos precisam ser feitos para validar efetivamente essas duas proteínas, sobretudo a Smp_131910, a qual parece ser um excelente candidato a um estudo mais aprofundado, em uma busca por similaridade preliminar, observamos que tal proteína não tem evidência significativa de expressão em

humano. Estes resultados preliminares mostram que a abordagem computacional pode ser combinada às técnicas existentes para obtenção de novas possibilidades.

5.3.2. Resultados *in silico*

O presente projeto resultou no desenvolvimento de uma metodologia em larga escala, sem precedentes em eucariotos, baseado em análises em *in silico* na busca de candidatos à vacina contra esquistossomose. Permitiu também a criação de um *pipeline* e de um *webserver* para automatização de todo o processo que está disponível no endereço <http://epifinder.cebio.org> e traz recursos para acelerar a pesquisa de vacinas em outros organismos, principalmente outras espécies de *Schistosoma*.

Uma vantagem do trabalho foi o fato de que a metodologia que propomos, tem permitido a rápida triagem do proteoma na classificação proteínas expostas de outros grupos e organismos na instituição. Embora ainda não tenhamos a comprovação efetiva em bancada, estamos convencidos de que os objetivos foram atingidos, visto que a diminuição do número de bons potencialmente candidatos, com relação ao total de dados no genoma, se mostra eficaz pois reduziu a um número plausível os testes experimentais.

Alguns candidatos na Tabela 7 merecem ser destacados: **Smp_147250.1** (multidrug resistance-associated protein, putative) e **Smp_147250.2** que são anotadas como genes de resistência a droga. Há registros na literatura sobre o resistência à drogas e o papel de vacinas no seu combate (Kaplan, Mason *et al.*, 2004).

Outra potencial candidata interessante que também está relacionada a resistência à droga é a **Smp_171740**, proteína de membrana que possui domínios funcionais relacionados com transporte e apresentação de antígenos. Um trabalho recente demonstra que tal proteína é altamente expressa em vermes jovens quando recebem o tratamento do PZQ (Kasinathan, Morgan *et al.*, 2010).

Podemos observar também que muitos candidatos da tabela, não possuem nenhuma anotação (expressed protein), e que merecem serem avaliados de forma mais crítica.

Outros exemplos que também merecem ser analisados são a **Smp_167690** (*ubiquitin-protein ligase, putative*) e a **Smp_169750** (*dicer-1, putative*), que são proteínas

reconhecidamente da maquinaria celular, mas que no caso da ubiquitina, há na literatura registros de uma nova classe desta proteína associada à secreção (Tytgat, Vanholme *et al.*, 2004). No entanto tais dados precisam ser melhores investigados, pois podem se tratar de erros nos preditores por localização ou limitações da técnica, aqui proposta, e que serão relatadas na seção 7.

Submetemos os candidatos clássicos (Tabela 1) a análise aqui proposta, dois deles merecem destaque: a *integral membrane protein Sm23* (Smp_017430) e **Sm29** (Smp_072190). A Sm23 é uma proteína de membrana e a Sm29 é uma proteína da superfície tegumentar do parasita, ambas apresentam peptídeo sinal e tem reconhecida resposta a infecção do parasito. Tanto a Sm23 quanto a Sm29 passam em todas as etapas de predição de nossa análise, mas não são vistas em nossa seleção final, pois apresentam um baixo número de ligantes ao MHC, respectivamente 2 e 1 epitopos. Outros candidatos da lista não são identificados em nossas predições, como por exemplo a proteína *CD9-like protein Sm-TSP-1* (Smp_095630) da família das tetraspaninas, em nossa análise podemos confirmar sua exposição ao hospedeiro, mas não foi possível identificar ligantes ao MHC nesta proteína. Quanto às demais proteínas da lista de alvos conhecidos, não há indícios de que sejam expostas e nossa técnica não é 100% inclusiva.

Embora nossa predição de epitopos não apresente resultados significativos, casos de sucesso na literatura apontam que estamos no caminho certo. Um estudo recente em *Schistoma japonicum* que utiliza uma abordagem *in silico* semelhante a aqui apresentada, juntamente com validação experimental, a partir do transcriptoma do organismo, selecionou 6 peptídeos expressos na fase pulmonar do parasito, dos quais 3 foram capazes de induzir resposta imune *in vitro* tipo CD4+ (Zhang, Jia *et al.*, 2012), tais resultados demonstram que potenciais epitopos podem ser identificados rapidamente por uma combinação de análise *in silico* e quando combinados com validação experimental produzem bons resultados.

1. CONCLUSÕES

Este trabalho foi um dos primeiros a aplicar uma abordagem de vacinologia reversa em parasita, e a usar uma metodologia que combinava diversas abordagens para obtenção de possíveis alvos de estudo para desenvolvimento de vacina eficaz contra esquistossomose. Assim:

- Esperamos que o EpiFinder seja útil a comunidade científica, e que está também possa sugerir novas funcionalidades ou características importantes para o estudo de alvos de vacinas.
- Além da predição de proteínas antigênicas, o trabalho resultou em análises importantes, como por exemplo, é possível extrair de nossas predições o secretoma de *S. mansoni* (conjunto de proteínas secretadas do organismo).
- O banco criado para nossas predições pode ser estendido para outras predições, aproveitando a metodologia de integração para descoberta de conhecimento.
- A metodologia descrita demonstrou que a técnica de integração de predições em larga escala, para seleção de alvos de vacinas, pode contribuir no processo de direcionamento e aceleração dos experimentos laboratoriais.
- Embora experimentos para a validação da metodologia estejam em andamento, acreditamos que esta abordagem seja de grande valor para a seleção de potenciais antígenos.

2. LIMITAÇÕES E PERSPECTIVAS

Durante o desenvolvimento do projeto algumas limitações da técnica aqui descrita merecem ser destacadas. O genoma de *Schistosoma mansoni* ainda não possui uma montagem final, o que nos leva a ter problemas recorrentes na determinação dos genes, principalmente no que diz respeito às regiões limítrofes (N-Terminal e C-Terminal) o que torna crítica a obtenção de proteínas secretadas, uma vez que o peptídeo sinal comumente se encontra nestes compartimentos. Porém com a melhora desta montagem podemos refazer estas predições afim de reduzir erros.

Vale ressaltar a limitação inerente na predição de epitopos, uma vez que todos os métodos têm modelos de alelos específicos, o que dificulta integrar e comparar tais análises e também nos impede de obter peptídeos ligantes de tamanho diferente do modelo usado.

Temos também encontrado algumas dificuldades em integrar a metodologia empregada aqui com outras abordagens na busca por alvos de vacina, por exemplo, a proteômica experimental que em muitas de suas técnicas não leva em conta proteínas de superfície, não encontrando interseções com nosso método para então a escolha de melhores candidatos assim, acreditamos que as abordagens experimental e computacional sejam complementares.

O kit da *GE Healthcare* que foi usado no experimento de proteômica para extração das proteínas hidrofóbicas, embora tenha apresentado um resultado preliminar interessante, não é o ideal para a avaliação experimental deste trabalho, embora o kit seja indicado para a extração de proteínas hidrofóbicas podemos verificar poucas proteínas com domínios de membrana de acordo com a predições por localização.

Como perspectivas, esperamos disponibilizar através do SchistoDB, a informação obtida, e também esperamos que novas análises possam ser incorporadas tanto ao banco quanto ao pipeline, afim de que possamos obter melhores candidatos.

Uma das análises já está sendo realizada em colaboração com grupo do Dr. Franco Falcone da Universidade de Nottingham conduzida pela estudante de Doutorado do nosso grupo Fernanda Ludolf. Tal validação será realizada tanto para proteínas identificadas em nossa abordagem computacional, como também aquelas obtidas por proteômica experimental.

O grupo do Dr. Falcone grupo possui uma abordagem *Cell-Free* para expressão de antígenos em arrays de peptídeo. Neste sistema os cDNAs dos genes selecionados para validação são amplificados por PCR usando clones sequenciados completos, de bibliotecas de

cDNA, ou DNA total, conforme apropriado. O grupo possui aproximadamente 14.000 clones de cDNA sequenciados individualmente (gentilmente doado pelo Dr. Alan Wilson), bibliotecas de clones produzidas localmente e uma biblioteca de DNA total cedida por outros grupos.

Aos iniciadores destinados à amplificação de cada gene é adicionado a sequência do promotor T7 e a metionina inicial, essencial durante o processo de iniciação da tradução. Aos iniciadores reversos é acrescentada à cauda de histidina para avaliar os níveis de proteína por Dot-Blot. A expressão da proteína é então avaliada pela incorporação de um corante fluorescente (FluoroTect™ GreenLys) e detectada diretamente no gel. A precisão da PCR é confirmada por sequenciamento. A otimização dos ensaios de transcrição e transcrição *in vitro* é realizada utilizando os genes selecionados que representam os antígenos melhor estudados. Todas as reações de tradução *in vitro* são aplicadas diretamente para um arranjo em triplicata e incubadas com soro de pacientes de área endêmica infectados. Estes resultados são comparados com a resposta à anticorpos descrita para esses antígenos na literatura.

Outra possibilidade de trabalho futuro que visa melhorar nossa abordagem, é a inclusão de métodos de predição de célula B, os quais foram descartados no início do projeto, por não apresentarem bons resultados de acordo com a literatura, mas que agora pode ser combinado com a predição de célula T que já dominamos, possibilitando novos critérios de filtragem em uma análise em larga escala.

Espera-se ainda que no fim deste trabalho todos os resultados e o conhecimento adquirido no desenvolvimento da metodologia aqui estabelecida, sejam futuramente incorporados ao SchistoDB.

Alguns projetos já aqui citados nos indicam que as predições computacionais identificaram com sucesso possíveis candidatos à vacina. Objetivando confirmar os resultados dessa metodologia, outros testes serão realizados, objetivando tanto agregar novos possíveis antígenos, como nos proverá mais informações sobre esta metodologia e sobre o organismo em estudo.

3. REFERÊNCIAS BIBLIOGRÁFICAS

ABBAS, A. K.; LICHTMAN, A. H.; PILLAI, S. **Cellular and molecular immunology**. 6th. Philadelphia: Saunders Elsevier, 2007. viii, 566 p. ISBN 9781416031222

AHMAD, G. et al. Sm-p80-based DNA vaccine formulation induces potent protective immunity against *Schistosoma mansoni*. **Parasite Immunol**, v. 31, n. 3, p. 156-61, Mar 2009. ISSN 1365-3024 (Electronic)

AITKEN, R. et al. Radiation-resistant acquired immunity of vaccinated mice to *Schistosoma mansoni*. **Am J Trop Med Hyg**, v. 37, n. 3, p. 570-7, Nov 1987. ISSN 0002-9637 (Print)

AL-SHERBINY, M. et al. In vitro cellular and humoral responses to *Schistosoma mansoni* vaccine candidate antigens. **Acta Trop**, v. 88, n. 2, p. 117-30, Oct 2003. ISSN 0001-706X (Print)

ARIEL, N. et al. Search for potential vaccine candidate open reading frames in the *Bacillus anthracis* virulence plasmid pXO1: in silico and in vitro screening. **Infect Immun**, v. 70, n. 12, p. 6817-27, Dec 2002. ISSN 0019-9567 (Print)

ARONSTEIN, W. S.; STRAND, M. Gender-specific and pair-dependent glycoprotein antigens of *Schistosoma mansoni*. **J Parasitol**, v. 70, n. 4, p. 545-57, Aug 1984. ISSN 0022-3395 (Print)

AURIAULT, C. et al. Epitopic characterization and vaccinal potential of peptides derived from a major antigen of *Schistosoma mansoni* (Sm28 GST). **Pept Res**, v. 4, n. 1, p. 6-11, Jan-Feb 1991. ISSN 1040-5704 (Print)

BAMBINI, S.; RAPPUOLI, R. The use of genomics in microbial vaccine development. **Drug Discov Today**, v. 14, n. 5-6, p. 252-60, Mar 2009. ISSN 1878-5832 (Electronic)

BENDTSEN, J. D. et al. Feature-based prediction of non-classical and leaderless protein secretion. **Protein Eng Des Sel**, v. 17, n. 4, p. 349-56, Apr 2004. ISSN 1741-0126 (Print)

BENDTSEN, J. D. et al. Improved prediction of signal peptides: SignalP 3.0. **J Mol Biol**, v. 340, n. 4, p. 783-95, Jul 16 2004. ISSN 0022-2836 (Print)

BERRIMAN, M. et al. The genome of the blood fluke *Schistosoma mansoni*. **Nature**, v. 460, n. 7253, p. 352-8, Jul 16 2009. ISSN 1476-4687 (Electronic)

BETTS, J. C. Transcriptomics and proteomics: tools for the identification of novel drug targets and vaccine candidates for tuberculosis. **IUBMB Life**, v. 53, n. 4-5, p. 239-42, Apr-May 2002. ISSN 1521-6543 (Print)

BISWAS, R., ORT, E. The Java Persistence API - A Simpler Programming Model for Entity Persistence. 2006. Acesso em: 03-27-2012.

BRASCHI, S.; WILSON, R. A. Proteins exposed at the adult schistosome surface revealed by biotinylation. **Mol Cell Proteomics**, v. 5, n. 2, p. 347-56, Feb 2006. ISSN 1535-9476 (Print)

BURNS, N. et al. Large-scale analysis of gene expression, protein localization, and gene disruption in *Saccharomyces cerevisiae*. **Genes Dev**, v. 8, n. 9, p. 1087-105, May 1 1994. ISSN 0890-9369 (Print)

CARDOSO, F. C. et al. *Schistosoma mansoni* tegument protein Sm29 is able to induce a Th1-type of immune response and protection against parasite infection. **PLoS Negl Trop Dis**, v. 2, n. 10, p. e308, 2008. ISSN 1935-2735 (Electronic)

CARDOSO, F. C. et al. Human antibody responses of patients living in endemic areas for schistosomiasis to the tegumental protein Sm29 identified through genomic studies. **Clin Exp Immunol**, v. 144, n. 3, p. 382-91, Jun 2006. ISSN 0009-9104 (Print)

CHAKRAVARTI, D. N. et al. Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. **Vaccine**, v. 19, n. 6, p. 601-12, Nov 8 2000. ISSN 0264-410X (Print).

CHEEVER, A. W. et al. Kinetics of egg production and egg excretion by *Schistosoma mansoni* and *S. japonicum* in mice infected with a single pair of worms. **Am J Trop Med Hyg**, v. 50, n. 3, p. 281-95, Mar 1994. ISSN 0002-9637 (Print)

CHITSULO, L. et al. The global status of schistosomiasis and its control. **Acta Trop**, v. 77, n. 1, p. 41-51, Oct 23 2000. ISSN 0001-706X (Print)

CHOU, K. C.; SHEN, H. B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. **J Proteome Res**, v. 5, n. 8, p. 1888-97, Aug 2006. ISSN 1535-3893 (Print)

CORREA-OLIVEIRA, R.; CALDAS, I. R.; GAZZINELLI, G. Natural versus drug-induced resistance in *Schistosoma mansoni* infection. **Parasitol Today**, v. 16, n. 9, p. 397-9, Sep 2000. ISSN 0169-4758 (Print)

CRISCIONE, C. D. et al. Genomic linkage map of the human blood fluke *Schistosoma mansoni*. **Genome Biol**, v. 10, n. 6, p. R71, 2009. ISSN 1465-6914 (Electronic)

CUI, J. et al. MHC-BPS: MHC-binder prediction server for identifying peptides of flexible lengths from sequence-derived physicochemical properties. **Immunogenetics**, v. 58, n. 8, p. 607-13, Aug 2006. ISSN 0093-7711 (Print)

DAVIS, A. H. Schistosomiasis. **Epidemiology and the Community Control of Disease in Warm Climate Countries**, n. 2nd ed., p. 389-412, 1984.

DE SILVA, N.; GUYATT, H.; BUNDY, D. Anthelmintics. A comparative review of their clinical pharmacology. **Drugs**, v. 53, n. 5, p. 769-88, May 1997. ISSN 0012-6667 (Print)

DEGRAVE, W. M. et al. Parasite genome initiatives. **Int J Parasitol**, v. 31, n. 5-6, p. 532-6, May 1 2001. ISSN 0020-7519 (Print)

DEITEL, H. M. **Java: Como Programar**. São Paulo: Pearson Education do Brasil, 2005.

DOENHOFF, M. J.; CIOLI, D.; UTZINGER, J. Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis. **Curr Opin Infect Dis**, v. 21, n. 6, p. 659-67, Dec 2008. ISSN 1535-3877 (Electronic)

DOENHOFF, M. J.; PICA-MATTOCCIA, L. Praziquantel for the treatment of schistosomiasis: its use for control in areas with endemic disease and prospects for drug resistance. **Expert Rev Anti Infect Ther**, v. 4, n. 2, p. 199-210, Apr 2006. ISSN 1744-8336 (Electronic)

DONNES, P.; KOHLBACHER, O. SVMHC: a server for prediction of MHC-binding peptides. **Nucleic Acids Res**, v. 34, n. Web Server issue, p. W194-7, Jul 1 2006. ISSN 1362-4962 (Electronic)

DRAWID, A.; GERSTEIN, M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. **J Mol Biol**, v. 301, n. 4, p. 1059-75, Aug 25 2000. ISSN 0022-2836 (Print)

DUNKLEY, T. P. et al. The use of isotope-coded affinity tags (ICAT) to study organelle proteomes in *Arabidopsis thaliana*. **Biochem Soc Trans**, v. 32, n. Pt3, p. 520-3, Jun 2004. ISSN 0300-5127 (Print)

EMANUELSSON, O. et al. Locating proteins in the cell using TargetP, SignalP and related tools. **Nat Protoc**, v. 2, n. 4, p. 953-71, 2007. ISSN 1750-2799 (Electronic).

ENGELS, D. et al. The global epidemiological situation of schistosomiasis and new approaches to control and research. **Acta Trop**, v. 82, n. 2, p. 139-46, May 2002. ISSN 0001-706X (Print)

FELDHAHN, M. et al. FRED--a framework for T-cell epitope detection. **Bioinformatics**, v. 25, n. 20, p. 2758-9, Oct 15 2009. ISSN 1367-4811 (Electronic)

FONSECA, C. T. et al. Identification of paramyosin T cell epitopes associated with human resistance to *Schistosoma mansoni* reinfection. **Clin Exp Immunol**, v. 142, n. 3, p. 539-47, Dec 2005. ISSN 0009-9104 (Print)

FRANCO, G. R. et al. The *Schistosoma* gene discovery program: state of the art. **Int J Parasitol**, v. 30, n. 4, p. 453-63, Apr 10 2000. ISSN 0020-7519 (Print)

GALAXY. Galaxy - Easy-to-use, open-source, scalable framework for tool and data integration. 2010.

GAMMA, E., HELM, R., JOHNSON, R., VLISSIDES, J. **Design Patterns: Elements of Reusable Object-Oriented Software**. Addison Wesley Professional, 1994. ISBN 978-0201633610.

GOLD. Genomes Online Database. 2011. Disponível em: < <http://www.genomesonline.org/cgi-bin/GOLD/bin/gold.cgi> >.

GOMES, Y. M. P. **Java na Web com JSF, Spring, Hibernate e Netbeans 6**. Rio de Janeiro: Editora Ciência Moderna, 2008.

GUAN, P. et al. MHCpred 2.0: an updated quantitative T-cell epitope prediction server. **Appl Bioinformatics**, v. 5, n. 1, p. 55-61, 2006. ISSN 1175-5636 (Print)

HAMMER, J. New methods to predict MHC-binding sequences within protein antigens. **Curr Opin Immunol**, v. 7, n. 2, p. 263-9, Apr 1995. ISSN 0952-7915 (Print)

HANSON, M. R.; KOHLER, R. H. GFP imaging: methodology and application to investigate cellular compartmentation in plants. **J Exp Bot**, v. 52, n. 356, p. 529-39, Apr 2001. ISSN 0022-0957 (Print)

HOGLUND, A. et al. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. **Bioinformatics**, v. 22, n. 10, p. 1158-65, May 15 2006. ISSN 1367-4803 (Print)

JOHNSON, R. **Expert One-on-One: JEE Design e Development**. 1st. Wrox, 2003. 750 ISBN 1861007841.

KANOI, B. N.; EGWANG, T. G. New concepts in vaccine development in malaria. **Curr Opin Infect Dis**, v. 20, n. 3, p. 311-6, Jun 2007. ISSN 0951-7375 (Print)

KAPLAN, S. L. et al. Decrease of invasive pneumococcal infections in children among 8 children's hospitals in the United States after the introduction of the 7-valent pneumococcal conjugate vaccine. **Pediatrics**, v. 113, n. 3 Pt 1, p. 443-9, Mar 2004. ISSN 1098-4275 (Electronic)

KARIUKI, T. M. et al. Parameters of the attenuated schistosome vaccine evaluated in the olive baboon. **Infect Immun**, v. 72, n. 9, p. 5526-9, Sep 2004. ISSN 0019-9567 (Print)

KASINATHAN, R. S.; MORGAN, W. M.; GREENBERG, R. M. Schistosoma mansoni express higher levels of multidrug resistance-associated protein 1 (SmMRP1) in juvenile worms and in response to praziquantel. **Mol Biochem Parasitol**, v. 173, n. 1, p. 25-31, Sep 2010. ISSN 1872-9428 (Electronic)

KATZ, N.; FARIA, I.; REIS, F. Modernos Conhecimentos sobre Esquistossomose Mansônica. **Academia Mineira de Medicina**, 1986.

KATZ, N.; PEIXOTO, S. V. [Critical analysis of the estimated number of Schistosomiasis mansoni carriers in Brazil]. **Rev Soc Bras Med Trop**, v. 33, n. 3, p. 303-8, May-Jun 2000. ISSN 0037-8682 (Print)

KENT, W. J. BLAT--the BLAST-like alignment tool. **Genome Res**, v. 12, n. 4, p. 656-64, Apr 2002. ISSN 1088-9051 (Print)

LAUEMOLLER, S. L. et al. Identifying cytotoxic T cell epitopes from genomic and proteomic information: "The human MHC project.". **Rev Immunogenet**, v. 2, n. 4, p. 477-91, 2000.

LOVERDE, P. T.; CHEN, L. Schistosome female reproductive development. **Parasitol Today**, v. 7, n. 11, p. 303-8, Nov 1991. ISSN 0169-4758 (Print)

LOVERDE, P. T. et al. Schistosoma mansoni genome project: an update. **Parasitol Int**, v. 53, n. 2, p. 183-92, Jun 2004. ISSN 1383-5769 (Print).

LUND, O. **Immunological bioinformatics**. Cambridge, Mass.: MIT Press, 2005. xii, 296 p., [24] p. of plates ISBN 0262122804 (alk. paper).

MARCOTTE, E. M. et al. Localizing proteins in the cell from their phylogenetic profiles. **Proc Natl Acad Sci U S A**, v. 97, n. 22, p. 12115-20, Oct 24 2000.

MCMANUS, D. P.; LOUKAS, A. Current status of vaccines for schistosomiasis. **Clin Microbiol Rev**, v. 21, n. 1, p. 225-42, Jan 2008. ISSN 1098-6618 (Electronic)

MEI, H.; LOVERDE, P. T. Schistosoma mansoni: the developmental regulation and immunolocalization of antioxidant enzymes. **Exp Parasitol**, v. 86, n. 1, p. 69-78, May 1997. ISSN 0014-4894 (Print)

MIDDLETON, D. et al. New allele frequency database: <http://www.allelefrequencys.net>. **Tissue Antigens**, v. 61, n. 5, p. 403-7, May 2003. ISSN 0001-2815 (Print)

MONTIGIANI, S. et al. Genomic approach for analysis of surface proteins in Chlamydia pneumoniae. **Infect Immun**, v. 70, n. 1, p. 368-79, Jan 2002. ISSN 0019-9567 (Print)

MYSQL. MySQL - The world's most popular open source database. 2010.

NAIR, R.; ROST, B. Inferring sub-cellular localization through automated lexical analysis. **Bioinformatics**, v. 18 Suppl 1, p. S78-86, 2002.

NCBI-DBEST. Expressed Sequence Tags database. 2008.

NETBEANS. **Netbeans IDE**: Available at: <http://netbeans.org/community/releases/67/> 2011.

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

NIELSEN, M. et al. Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan. **PLoS Comput Biol**, v. 4, n. 7, p. e1000107, 2008. ISSN 1553-7358 (Electronic).

NIELSEN, M.; LUNDEGAARD, C.; LUND, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. **BMC Bioinformatics**, v. 8, p. 238, 2007. ISSN 1471-2105 (Electronic)

OLIVEIRA, G. The Schistosoma mansoni transcriptome: An update. **Exp Parasitol**, Jun 12 2007. ISSN 0014-4894 (Print).

OLIVEIRA, G.; JOHNSTON, D. A. Mining the schistosome DNA sequence database. **Trends Parasitol**, v. 17, n. 10, p. 501-3, Oct 2001. ISSN 1471-4922 (Print).

PEARCE, E. J.; MACDONALD, A. S. The immunobiology of schistosomiasis. **Nat Rev Immunol**, v. 2, n. 7, p. 499-511, Jul 2002. ISSN 1474-1733 (Print)

PEARSON, R. D. Schistosomiasis (Bilharziasis). **The Merck Manual**, 2009. Acesso em: 15-11.2011.

PERL. The perl programming language. 2010.

PICA-MATTOCCIA, L.; CIOLI, D. Sex- and stage-related sensitivity of Schistosoma mansoni to in vivo and in vitro praziquantel treatment. **Int J Parasitol**, v. 34, n. 4, p. 527-33, Mar 29 2004. ISSN 0020-7519 (Print)

PIZZA, M. et al. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. **Science**, v. 287, n. 5459, p. 1816-20, Mar 10 2000. ISSN 0036-8075 (Print).

PROTASIO, A. V. et al. A systematically improved high quality genome and transcriptome of the human blood fluke *Schistosoma mansoni*. **PLoS Negl Trop Dis**, v. 6, n. 1, p. e1455, Jan 2012. ISSN 1935-2735 (Electronic)

RAMMENSEE, H. et al. SYFPEITHI: database for MHC ligands and peptide motifs. **Immunogenetics**, v. 50, n. 3-4, p. 213-9, Nov 1999.

RAPPUOLI, R. Reverse vaccinology, a genome-based approach to vaccine development. **Vaccine**, v. 19, n. 17-19, p. 2688-91, Mar 21 2001. ISSN 0264-410X (Print)

RECHE, P. A.; GLUTTING, J. P.; REINHERZ, E. L. Prediction of MHC class I binding peptides using profile motifs. **Hum Immunol**, v. 63, n. 9, p. 701-9, Sep 2002. ISSN 0198-8859 (Print)

ROLLINSON, D.; SIMPSON, A. J. G. The Biology of Schistosomes: from Genes to Latrines. **The American Journal of Tropical Medicine and Hygiene**, 1988.

ROSS, B. C. et al. Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. **Vaccine**, v. 19, n. 30, p. 4135-42, Jul 20 2001. ISSN 0264-410X (Print)

SHATKAY, H. et al. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. **Bioinformatics**, v. 23, n. 11, p. 1410-7, Jun 1 2007. ISSN 1367-4811 (Electronic)

SINGH, H.; RAGHAVA, G. P. ProPred: prediction of HLA-DR binding sites. **Bioinformatics**, v. 17, n. 12, p. 1236-7, Dec 2001. ISSN 1367-4803 (Print)

SMYTH, D. et al. Isolation of cDNAs encoding secreted and transmembrane proteins from *Schistosoma mansoni* by a signal sequence trap method. **Infect Immun**, v. 71, n. 5, p. 2548-54, May 2003. ISSN 0019-9567 (Print)

SONNHAMMER, E. L.; VON HEIJNE, G.; KROGH, A. A hidden Markov model for predicting transmembrane helices in protein sequences. **Proc Int Conf Intell Syst Mol Biol**, v. 6, p. 175-82, 1998. ISSN 1553-0833 (Print)

SQLYOG. MySQL GUI Tools. MySQL Monitor and Menager. 2010.

STELMA, F. F. et al. Efficacy and side effects of praziquantel in an epidemic focus of *Schistosoma mansoni*. **Am J Trop Med Hyg**, v. 53, n. 2, p. 167-70, Aug 1995. ISSN 0002-9637 (Print)

STURNIOLO, T. et al. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. **Nat Biotechnol**, v. 17, n. 6, p. 555-61, Jun 1999. ISSN 1087-0156 (Print)

TENDLER, M. et al. Vaccination against schistosomiasis and fascioliasis with the new recombinant antigen Sm14: potential basis of a multi-valent anti-helminth vaccine? **Mem Inst Oswaldo Cruz**, v. 90, n. 2, p. 255-6, Mar-Apr 1995. ISSN 0074-0276 (Print)

TRAN, M. H. et al. Tetraspanins on the surface of *Schistosoma mansoni* are protective antigens against schistosomiasis. **Nat Med**, v. 12, n. 7, p. 835-40, Jul 2006. ISSN 1078-8956 (Print)

TYTGAT, T. et al. A new class of ubiquitin extension proteins secreted by the dorsal pharyngeal gland in plant parasitic cyst nematodes. **Mol Plant Microbe Interact**, v. 17, n. 8, p. 846-52, Aug 2004. ISSN 0894-0282 (Print)

VERJOVSKI-ALMEIDA, S. et al. Transcriptome analysis of the acelomate human parasite *Schistosoma mansoni*. **Nat Genet**, v. 35, n. 2, p. 148-57, Oct 2003. ISSN 1061-4036 (Print).

VON HEIJNE, G. The signal peptide. **J Membr Biol**, v. 115, n. 3, p. 195-201, May 1990. ISSN 0022-2631 (Print)

WAN, J. et al. SVRMHC prediction server for MHC-binding peptides. **BMC Bioinformatics**, v. 7, p. 463, 2006. ISSN 1471-2105 (Electronic)

WHO. Preventive Chemotherapy Databank. 2011. Disponível em: <
http://www.who.int/neglected_diseases/preventive_chemotherapy/databank/en/index.html>.
Acesso em: 17-11-2011.

WIZEMANN, T. M. et al. Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. **Infect Immun**, v. 69, n. 3, p. 1593-8, Mar 2001. ISSN 0019-9567 (Print)

WRIGHT, M. D. et al. Serologic reactivities of the 23-kDa integral membrane proteins of schistosomes. **J Immunol**, v. 147, n. 12, p. 4338-42, Dec 15 1991. ISSN 0022-1767 (Print)

ZERLOTINI, A. et al. SchistoDB: a *Schistosoma mansoni* genome resource. **Nucleic Acids Res**, v. 37, n. Database issue, p. D579-82, Jan 2009. ISSN 1362-4962 (Electronic)

ZHANG, G. L. et al. MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides. **Nucleic Acids Res**, v. 33, n. Web Server issue, p. W172-9, Jul 1 2005. ISSN 1362-4962 (Electronic)

ZHANG, Y. et al. Identification of Th1 epitopes within molecules from the lung-stage schistosomulum of *Schistosoma japonicum* by combining prediction analysis of the transcriptome with experimental validation. **Parasitol Int**, May 19 2012. ISSN 1873-0329 (Electronic)

4. ANEXOS

Disponibilizados em CD.