Ministério da Saúde

Fundação Oswaldo Cruz

Centro de Pesquisa René Rachou

Programa de Pós-graduação em Ciências da Saúde

Desenvolvimento de uma estratégia automática para busca de potenciais antígenos em proteomas preditos de *Plasmodium* spp

por

Ricardo de Souza Ribeiro

Belo Horizonte
Fevereiro/2013

Ministério da Saúde

Fundação Oswaldo Cruz

Centro de Pesquisa René Rachou

Programa de Pós-graduação em Ciências da Saúde

Desenvolvimento de uma estratégia automática para busca de potenciais antígenos em proteomas preditos de *Plasmodium* spp

por

Ricardo de Souza Ribeiro

Dissertação apresentada com vistas à obtenção do Título de Mestre em Ciências na área de concentração em Biologia Celular e Molecular.

Orientação: Dra. Cristiana Ferreira Alves de Brito

Belo Horizonte

Fevereiro/2013

Catalogação-na-fonte

Rede de Bibliotecas da FIOCRUZ

Biblioteca do CPqRR

Segemar Oliveira Magalhães CRB/6 1975

R484d

Ribeiro, Ricardo de Souza.

2013

Desenvolvimento de uma estratégia automática para busca de potenciais antígenos em proteomas preditos de *Plasmodium* spp / Ricardo de Souza Ribeiro. – Belo Horizonte, 2013.

xii, 73 f.: il.; 210 x 297mm.

Bibliografia: f.: 82 - 85

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

1. Malária/imunologia 2. *Plasmodium*/parasitologia 3. Eritrócitos/parasitologia I. Título. II. Brito, Cristiana Ferreira Alves de (Orientação).

CDD – 22. ed. – 616.936 2

Ministério da Saúde

Fundação Oswaldo Cruz

Centro de Pesquisas René Rachou

Programa de Pós-graduação em Ciências da Saúde

"Desenvolvimento de uma estratégia automática para busca de potenciais antígenos em proteomas preditos de *Plasmodium* spp"

por

Ricardo de Souza Ribeiro

Foi avaliada pela banca examinadora composta pelos seguintes membros:

Profa. Dra. _Cristiana Ferreira Alves de Brito_____ (Presidente)

Prof. Dr. _Francisco Pereira Lobo_____

Prof. Dr. __Jeronimo Conceição Ruiz_____

Suplente: _Rômulo Lucio Vale de Moraes_____

Dissertação defendida e aprovada em: _27_/_02_/_2013_

_____

Agradecimentos

Agradecer não é uma coisa tão fácil quanto parece. Ainda mais quando você quer agradecer muitas pessoas por várias coisas durante um período de tempo de pouco mais de 2 anos. Dessa forma vou tentar colocar o mínimo de nomes possíveis, para que eu não seja injusto com ninguém.

Vou começar agradecendo a pessoa que sempre me deu motivos para agradecê-la independente do momento da minha vida: minha Mãe. Obrigado pelo apoio e por me proporcionar a oportunidade de conhecer o mundo da forma como eu o conheço. Mesmo eu não compartilhando de muitas idéias e crenças, você sempre me apoiou e me incentivou nas minhas decisões. Se eu cheguei até aqui é porque você me ajudou muito.

Gostaria de agradecer meu Pai. Obrigado pelo apoio e por acreditar que eu era capaz. Seu apoio também foi muito importante para que eu conseguisse. Obrigado. Quero agradecer também ao meu irmão pelo incentivo.

Eu não poderia deixar de agradecer a coisa mais valiosa que eu tenho na minha vida: meu filho João Gabriel. Obrigado por existir e por às vezes tornar as coisas um pouquinho mais difíceis. Porque se foi difícil com você, sem você seria impossível. Você é muito importante pra mim.

A história começou em 2008 quando eu fui parar no Laboratório de Malária como estudante de iniciação científica sob a orietação da Dra. Cristiana Brito. Ao longo destes quase cinco anos me faltam palavras para descrever a pessoa fantástica que eu tive a oportunidade de conviver. Entre meio a filho, formatura, seleção, divisões, você sempre me ajudou de várias formas. Muito obrigado Cris, primeiro por ser quem você é e segundo por ter me orientado neste trabalho. Neste período em que fui seu aluno você se mostrou ser muito mais do que uma excelente pesquisadora, professora e orientadora, você se mostrou ser um ser humano. Obrigado por me mostrar que é possível fazer diferente!

Gostaria de agradecer aos meus colegas do Laboratório, que é incrível! Muitos já se foram e muitos chegaram nesse período, por isso talvez fosse injusto colocar alguns nomes. Quem é especial pra mim por algum motivo sabe disso e eu gostaria de agradecer muito por ter tido a oportunidade de conhecer vocês. Eu gosto muito de fazer parte desse time! Só para deixar alguns com ciúmes eu gostaria de fazer um agradecimento especial a uma figura emblemática do laboratório, o Geraldo, por fazer a nossa alegria sempre, mesmo quando não está presente.

Quero agradecer também a todos do Cebio, que durante uma parte do tempo também foi minha casa. A todos muito obrigado pelo convívio. Aos colegas do Centro de Pesquisas René Rachou obrigado por tornar o ambiente de trabalho mais agradável.

Existem muitas outras pessoas fora do meio acadêmico que também foram igualmente importantes nesse período. A todos os meus amigos obrigados pelos momentos de descontração e pelas conversas impagáveis. Sem elas as coisas seriam mais tristes. Queria agradecer também à Tati, que acompanhou boa parte desse processo. Valeu pela paciência e pelos momentos de paz. Obrigado por tudo.

Sumário

Lista de figuras

Lista de tabelas

Lista de abreviaturas e símbolos

OMS: Organização Mundial de Saúde

SVS: Secretaria de Vigilância em Saúde

DDT: Dicloro Difenil Tricloroetano

PNCM: Programa Nacional de Controle da Malária

EERFs: Formas Replicativas Exoeritrocíticas

VP: Vacúolo Parasitóforo

RBC: Célula vermelha do sangue

RE: Retículo Endoplasmático

G: Complexo de Golgi

PS: Peptídeo Sinal

PEXEL: Plasmodium Export Element

PNEP: Pexel Negative Exported Protein

VTS: Vacuolar Transport Signal

NGS: Next Generation Sequencing

DNA: Ácido Desoxirribonucléico

ESPs: Proteínas Escretadas e Secretadas

IEDB: Immune Epitope Database

MHC: Major Histocompatibility Complex

HMM: Hidden Markov Model

XML: Extensible Markup Language

Resumo

A malária é uma doença que provoca um enorme impacto em todo mundo e muitos esforços tem sido feitos na tentativa de eliminar esta doença há séculos. O desenvolvimento de uma vacina eficaz contra esta doença se tornou prioridade para muitos grupos de pesquisa, entretanto apenas um antígeno apresentou resultados promissores em ensaios clínicos de fase III. O processo de identificação de candidatos promissores para comporem uma vacina é muito laborioso e demanda um tempo muito grande, principalmente devido à biologia complexa destes parasitos intracelulares. Identificar proteínas nesses tipos de patógenos intracelulares é um grande desafio e requer a análise de diferentes fatores ao mesmo tempo. Um bom candidato a antígeno deve possuir características que façam com que a proteína ou pelo menos parte dela, entre em contato com o sistema imune do hospedeiro em algum momento do ciclo de vida do parasito e que este contato seja capaz de gerar uma resposta imune capaz de neutralizar o parasito e acabar com a infecção. As proteínas secretadas/exportadas se encaixam perfeitamente nessas características e foi justamente com o objetivo de identificá-las que foi desenvolvida uma estratégia automática integrando diferentes programas para buscar estas proteínas. Dessa forma, o proteoma predito de 5 espécies diferentes do gênero Plasmodium foi submetidas à análise de três programas que procuram por características importantes para proteínas exportadas/secretadas nesse gênero. O peptídeo sinal foi investigado utilizando o SignalP, um script em Perl foi desenvolvido para buscar um motivo que é crucial para a exportação de algumas proteínas (PEXEL) e os domínios transmembrana foram buscados utilizando o TMHMM. Algumas proteínas selecionadas foram submetidas à predição de epitopos de células B. Com esta abordagem foi possível identificar algumas famílias de proteínas que já são conhecidas como proteínas exportadas, como Rifin e Stevor, o que mostra que esta abordagem pode ser uma ferramenta útil na identificação de novos prováveis alvos para o desenvolvimento de uma vacina eficaz. Utilizando esta metodologia, esperamos contribuir para aumentar o número de proteínas em testes para formulação de um antígeno que seja capaz de ajudar no controle e erradicação desta doença que causa tantos prejuízos para a humanidade.

Abstract

Malaria is a disease that causes a huge impact around the world and many efforts have been made in an attempt to eliminate this disease for centuries. The development of an effective vaccine against this disease has become a priority for many research groups, however only one antigen showed promising results in phase III clinical trials. The process of identifying promising candidates to compose a vaccine is very laborious and requires a very long time, mainly due to the complex biology of these intracellular parasites. Identify proteins in these types of intracellular pathogens is a major challenge and requires the analysis of different factors simultaneously. A good candidate antigen must possess characteristics that make the protein or at least part of it, contact the host immune system at some point in the life cycle of the parasite and that this contact is able to generate an immune response capable to neutralize the parasite and stop the infection. The proteins secreted / exported fit perfectly and it was precisely these characteristics in order to identify them we developed a strategy that automatically integrating different programs to find these proteins. Thus, the predicted proteome of 5 different species of the genus Plasmodium was subjected to analysis of three programs looking for important features for proteins exported / secreted in this genre. The signal peptide was investigated using SignalP, a Perl script was developed to find a subject that is crucial for the export of some proteins (PEXEL) and transmembrane domains using TMHMM were searched. Some selected proteins were subjected to prediction of B cell epitopes With this approach it was possible to identify some protein families that are already known as proteins exported as Rifin and Stevor, showing that this approach can be a useful tool in identifying new likely targets for the development of an effective vaccine. Using this methodology, we hope to contribute to increasing the number of proteins in tests for formulation of an antigen that is able to help control and eradicate the disease that causes so much harm to humanity.

1 INTRODUÇÃO

1.1 Malária

A malária é uma doença humana bastante antiga, mortes associadas a febres periódicas e esplenomegalia foram mencionadas em escritos egípcios e chineses em 2700 a.C. Estima-se que a doença chegou a Roma em 200 a.C., se espalhou por toda a Europa durante o século XII, e chegou a Inglaterra por volta do século XIV (Garcia, 2010). A malária é uma doença parasitária causada por protozoários do gênero *Plasmodium*. Existem várias espécies dentro desse gênero causadoras da doença em hospedeiros vertebrados, dentre as quais quatro são as principais causadoras de malária em humanos: *Plasmodium falciparum, Plasmodium vivax, Plasmodium ovale* e *Plasmodium malarie*. Uma quinta espécie, *Plasmodium knowlesi*, originalmente descrita como parasito de macacos de cauda longa, também pode infectar humanos em algumas áreas como a Malásia (Duval *et al.*, 2010). Dentre estas espécies, *P. falciparum* é a mais virulenta e responsável pela grande maioria das mortes. Esta espécie se distingue pela capacidade de se ligar ao endotélio durante a fase sanguínea da infecção e de serem seqüestradas em órgãos, incluindo o cérebro. *P. vivax* é um grave problema de saúde pública, pois possui uma ampla distribuição mundial e está associado a uma alta taxa de morbidade global (Greenwood *et al.*, 2008). Além dos fatores sociais, geográficos, políticos e econômicos que contribuem para a catástrofe global causada pela malária, a própria biologia do parasito já é um dos fatores que mantém esta doença parasitária no topo das listas de prioridades de saúde pública em muitos lugares do planeta. A habilidade do *P. vivax* e também do *P. ovale* em se manterem dormentes por meses como hipnozoítos no fígado torna a infecção por essas espécies mais difícil de ser erradicada.

De acordo com a OMS (2011), a malária mata em torno de um milhão de pessoas por ano e mais de 3,2 bilhões de pessoas vivem em 107 países ou territórios em risco de infecção por esta doença. Estima-se que a doença afeta cerca de 300-500 milhões de pessoas nas áreas tropicais e subtropicais do planeta, resultando em cerca de 800.000 mortes a cada ano, na grande maioria, crianças com menos de cinco anos na África sub-Sahariana.

Plasmodium vivax, a espécie mais amplamente distribuída no planeta, é um dos mais importantes desafios para a Saúde Pública nas Américas do Sul e Central, Ásia Central, Sul e Sudeste Asiáticos, Oceania, Oriente Médio e África Oriental, onde 2,85 bilhões de pessoas vivem atualmente sob risco de infecção e 70-80 milhões de casos clínicos são relatados anualmente. Os recentes indícios de surgimento de cepas resistentes às drogas e de formas graves (às vezes fatais) da doença desafiam a visão tradicional da malária vivax como uma infecção benigna. A agenda de pesquisa para erradicação da malária (*Malaria Eradication Research Agenda, MalERA*) coloca o *P. vivax* no topo da lista de prioridades. Fora da África esse parasito é responsável por mais de 50% de todos os casos de malária com a maioria 80-90% ocorrendo na Ásia, Oeste do Pacífico e Leste da Ásia e 10-20% na América do Sul e Central. Normalmente é considerada uma doença benigna, entretanto existem casos de aumento de severidade clínica e resistência a antimaláricos na América do Sul e Sudeste Asiático, incluindo mortes associadas exclusivamentes a esta espécie (Mendis *et al.*, 2001; Oliveira-ferreira *et al.*, 2010; Price *et al.*, 2007).

No Brasil, de acordo com dados da SVS (2011) cerca de 49 milhões de pessoas vivem atualmente em áreas onde existe o risco de transmissão da doença, sendo que 99,8% dos casos de malária se concentram na Amazônia Legal, que inclui os estados da região Norte além do Maranhão e Mato Grosso.

Diferentes estratégias de combate à doença têm sido propostas, visando à interrupção de sua transmissão. Entre elas destacam-se o Programa de Erradicação da Malária, proposto em 1955 pela Organização Mundial da Saúde (OMS), centrado principalmente em ações verticais, incluindo a borrifação de paredes com inseticida de ação residual (DDT) e o tratamento em massa utilizando-se um antimalárico de baixa toxicidade (cloroquina). Apesar dessa estratégia mundial voltada para o combate da doença, ter sido bem-sucedida em vários países, apresentou efeitos limitados em algumas regiões da África, Ásia e América do Sul, incluindo a Amazônia brasileira. No Brasil, a incidência anual de infecção por *P. falciparum* (espécie do parasito de malária predominante entre 1985 e 1990) diminuiu de forma constante durante a década de 1990, enquanto a de *P. vivax* manteve uma tendência ascendente, mostrando que as

estratégias de controle adotadas para esta espécie podem não ser as mais adequadas. Embora tanto *P. falciparum* quanto *P. vivax* ainda sejam transmitidos em toda a bacia Amazônica, com raras infecções por *P. malariae*, atualmente *P. vivax* é responsável por cerca de 85% dos 315.000 casos de malária relatados neste país por ano (Oliveira-ferreira *et al.*, 2010), sugerindo que esta espécie deve ser menos suscetível às estratégias de controle da malária atualmente aplicadas no Brasil.

Adicionalmente, a resistência de parasitos aos antimaláricos e as limitações do uso de inseticidas, associadas às questões político-econômicas mundiais, desencadearam um agravamento da situação epidemiológica da malária nas três ultimas décadas. Felizmente o reconhecimento dessas questões fez com que a estratégia de enfrentamento do problema fosse modificada ao longo dos anos. Atualmente a OMS possui um plano de ação denominado Estratégia Global da Malária, o qual prioriza a integração das atividades de controle dos serviços gerais da saúde, reconhecendo os pontos específicos de cada situação a ser enfrentada. O Brasil conta com o Programa Nacional de Controle da Malária (PNCM), que tem como principal objetivo reduzir a morbimortalidade por malária com estratégias de intervenção de forma integrada. A principal destas estratégias é a realização de diagnóstico rápido e tratamentos adequados e oportunos totalmente financiados pelo governo.

1.2 Ciclo de vida do parasito

Os parasitos causadores da malária apresentam um ciclo de vida heteroxênico, ou seja necessitam, obrigatoriamente, de dois hospedeiros, um vertebrado e um invertebrado, sendo o último o hospedeiro definitivo, onde a fase sexual acontece. No caso da malária humana, os hospedeiros invertebrados são exclusivamente fêmeas de mosquitos anofelinos (**Figura 1**).

Durante o repasto sanguíneo as fêmeas infectadas transferem, através da saliva, dezenas de esporozoítos para a epiderme do hospedeiro vertebrado os quais, segundo Mota e colaboradores (Mota and Rodriguez, 2001), são capazes de atravessar várias células do hospedeiro mantendo-se íntegros. Após transpor a barreira epitelial, uma parte desses esporozoítos alcança a corrente sanguínea enquanto outros

são drenados para vasos linfáticos (Amino *et al.*, 2006). Uma vez dentro do sistema circulatório sanguíneo os esporozoítos atingem o fígado, infectando os hepatócitos.

O processo de invasão do fígado é complexo, e depende de várias interações parasito-célula hospedeira. Os esporozoítos deslizam até encontrarem uma célula de Kupfer, a qual invade e atravessa dentro de um vacúolo não fusiogênico até atingirem o espaço de Disse (Frevert *et al.*, 2005). Uma vez dentro do parênquima hepático, os esporozoítos invadem e atravessam múltiplas células hepáticas até alcançarem um hepatócito final, onde se desenvolvem, dentro de um vacúolo parasitóforo, em formas replicativas exo-eritrocíticas (EEFs) (Mota and Rodriguez, 2001). A passagem transversal por vários hepatócitos parece ser um processo obrigatório para tornar os esporozoítos capazes de estabelecer a infecção em um hepatócito-alvo, culminando com a formação de um vacúolo no qual os parasitos irão se replicar e desenvolver(Mota, Hafalla and Rodriguez, 2002).

O estabelecimento da infecção hepática dá origem a esquizontes multinucleados que, quando maduros, se diferenciam em milhares de merozoítos, as formas invasivas que parasitam as hemácias circulantes. Os esquizontes liberam, nos sinusóides hepáticos, vesículas provenientes dos hepatócitos infectados contendo até milhares de merozoítos. Estas vesículas foram denominadas merossomos e seriam um mecanismo de defesa para evitar a fagocitose de merozoítos assim que deixassem o fígado, aumentando assim as chances de invasão de eritrócitos e de sobrevivências dos parasitos (Sturm *et al.*, 2006). Existem duas espécies que são capazes de desenvolver formas latentes que podem ficar longos períodos de tempo neste órgão. Trata-se de *P. vivax* e *P. ovale*. Estas formas dormentes tem a capacidade de permanecerem inativas e se ativarem, restabelecendo a infecção em pacientes que tiveram cura terapêutica, mesmo não tendo visitado uma área onde ocorre transmissão da doença. A fase de infecção assintomática no fígado dura cerca de 6 dias, com cada esporozoíto produzindo dezenas de milhares merozoítos que invadem e se desenvolvem dentro dos eritrócitos. Os estágios sanguíneos da infecção incluem formas assexuadas do parasito submetidas a ciclos repetitivos de multiplicação, bem como do sexo masculino e formas sexuais femininos, chamados gametócitos, que aguardam pela ingestão por

mosquitos antes de se desenvolver ainda mais. Parasitos na fase sanguínea assexuada produzem de 8-20 novos merozoítos a cada 48 horas (ou 72 horas para *P. malariae*). Os parasitos, chamados agora de anéis (trofozoítos jovens), começam a se desenvolver no interior das hemácias. Os anéis se desenvolvem em trofozoítos maduros e estes dão origem a esquizontes através de múltiplas divisões nucleares. Novos merozoítos são formados no interior dos esquizontes e são liberados pela ruptura destes, para darem início a uma nova rodada de invasões de eritrócitos, fechando o ciclo de reproduções assexuadas eritrocíticas característico de *Plasmodium* spp. (Miller *et al.*, 2002).

Alguns merozoítos, após invadirem as células sanguíneas, ao invés de transformarem-se em esquizontes e perpetuarem o ciclo de reprodução assexual, passam por um desenvolvimento diferencial que resulta na formação de células sexuais especializadas na transição entre o hospedeiro vertebrado e o invertebrado (Talman *et al.*, 2004). Estas células são chamadas gametócitos e existem dois tipos, os microgametócitos e os macrogametócitos. Sua maturação é descrita em 5 estágios de desenvolvimento (Field e Shute, 1956). São os gametócitos maduros, ou do quinto estágio, que são transferidos no momento do repasto sanguíneo para as fêmeas dos mosquitos, iniciando então o ciclo sexual dos plasmódios.

Aproximadamente nos próximos trinta minutos ocorre a fecundação dos gametas e a formação do zigoto diplóide. Durante as próximas horas, o zigoto, uma célula de morfologia esférica, irá se transformar em um oocineto, uma célula alongada, com diferenciação antero-posterior e móvel. Esta metamorfose é amparada por uma reorganização do citoplasma, com o aparecimento de um complexo pelicular sob a membrana plasmática, a estruturação de um característico citoesqueleto e a formação de estruturas apicais típicas de formas invasivas (Sinden, 1999). Este atravessa a matriz peritrófica e atinge a parede do intestino médio, onde se encista na camada epitelial do órgão, passando a ser chamado de oocisto. Inicia-se então o processo de divisão esporogônica que, após o período de alguns dias, ocorre a ruptura da parede do oocisto, liberando, assim, os esporozoítos (Barillas-Mury and Kumar, 2005). Estes são disseminados por todo o corpo do inseto através da hemolinfa, até atingirem as

células das glândulas salivares, onde são inoculados no hospedeiro vertebrado no momento do repasto sanguíneo do vetor. Na fase sexual os parasitos não são patogênicos, mas são transmissíveis ao vetor *Anopheles*, onde se recombinam durante um breve período de fase diplóide e geram formas geneticamente distintas, os esporozoítos. O mosquito se torna apto a infectar o sangue do hospedeiro aproximadamente duas semanas após a ingestão de gametócitos, um período de tempo que é influenciada pela temperatura externa. O desenvolvimento de *P. vivax* pode ocorrer dentro do mosquito em uma menor temperatura ambiente do que a exigida para o desenvolvimento de *P. falciparum,* explicando a predominância de infecções por *P. vivax* fora regiões tropicais e subtropicais (Greenwood *et al.*, 2008). A fase patogênica é a assexuada eritrocítica e indivíduos infectados podem apresentar-se com diversas sequelas que afetam diferentes órgãos.

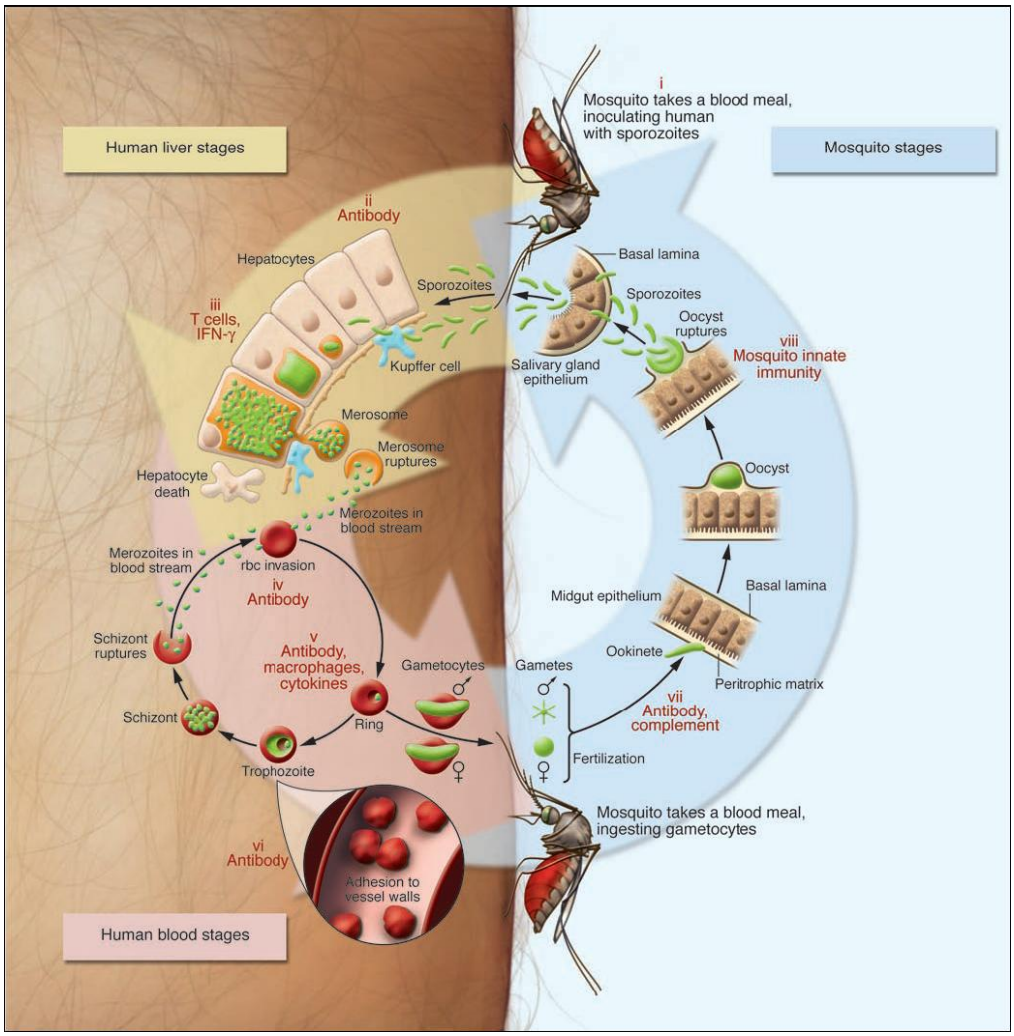

Figura 1: Ciclo de vida do *Plasmodium*, parasito causador da malária. Figura de Greenwood *et al*. 2008.

1.3 Sistema de transporte e tráfego de proteínas

Uma vez dentro da célula hospedeira, os parasitos agora precisam de estratégias para conseguir recursos nutricionais para seu desenvolvimento e multiplicação. Os parasitos da malária infectam eritrócitos que em grande parte não possuem as vias de biossíntese, são metabolicamente lentos e desprovidos de compartimentos intracelulares (Dhangadamajhi, Kar and Ranjit, 2010). Devido a esses fatores, muitas mudanças acontecem desde o momento da invasão, até o momento da ruptura do eritrócito, quando novos merozoítos são liberados na corrente sanguínea e

estão prontos para invadir novas células e dar continuidade à infecção. A partir do momento em que o parasito invade a célula, ele promove alterações drásticas nas características físicas da mesma, por exemplo, estabelece um sistema de membranas elaborado no citosol do eritrócito (Hanssen *et al.*, 2008) e altera a sua superfície de tal forma que surgem saliências elétron-densas como botões (Crabb *et al.*, 1997). Com essas mudanças físicas surgem novas propriedades biológicas, muitas das quais são patogênicos na natureza, como o aumento da rigidez, capacidade de aderir às paredes vasculares de células endoteliais e uma capacidade de variar o revestimento antigênico do eritrócito infectado para evitar anticorpos (Crabb, Koning-Ward, de and Gilson, 2010).

Dentro do vacúolo parasitóforo (VP), o parasito precisa criar um sistema de transporte eficiente, garantindo que as proteínas irão ultrapassar as múltiplas membranas até encontrar o citosol da célula hospedeira. Esse fato tem extrema importância para os parasitos da malária em mamíferos, uma vez que estes são parasitos intracelulares e necessitam invadir uma célula vermelha do sangue (RBC) para conseguirem sobreviver no hospedeiro. As RBC de mamíferos são anucleadas e não possuem um sistema de secreção, o que faz com que os parasitos necessitem de um sistema de tráfego de proteínas próprio. Assim, os parasitos apresentam diferentes vias de tráfego de proteínas: (i) Através do retículo endoplasmático (ER) e complexo de golgi (G), utilizada para a maioria das proteínas secretadas; (ii) Através das organelas chamadas Maurer's cleft, que exporta para o citoplasma da célula hospedeira através do vacúolo parasitóforo; (iii) Vias alternativas independentes de ER-G, casos específicos (Lingelbach, 1993; Nacer *et al.*, 2001). Apesar de muitos trabalhos importantes contribuírem para o entendimento do sistema de transporte de proteínas em *Plasmodium*, algumas questões ainda necessitam de ser elucidadas. Por apresentarem compartimentos celulares incomuns: organelas secretoras como as roptrias, os micronemas e os grânulos densos; um vacúolo digestivo e o apicoplasto (Figura 2); os parasitos da malária apresentam formas atípicas de transporte e exportação de proteínas, impulsionados principalmente pelo ambiente em que este parasito tem que sobreviver e se desenvolver.

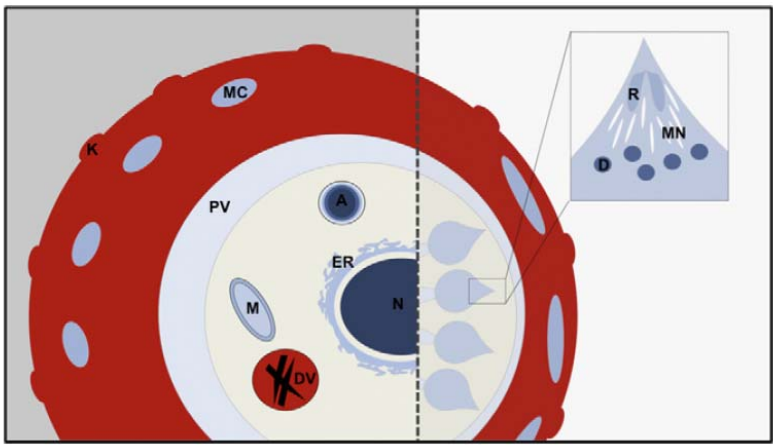

Figura 2: Destinos para o tráfego de proteínas em um eritrócito infectado por *Plasmodium*. O lado esquerdo representa um trofozoíto e o lado direito um esquizonte. Apicoplasto (A), Maurer´s Cleft (MC), Knobs (K), Vacúolo parasitóforo (PV), Mitocôndria (M), Vacúolo digestivo (DV), Núcleo (N), Retículo endoplasmático (ER) do lado esquerdo e do lado direito Roptrias (R), Micronemas (MN) e Grânulos densos (D). Fonte: De Ponte et al. 2012.

Para a grande maioria dos tipos de células eucarióticas, a entrada para a via secretora é determinada pela presença de uma sequência na extremidade N-terminal da pré-proteína, conhecida como peptídeos sinal (PS). Estas sequências, que geralmente são hidrofóbicas e possuem de 15 a 30 aminoácidos, emergem do ribossomo e são ligados à partícula de reconhecimento de sinal, que interrompe a tradução da proteína e dirige a cadeia polipeptídica nascente para o retículo endoplasmático (RE). Após a ligação da partícula ao seu receptor na membrana do RE, a tradução de proteínas continua e a proteína madura pode entrar no lúmen ER, ou na membrana do RE. Em muitos casos, o PS é depois clivado a partir da pré-proteína por uma peptidase sinal. O sistema de secreção de proteínas em *Plasmodium* parece ser na maioria das vezes parecido com o dos outros tipos de células eucarióticas. Porém existem algumas diferenças que já foram observadas. Uma diz respeito à localização do parasito. Por se encontrar dentro do vacúolo parasitóforo, o destino final da proteína ou é o lúmen do próprio VP, ou ainda o citoplasma ou a membrana da célula hospedeira. Outra característica que também foi observada é o fato de que muitas proteínas secretadas pelo parasita contêm PS incomuns (n-terminais atípicos ou C-terminais)(Nacer *et al.*, 2001). Alguns PS apesar de possuírem as características hidrofóbicas comuns elas geralmente estão localizadas um pouco recuadas em relação

a esta região, entre 15 e 50 aminoácidos. O motivo pelo qual estes parasitos possuem diferenças neste tipo de transporte ainda necessita de esclarecimentos, porém muitos esforços têm sido investidos neste sentido (Hiss *et al.*, 2008; Lingelbach and Przyborski, 2006).

Além destas peculiaridades em relação ao peptídeo sinal de algumas proteínas de *Plasmodium*, foi descoberto um sistema de transporte de proteínas que até agora se mostrou ser exclusivo de espécies deste gênero. Dois estudos independentes identificaram a presença de um motivo de cinco aminoácidos, que estaria ligado à exportação de algumas proteínas e por isso recebeu o nome de PEXEL (*Plasmodium EXport ELemet*) ou VTS (*Vacuolar Transport Signal*) (Hiller *et al.*, 2004; Marti *et al.*, 2004). O motivo foi identificado utilizando uma análise bioinformática dos alinhamentos da região N-terminal de proteínas que eram já conhecidas como sendo exportadas do VP para o eritrócito. O consenso do motivo é R/QxLxE/Q/D, onde x pode ser qualquer aminoácido na posição onde ele ocorre, e ele está localizado aproximadamente 20 aminoácidos *downstream* do PS e é funcional tanto em proteínas solúveis quanto de membrana (Haase and Koning-Ward, de, 2010). Levando-se em conta o fato de que a distinção entre as proteínas exportadas e aquelas que não serão ocorre na membrana do vacúolo parasitóforo (MVP), a maioria dos motivos PEXEL foram descritos como sendo removidos no RE do parasito. Isto ocorre através da clivagem do motivo entre a posição 3 e 4 pela protease plasmepsina V (**Figura 3**) (Chang *et al.*, 2008; Deponte *et al.*, 2012). A identificação deste motivo permitiu a utilização de ferramentas que diminuem o tempo para encontrar novas proteínas que compartilham a mesma via de exportação e alguns esforços tem sido feitos para identificar e caracterizar estas proteínas (Hiss *et al.*, 2008; Ooij, van *et al.*, 2008).

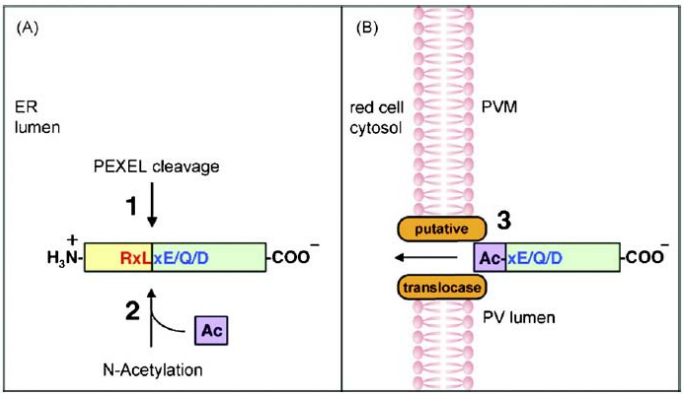

Figura 3: Modelo mostrando o motivo PEXEL (A) e sua via de processamento (B). Fonte: Chang et al. 2008.

1.4 Tratamento: Vacinas e drogas

A malária é uma doença extremamente complexa e diferentes aspectos relacionados principalmente à biologia do parasito, tornam o controle e erradicação dessa doença um grande desafio para a saúde pública. A OMS preconiza o diagnóstico rápido e o tratamento dos doentes como as principais formas de atuação para o controle da doença. Existem também medidas auxiliares de controle como o uso de mosquiteiros impregnados com inseticidas. Existem algumas limitações com relação ao diagnóstico e tratamento que tornam a tarefa ainda mais complicada. Para o diagnóstico, é utilizada a gota espessa visualizada pelo microscópio, o que requer um profissional de qualidade treinado, para que não aconteçam diagnósticos inadequados, pois é a partir dele que será feita a decisão de qual medicamente o paciente deverá tomar. No caso do uso de antimaláricos um problema que tem sido relatado é a seleção de cepas resistentes às principais drogas disponíveis para o tratamento da malária.

Para muitos especialistas, o controle e erradicação da malária passam diretamente pelo desenvolvimento de uma vacina (Arévalo-herrera, Chitnis and Herrera, 2010; Schwartz *et al.*, 2012). Uma vacina eficaz contra a malária é uma prioridade e muitos esforços têm sido feitos para identificar candidatos promissores. Para muitas autoridades de saúde pública, a vacina é um componente chave para o sucesso e manutenção de um programa eficaz de controle da malária. Existe

atualmente um antígeno de *P. falciparum* em teste na África. Recentemente foram publicados os resultados da fase III dessa vacina e ela conseguiu garantir uma proteção de mais de 50% nas crianças que foram vacinadas (Asante *et al.*, 2011). Apesar destes esforços, em relação ao *P. vivax* poucas proteínas estão sendo testadas como candidatos à vacina.

1.5 Uso de ferramentas de bioinformática para busca de novos alvos terapêuticos

Com o advento do seqüenciamento de DNA em larga escala, principalmente utilizando tecnologias de sequenciamento de última geração (*Next generation sequencing* – NGS), um número cada vez maior de dados biológicos se encontra disponível para ser explorado para responder diferentes questões biológicas. Com a disponibilidade destes dados, surgiu a possibilidade de se empregar metodologias baseadas na análise de sequências para o estudo de diferentes organismos. Para os parasitos do gênero *Plasmodium* existe um banco de dados público, o PlasmoDB (Aurrecoechea *et al.*, 2009), onde é possível obter qualquer sequência biológica que está atualmente disponível para este gênero. Juntamente com este aumento do número de dados, surge também a necessidade de buscar novas metodologias para lidar com este novo cenário, acessando as informações necessárias. Hoje em dia é extremamente comum o emprego de ferramentas de bioinformática para inúmeras tarefas relacionadas à mineração, manipulação e processamento de sequências biológicas. O uso destas ferramentas pode muitas vezes reduzir o tempo para ensaios biológicos e também eliminar moléculas que não seriam bons candidatos para determinado tipo de estudo. No caso de busca por novas drogas e vacinas, o uso desta abordagem pode aumentar o número de candidatos na fase pré-clínica, aumentando as chances de se encontrar um candidato promissor que passará para as próximas etapas de testes.

Para ser um bom candidato à vacina, uma proteína deve ter determinadas características que vão garantir que esta esteja acessível para o sistema imune do hospedeiro e que ele seja capaz de induzir uma resposta duradoura contra este

antígeno. Uma característica importante que um candidato a vacina deve possuir é estar acessível ao sistema imune do hospedeiro, e no caso de parasitos intracelulares, como no caso dos plasmódios, a proteína necessita ser exportada ou secretada. Proteínas excretadas/secretadas (ESPs) estão localizadas ou são liberadas da superfície celular, fazendo com que elas estejam acessíveis para o sistema imune ou para alguma droga. Pensando nestas características é necessário selecionar programas que irão procurar nas sequências de proteínas do parasito, marcadores que irão conferir o status de ESP. Para isto a proteína precisa ter primeiramente um peptídeo sinal. No caso dos plasmódios, existe o motivo PEXEL, já discutido anteriormente como sendo essencial para o transporte de determinadas proteínas nestes organismos. Com o objetivo de procurar por quais destas ESPs do parasito serão expostas na membrana da célula hospedeira, é interessante que esta proteína possua também um domínio transmembrana que permitirá a ela se ancorar na membrana da célula onde reside.

O uso de ferramentas de bioinformática requer um conhecimento mínimo para manipulação de dados, principalmente em análises em larga escala como no caso de genomas completos. Muitos programas precisam ser rodados em linha de comando, têm formatos de arquivos de entrada específicos e muitos parâmetros que podem ser alterados de acordo com o tipo de análise que se deseja fazer. Muitas equipes não contam com alguma pessoa com este tipo de conhecimento e muitas vezes análises interessantes, que poderiam reduzir o tempo e os custos da pesquisa, deixam de ser feitas. Pensando nisso, alguns esforços têm sido feitos para tentar melhorar a comunicação entre os pesquisadores e os programas disponíveis, que são desenvolvidos por pessoas da área da computação. É comum o uso do termo pipeline, que seria o uso de diferentes programas em sequência, sem que seja necessário alterar nenhum arquivo de entrada e saída de nenhum dos programas que compõe esta sequência. Dessa forma, os não-especialistas em bioinformática, poderiam ter maior acesso ao emprego de múltiplas ferramentas ao mesmo tempo, sem ter a necessidade de conhecer profundamente cada uma delas. Alguns esforços têm sido feitos no sentido de tentar melhorar esta situação. Uma das tentativas de melhorar a

interface dos usuários não familiarizados com os programas de bioinformática é a plataforma Galaxy (Giardine *et al.*, 2005). Trata-se de uma plataforma para integração de diferentes programas para manipulação e análise de sequências biológicas. O Galaxy pode ser baixado e sua versão básica conta com diversas ferramentas para trabalhar inclusive com dados oriundos de seqüenciamento de nova geração. A interface funciona no navegador e todos os dados e todas as ações ficam armazenadas em forma de histórico. Isto permite ao usuário fazer diferentes buscas e combinar os diferentes resultados destas buscas de forma simples e rápida. Operações como interseção, união, subtração, dentre outras, podem ser feitas utilizando esta interface. O programa também permite que novas ferramentas sejam adicionadas à versão básica, de acordo com a necessidade do usuário. Para esta etapa é necessário um conhecimento sobre programação e linguagem web, porém uma vez inserida a ferramenta fica disponível para todos os usuários utilizarem. O galaxy permite a integração de diferentes ferramentas. É possível criar um usuário e a partir desta ação a opção de criar um *workflow* fica disponível. Um *workflow* consiste em uma sequência de diferentes ferramentas que são adicionadas como caixas, juntamente com setas, que ligam estas caixas, de uma forma que uma sequência de passos é seguida e as análises passam então a ser feitas automaticamente. Esta possibilidade é interessante, principalmente quando se deseja fazer análises comparativas em diferentes espécies ou organismos, sem que seja necessário repetir todos os passos manualmente. Basta carregar os novos dados e com um clique todos os programas irão rodar automaticamente.

2. JUSTIFICATIVA

A malária é a doença parasitária de maior impacto no mundo. Dois fatores são determinantes para esse fato. A distribuição da doença em todo mundo, onde mais de 3 bilhões de pessoas vivem em áreas onde existe o risco de se contrair a doença, o que representa cerca de 43% da população do planeta; e o grande número de mortes associados, chegando a quase 1 milhão por ano principalmente em crianças com menos de cinco anos no continente africano. Diante desse impacto, a doença tem se mostrado um dos grandes desafios da saúde pública mundial e muitos esforços tem sido feitos na tentativa de controlar e algumas vezes até erradicar a doença. Apesar dos avanços conseguidos ao longo destes anos de pesquisa, uma vacina eficaz contra estes parasitos ainda não está disponível. Conseguir uma vacina eficaz, principalmente contra parasitos tão complexos quanto os causadores da malária não é uma tarefa fácil demandando muito tempo e dinheiro. Qualquer ferramenta que possa ser utilizada para acelerar alguma parte do processo de obtenção de uma vacina pode ser de extrema utilidade e ter importância crucial na diminuição do tempo para se chegar a uma vacina eficaz. Assim, o desenvolvimento de interfaces de comunicação entre os programas de bioinfomática mais amigáveis e de fácil manipulação articulando diferentes programas é de grande utilidade na busca de novos candidatos vacinais e alvos de drogas.

O principal foco da pesquisa da busca por antígenos em malária é o *P. falciparum*, porém cada vez mais grupos tem concordado que a eliminação e erradicação da malária passa pelo controle efetivo também do *P. vivax*. Devido a estes fatores e também por ser a principal espécie causadora da malária no Brasil, um esforço maior é necessário na elaboração de metodologias que permitam aumentar o número de possíveis antígenos que podem futuramente vir a ser um componente de uma vacina que irá ajudar a eliminar esta doença que causa tantas perdas para a humanidade.

3 OBJETIVOS

3.1 Objetivo geral

Analisar características antigênicas importantes em *Plasmodium* e desenvolver um pipeline para identificação de diferentes alvos para vacinas contra parasitos deste gênero.

3.2 Objetivos específicos

Analisar o preditor de peptídeo sinal SignalP.

Estudar os peptídeos sinal preditos em diferentes espécies de Plasmodium;

Avaliar a presença, conservação e localização do motivo PEXEL nas diferentes espécies de *Plasmodium*.

Avaliar a predição de segmentos transmembrana em proteínas de *Plasmodium* spp.

Avaliar a predição de epitopos através de diferentes programas.

Integrar diferentes programas de bioinformática numa plataforma automática.

Identificar novos alvos com potencial para serem avaliados como candidatos à vacina.

4. MATERIAIS E MÉTODOS

4.1 Bancos de dados

Para procurar pelos candidatos promissores, recorremos à base de dados de *Plasmodium*, o PlasmoDB v7.1 (http://www.plasmodb.org/). Utilizando a ferramenta de busca de genes por organismos foram selecionadas todas as sequências preditas de proteínas para cinco espécies: *P. vivax*, *P. falciparum*, *P. knowlesi, P. yoelli* e *P. berghei*. Apesar de possuir sequências depositadas neste repositório, sequências de *P. chabaudi* foram excluídas da análise, pois os resultados foram discrepantes em relação às outras espécies. Infelizmente não foi possível utilizar esta espécie, porém a partir do momento em que os dados do banco sejam corrigidos, incluí-la no pipeline não seria nenhum problema e não existem razões para acreditar que os resultados do trabalho não sejam válidos. A partir destas sequências protéicas, diferentes programas foram utilizados para buscar as características desejadas. Foi construído um banco de dados local utilizado MySQL v5.1 para armazenar e acessar as sequências. Os resultados de cada análise foram carregados em uma tabela do banco para que posteriormente, uma análise de combinação dos resultados pudesse ser feita.

4.2 Predição de peptídeo sinal

Para buscar pelas proteínas que potencialmente poderiam estar acessíveis ao sistema imune do hospedeiro, diferentes programas foram selecionados. Para identificar as proteínas que poderiam ser exportadas/secretadas ou de membrana utilizamos o programa SignalP v3.0 (Bendtsen *et al.*, 2004), o preditor de peptídeo sinal mais amplamente utilizado. O SignalP utiliza duas diferentes abordagem para identificar um peptídeo sinal. Ele emprega tecnologias baseadas em aprendizagem de máquinas sendo uma rede neural (*Neural Networks)* e também um modelo oculto de Markov (*Hidden Markov Model*), que consegue discriminar as proteínas cujo peptídeo sinal não é clivado e permanecem ancoradas na membrana. O arquivo de entrada do programa pode ser um conjunto ou uma única sequência protéica. Após a análise do algoritmo, diferentes valores de scores são dados para cada sequência analisada. Dentre estes parâmetros, o *D-score* (escore de discriminação) mostrou ser o mais

indicado para analisar. Este score é obtido através de operações com os outros scores: *Y-score* e *S-score*. O primeiro é o score combinado do sítio de clivagem e o outro é o score de peptídeo sinal. O *D-score* é a média ponderada da média do *S-score* e do *Y-score* máximo. Este é o resultado de que é usado para discriminar peptidos sinal de não péptidos sinal. O seu valor varia de 0 a 1, sendo que o CUT-off padrão do programa é 0,43 para ser considerada como positiva. Todos os parâmetros foram utilizados com os valores padrão. Para cada espécie foi gerado um arquivo com os resultados que foram carregados em uma tabela do banco de dados local.

4.3 Busca pelo motivo PEXEL

Com a descoberta do motivo PEXEL, foi possível utilizar uma estratégia de bioinformática para procurar por este consenso em sequências de proteínas. Para esta busca, foi desenvolvido um script em Perl (ver lista em anexo) que procurava pelo motivo PEXEL (R/KxLxE/Q/D) nas sequências de proteínas no formato fasta. O resultado gerado retorna o identificador da sequência seguido pelo motivo encontrado, a posição e o número de vezes que este motivo foi identificado em cada sequência. Esta lista foi gerada e armazenada em uma tabela do banco de dados local.

Para a análise de conservação do motivo PEXEL, alinhamentos dos motivos encontrados nas proteínas foram feitos utilizando o programa clustalw2. O alinhamento dos motivos foi submetido ao programa weblogo3.1 (Crooks *et al.*, 2004) para geração dos gráficos para representar os possíveis padrões de conservação nos alinhamentos.

4.4 Busca por regiões transmembrana

Para procurar proteínas que poderiam ser integrais de membrana foi utilizado o TMHMM v2.0 (Krogh *et al.*, 2001), que busca por domínios transmembrana em sequências de proteínas. O programa utiliza um algoritmo baseado em um Modelo Oculto de Markov para predizer a localização de uma possível região transmembrana além da topologia das estruturas secundárias das proteínas, indicando qual região está voltada para o interior ou exterior da célula. O arquivo de entrada do programa é uma ou um conjunto de sequências de proteínas. Não existem parâmetros para serem

ajustados no programa. O resultado é uma com todos os identificadores, seguido do tamanho da sequência, número esperado de aminoácidos em hélices transmembrana, o número esperado de aminoácidos em hélices nas primeiras 60 posições, o número de hélices preditas e a topologia da hélice, indicando quais regiões estão dentro ou fora. A lista com os identificadores das proteínas positivas para esta predição foi carregada em uma tabela do banco de dados local.

4.5 Análise combinatória dos resultados

Para comparar os resultados das predições e buscar pelas proteínas que eram positivas para as predições de PS, motivo PEXEL e domínio transmembrana foi utilizada a ferramenta *select* do MySQL. Dessa forma no final uma lista com os identificadores foi gerada. A partir desta lista, as sequências das proteínas de cada identificador foram recuperadas. Estas sequências foram utilizadas para a predição de presença de epitopos de células B.

4.6 Seleção das proteínas e Predição de Epitopos

Após cruzar os resultados das predições, as proteínas que tiveram predição positiva para as diferentes características foram selecionadas. Antes de procurar pelos possíveis epitopos nas sequências foi feita uma investigação no PlasmoDB, para saber se alguma dessas proteínas já tinham alguma informação relacionada à presença de epitopos. Estas informações foram acessadas de outro banco de dados, o IEDB, que fornece uma variedade de epitopos de células B e T experimentalmente caracterizados, além de dados de experimentos de ligação ao MHC. No próprio PlasmoDB existe o peptídeo para a determinada proteína que já tenha sido experimentalmente validado e depositado no banco. Aquelas que já tinham alguma informação foram excluídas da nossa análise. Além disso, foram priorizadas proteínas hipotéticas, excluindo assim proteínas que pudessem ter evidências indiretas de ser um antígeno. Após esta etapa, dez proteínas de *P. vivax* foram selecionadas para serem submetidas aos programas de predição de epitopos. Esta espécie foi selecionada, por ser a principal espécie que causa a malária em humanos no Brasil, e também porque o laboratório no qual este projeto foi desenvolvido, contar com soros de pacientes da área endêmica infectados

por esta espécie, o que torna possível um ensaio de validação. Foram utilizados apenas preditores de célula B, devido à facilidade do teste de validação. Para a predição de epitopos de células B, foram utilizados três diferentes programas: Bepipred (Larsen, Lund and Nielsen, 2006), que combina duas diferentes metodologias para predizer epitopos lineares. Este programa utiliza as predições de um HMM com escala de propensão proposta anteriormente por Parker et al. (Parker, Guo and Hodges, 1986); BCPREDS, que emprega uma metodologia de aprendizado de máquina para a predição. O programa utiliza um classificador, uma Máquina de Suporte de Vetor (*Support Vector Machine*), utilizando *strings* como função de kernel (El-Manzalawy, Dobbs and Honavar, 2008); e uma ferramenta de predição do IEDB (Vita *et al.*, 2010) através do método de Kolaskar & Tongaonkar (Kolaskar and Tongaonkar, 1990), que é uma metodologia semi-empírica, que utiliza dados das propriedades físico-químicas dos resíduos juntamente com as frequências em que eles ocorrem em epitopos experimentalmente conhecidos. Os epitopos selecionados foram aqueles que se encontravam em uma região predita pelas três abordagens e tinham entre 17 e 25 aminoácidos. Além do grupo de proteínas preditas como potenciais antígenos, para validação dos resultados de predição foram selecionados outros dois grupos controle de proteínas: positivo e negativo. O grupo controle positivo contém proteínas que são conhecidas como sendo imunogênicas de trabalhos anteriores cujos epitopos já foram identificados. O grupo controle negativo conta com proteínas que foram negativas para todas as predições feitas pelos programas (PS, PEXEL e TMHMM). As três listas de proteínas foram utilizadas na predição de epitopos, sendo que para o grupo teste os epitopos foram identificados pelos 3 programas diferentes. Algumas proteínas utilizadas na construção da lista de controles positivos, foi feita a investigação a respeito de epitopos depositados no IEDB. O próprio PlasmoDB tem esta informação para cada proteína. Foi feita a verificação e para aquelas que já tinham epitopos depositados, estes foram utilizados, caso contrário a proteína era submetida à análise do preditor do IEDB.

Uma análise quanto ao número de epitopos preditos pelos métodos foi feita para as proteínas selecionadas pelo pipeline. As proteínas selecionadas e submetidas aos

programas foram investigadas para saber o número de epitopos por proteína para cada método procurou-se também por epitopos comuns a 2 e comum aos 3 métodos utilizados. Espera-se que com esta abordagem, a probabilidade da região realmente ser um epitopo seja maior, uma vez que os 3 programas concordaram na predição.

4.7 Desenvolvimento do pipeline

Para tornar a combinação das análises mais familiar aos não especialistas em bioinformática foi utilizada a plataforma Galaxy (Giardine *et al.*, 2005) para criar um pipeline que pudesse automatizar a seleção das sequências e submissão pra análise nos diferentes programas utilizados. Para isto, os programas foram adicionados à versão básica. Para adicionar uma ferramenta são necessários dois diferentes tipos de arquivos: um arquivo que irá chamar a linha de comando do programa, chamado de wraper e outro arquivo que será carregado no galaxy ao iniciá-lo, que tem formato XML. A partir daí, foi construído um workflow (seleção de diferentes passos para serem automatizados) onde cada caixa corresponde a um programa. A parte das análises utilizando o banco foi substituída por diferentes programas usando a ferramenta *Text manipulation tools* (que já vem instalada na versão básica) onde você pode selecionar IDs comuns a partir de duas ou mais listas de identificadores. Dessa forma foi possível fazer as combinações dos resultados utilizando os arquivos de saída dos programas.

4.8 Análises Estatísticas

Para análise estatística foi utilizado o programa GraphPad Prism v5.0. Foi utilizado o teste de Kruskal-Wallis, para verificar a diferença entre as medianas dos D-scores. Foi feito também um teste *post hoc,* para verificar em quais pares se encontravam as diferenças, o teste de Dunn. Foi feito também o teste de ANOVA para comparar as médias das distâncias do sítio de clivagem e o primeiro aminoácido do motivo PEXEL. Todos os testes foram considerados como significativos quando o valor p fosse menor que 0.05.

5. RESULTADOS

Parte I: Análise dos motivos nas sequências

5.1 Bancos de dados locais

Para construir o espaço de busca, foi utilizado o PlasmoDB (Aurrecoechea *et al.*, 2009), que é uma base de dados pública que contém informações e sequências de algumas espécies do gênero *Plasmodium.* Utilizando a estratégia de busca de genes por organismos foram recuperadas todas as possíveis sequências preditas de proteínas para cinco espécies diferentes: *P. falciparum, P. vivax, P. knowlesi, P. berghei* e *P. yoelli*. O número de proteínas recuperado para cada espécie está na Tabela 1.

5.2 Predição de peptídeo sinal

5.2.1 Análises do preditor

Antes de usar os dados de predição do SignalP, foi feita uma comparação dos resultados do programa entre diferentes espécies. Como esta foi a etapa inicial, foi feita uma análise mais profunda do preditor, para sabermos se este realmente seria um bom preditor para os organismos que estamos trabalhando. Assim, utilizamos a espécie *Homo* sapiens, que é a espécie que possui o maior número de sequências utilizadas no treinamento do algoritmo do programa, comparando com as cinco espécies de *Plasmodium*. A Figura 1 mostra a mediana do D-score para as espécies de *Plasmodium* e para *Homo sapiens*. Foi possível observar, que os valores das medianas variaram significativamente entre estas espécies através do teste de Kruskal-Wallis ($p < 0.0001$). Uma análise par a par foi feita através do teste de Dunn mostra que, com exceção das comparações *P. vivax* e *P. knowlesi*, *P. vivax* e *P. berghei*, *P. vivax* e *P. yoellii*, todas as outras comparações foram estatisticamente significativas.

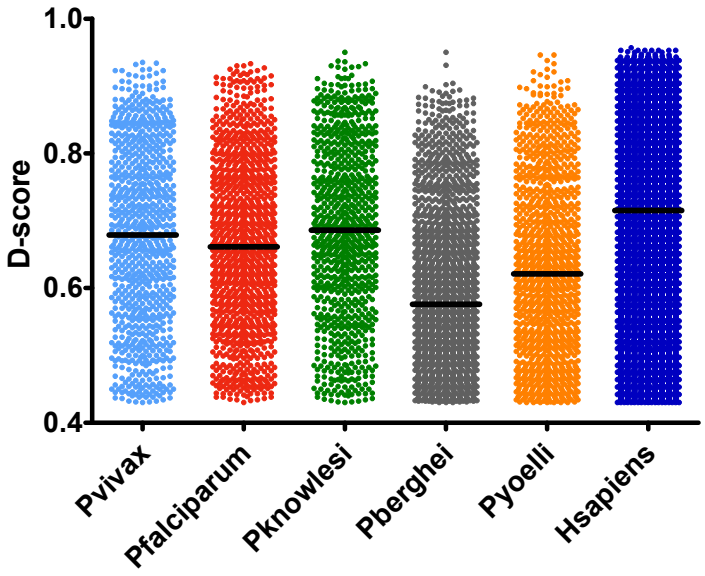

Figura 4: Mediana do D-scores (SignalP NN) das espécies de *Plasmodium* e *Homo sapiens*. Kruskal-Wallis test ($p<0.0001$). Teste de Dunn significativo, com exceção PVxPK; PVxPB e PVxPY

Outra comparação feita foi com relação à distribuição dos valores do D-score. Foi construído um gráfico **(Figura 2)** mostrando a distribuição dos valores para cada uma das espécies. Podemos observar que apesar dos valores da mediana serem menores para as espécies de *Plasmodium*, o padrão de distribuição destes valores se mostrou parecido com a espécie *H. sapiens.* Apesar de graficamente não parecer tão claro para algumas espécies, todas elas apresentaram uma distribuição de frequências bimodal, com uma moda maior nos menores valores do *Dscore* correspondente às proteínas sem peptídeo sinal e outra após o cut off (0,43) correspondente às proteínas com peptídeo sinal. Estes dados sugerem uma discriminação adequada das duas classes de proteínas em todas as espécies. As maiores modas presentes nos menores valores do *Dscore* variaram de 0,02 a 0,03 nas diferentes espécies e segunda moda, correspondente aos maiores valores do *Dscore* variaram de 0,53 a 0,87 com média igual a 0,70. Diante destes resultados, o SignalP v3.0, foi o programa utilizado para fazer as predições de presença do peptídeo sinal.

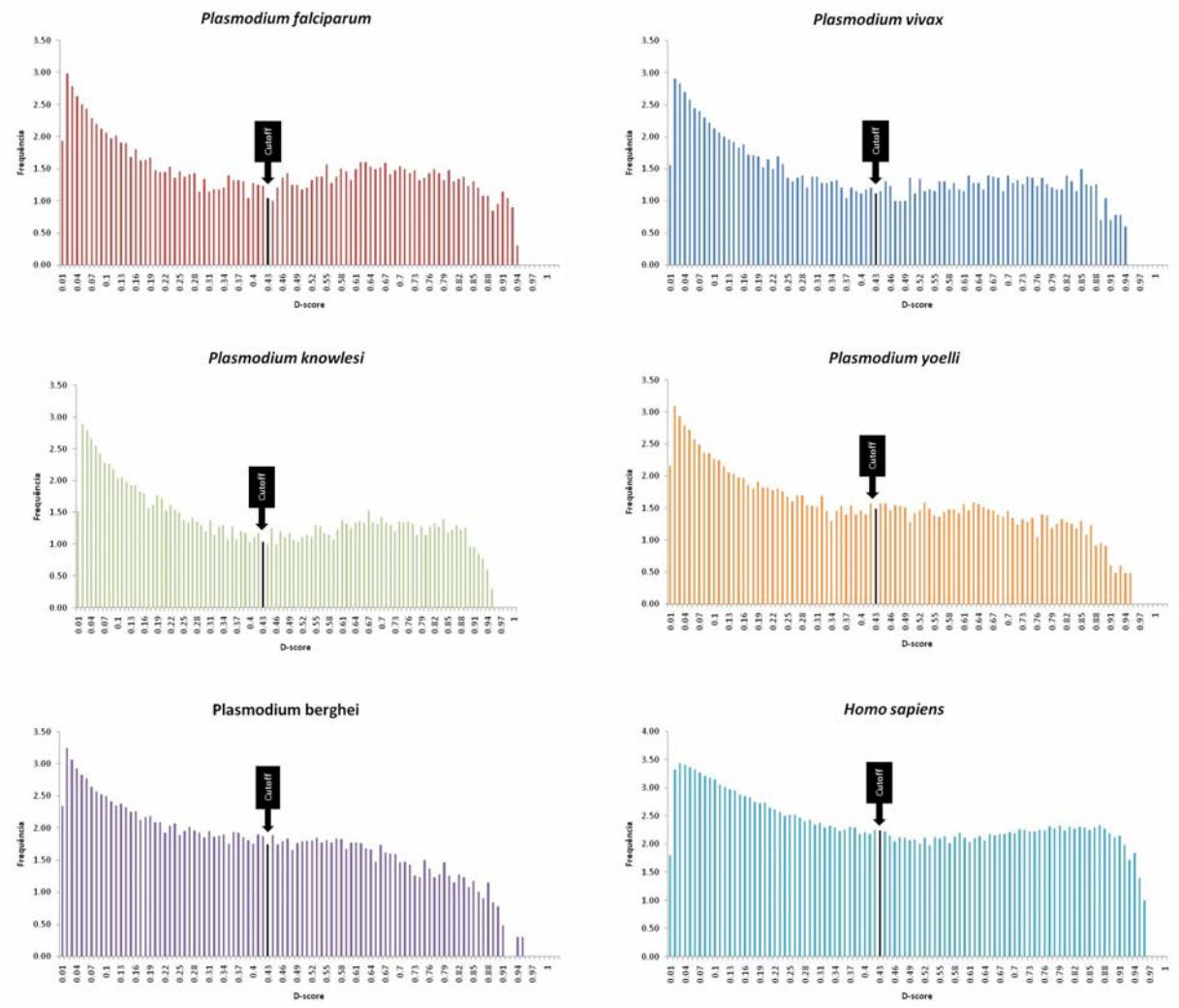

Figura 5: Distribuição da frequência dos valores de D-score para diferentes espécies. O valor do cut off (0,43) está indicado pela seta e pela barra preta.

5.2.2 Predição de peptídeo sinal em espécies de Plasmodium

Com o intuito de identificar as proteínas que potencialmente poderiam ser exportadas/secretadas, as sequencias de proteínas preditas das espécies de *Plasmodium*, foram submetidas ao SignalP (Tabela 1). A porcentagem de proteínas positivas preditas variou de 15 a 20%, sendo que *P. falciparum* foi a espécie com o maior número de proteínas com esta característica.

Tabela1: Número de proteínas para cada característica estudada por espécie.

Espécie	Nº total proteínas	PS	PEXEL	TM
P. falciparum	5524	1125 (20%)	1164 (21%)	1741 (32%)
P. vivax	5435	813 (15%)	922 (17%)	1286 (24%)
P. knowlesi	5197	827 (16%)	865 (17%)	314 (6%)
P. berghei	12235	1807 (15%)	737 (6%)	1285 (10%)
P. yoelli	7724	1128 (15%)	945 (12%)	2424 (31%)
Total	36115	5700 (16%)	4633 (13%)	7050 (20%)

PS: número de proteínas positivas na predição de peptídeo sinal; PEXEL: número de proteínas com pelo menos um motivo PEXEL nos 80 primeiro aminoácidos e TM: número de proteínas com pelo menos um motivo transmembrana predito.

5.3 Busca pelo motivo PEXEL

Na tentativa de encontrar proteínas que potencialmente poderiam ser exportadas, foi investigada a presença do motivo PEXEL no proteoma predito das espécies de *Plasmodium*. A tabela 1 mostra o número e porcentagem de proteínas por espécie que possuem um motivo PEXEL nos primeiro 80 aminoácidos. O número de proteínas com pelo menos um motivo variou de 737 a 1164 nas diferentes espécies, sendo que *P. falciparum* foi a espécie com o maior porcentagem de proteínas contendo este motivo.

Para estudar a conservação do PEXEL nas diferentes espécies, foi feito um alinhamento do consenso dos motivos encontrados nas sequências das diferentes espécies utilizando o programa WebLogo v3.1. A figura 3 mostra o consenso dos alinhamentos dos motivos. O aminoácido leucina (L) na posição 3 do motivo é o único 100% conservado em todas as espécies. Na posição 1 há uma discreta variação entre as frequências entre arginina (R) e lisina (K), sendo ambos aminoácidos básicos. Esta variação é mais evidente para o *P. falciparum* com o predomínio de arginina, este resultado é inesperado, pois devido ao alto conteúdo de A/T no genoma deste organismo seria esperado maior conteúdo de lisina. Na posição 5 do motivo houve uma variação nas frequências dos aminoácidos ácido aspártico (D), ácido glutâmico (E) e glutamina (Q). Geralmente os aminoácidos ácidos predominam nesta posição,

principalmente o ácido glutâmico. Nas posições 2 e 4 não foi observado nenhum aminoácido mais frequente.

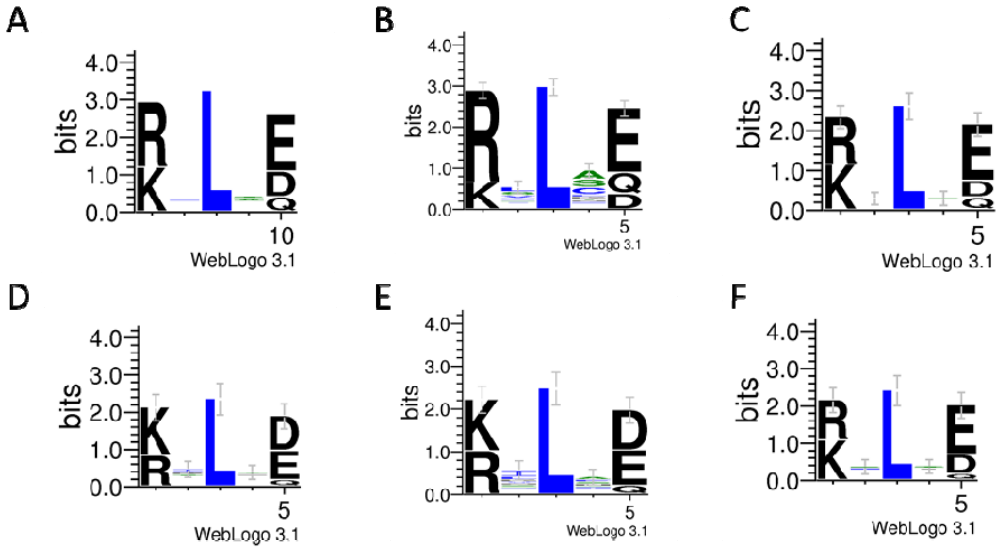

Figura 6: Consenso do motivo PEXEL em proteínas de species de *Plasmodium*: todas as proteínas (A), *P. falciparum* (B), *P. vivax* (C), *P. berguei* (D), *P. yoelii* (E) e *P. knowlesi* (F). O tamanho de cada letra é proporcional à frequência do aminoácido na determinada posição.

5.4 Busca por regiões transmembrana

A tabela 1 mostra os resultados do TMHMM para cada espécie. Esta característica foi a que mais variou entre as espécies. A porcentagem de proteínas positivas para a predição variou de 6%, no caso de *P. knowlesi*, a 32%, no caso de *P. falciparum*, novamente a espécie com maior número de proteínas positivas para a predição. A tabela 2 mostra uma análise do número de domínios transmembrana por proteína. A maioria das proteínas possuem 1 domínio transmembrana (de 45 a 66% dependendo da espécie), seguido pela presença de 2 domínios.

Tabela 2: Número de domínios transmembrana por proteínas.

Espécie	1 TM	2 TM	3 TM	4 - 9 TM	10- 19 TM	Total
P. falciparum	830 (48%)	342 (20%)	146 (8%)	320 (18%)	103 (6%)	1741(25%)
P. vivax	626 (47%)	231 (18%)	94 (7%)	253 (20%)	82 (6%)	1286 (18%)
P. knowlesi	142 (45%)	62 (20%)	23 (7%)	64 (20%	23 (7%)	314 (4%)
P. berghei	580 (45%)	256 (20%)	81 (6%)	267 (20%)	101 (8%)	1285 (18%)
P. yoelli	1591 (66%)	445 (18%)	117 (5%)	221 (9%)	50 (2%)	2424 (34%)
Total	3769 (53%)	1336 (19%)	461 (6,5%)	1125 (16%)	359 (5%)	7050

TM: número de domínios transmembrana

5.5 Combinações dos resultados

Para identificar as proteínas de interesse, foi necessário combinar os resultados, uma vez que cada programa procurava por uma característica de cada vez. A primeira análise feita foi em relação ao PS e o motivo PEXEL. Como o motivo PEXEL e o PS geralmente se encontram próximos, a primeira análise foi a medida da distância entre o ponto de clivagem predito do PS e o primeiro aminoácido do motivo PEXEL. A figura 2 mostra um gráfico da mediana das distâncias para cada espécie. A mediana da distância entre o sítio de clivagem do PS e o começo do PEXEL foi de 22 aminoácidos, variando de 21 a 24 nas diferentes espécies.

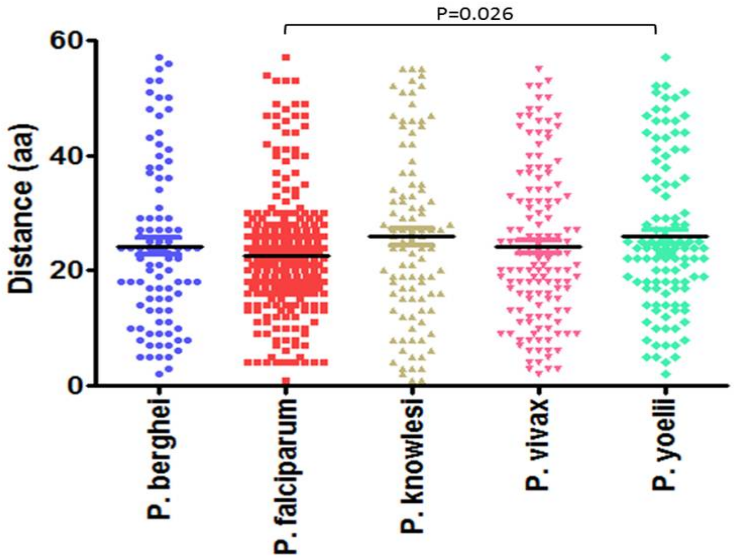

Figura 7: Distância entre o PS e o motivo PEXEL em cada espécie de *Plasmodium*. Significância estatística usando o Teste *t* não pareado. A barra indica o valor da mediana dos dados.

A partir da análise da distância, foi investigada a presença do motivo PEXEL na região do PS. A tabela 3 mostra a combinação das predições de peptídeo sinal, motivo PEXEL e domínio transmembrana. A porcentagem de proteínas com PEXEL na região do PS variou de 1 a 6%, sendo que *P. falciparum* foi a espécie com maior porcentagem. No caso de SP e domínio transmembrana, a porcentagem variou de 1,5 a 8%. Verificou-se que de 1 a 8% das proteínas das espécies estudadas, tinham PEXEL juntamente com domínio transmembrana. Finalmente menos de 1% das proteínas de todas as espécies tiveram a predição positiva para as 3 características analisadas.

Tabela 3: Combinação das predições

Espécies	N. de proteínas	Pexel + PS	SP + DTM	PEXEL + DTM	PS + PEXEL + DTM
P. falciparum	5524	305 (6)	437 (8%)	442 (8%)	43 (0,7%)
P. vivax	5435	92 (2)	204 (4%)	191 (3,5%)	35 (0,6%)
P. knowlesi	5197	97 (2%)	77 (1,5%)	48 (1%)	19 (0,3%)
P. berghei	12235	93 (2)	229 (2%)	163 (1%)	40 (0,3%)
P. yoelii	7724	109 (1,5%)	320 (4%)	244 (3%)	36 (0,4%)
Total	36115	696 (2%)	1267 (3,5%)	1088 (3%)	173 (0,6%)

N. de proteínas: número total de proteínas; PEXEL + PS: número de proteínas com domínio PEXEL na região do PS; SP + DTM: número de proteínas com PS e domínio transmembrana; PEXEL + DTM: número de proteínas com domínio transmembrana e PEXEL; PS + PEXEL + DTM: número de proteínas com as 3 características.

Como já discutido anteriormente, as análises de seleção dos possíveis candidatos e predição e síntese dos epitopos para validar a abordagem proposta neste trabalho, foram feitas apenas para a espécie *P. vivax*. Dessa forma, os resultados a seguir dizem respeito somente a esta espécie. Finalmente com o objetivo de identificar as possíveis proteínas que seriam exportadas e estariam ancoradas na superfície do eritrócito infectado, foram combinados os resultados das predições das três características: peptídeo sinal, motivo PEXEL e domínio transmembrana. A Figura 8 abaixo mostra um diagrama de Venn com a combinação dos resultados.

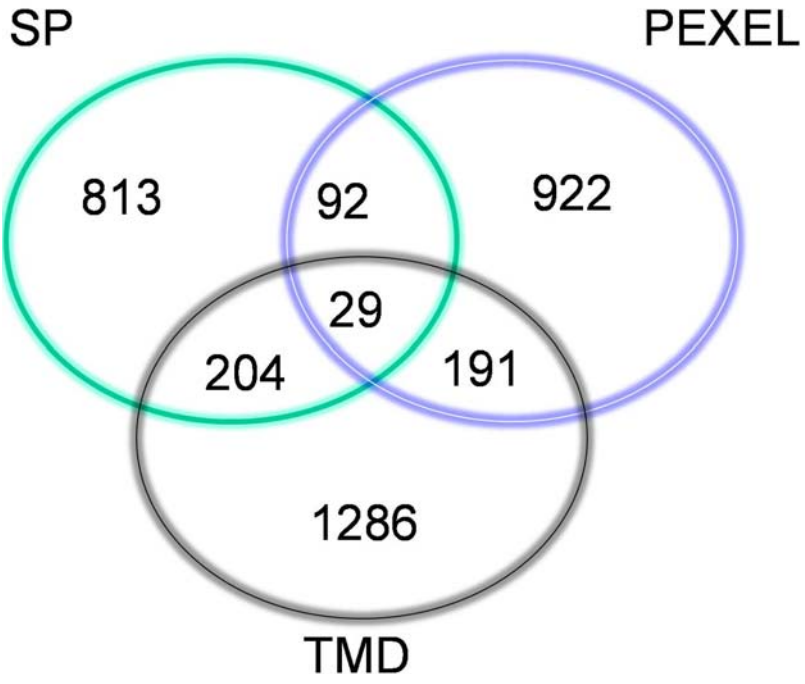

Figura 8: Diagrama de Venn mostrando o número de proteínas de cada grupo. SP: peptídeo sinal; TMD: domínios transmembrana. Este diagrama contém apenas sequências de *P. vivax*.

Após a análise destes resultados, o número total de proteínas com as três características estudadas foi de 29 para *P. vivax*. A Tabela 4 mostra o identificador destas proteínas juntamente com a descrição de cada uma delas. Antes de selecionar as proteínas que seriam submetidas à predição de epitopos, uma análise da lista foi feita para investigar quais destas proteínas já eram conhecidas. O objetivo foi gerar informações novas, por isso as proteínas selecionadas, foram aquelas que não tinham nenhum tipo de informação depositada no PlasmoDB tais como as proteínas hipotéticas.

Tabela 4: Lista com identificadores e descrição das proteínas de *P. vivax* de acordo com a presença das características analisadas.

ID	Descrição do produto
PVX_112665	TRAG (Pv-fam-a)
PVX_001020	Proteína hipotética
PVX_001015	Proteína hipotética
PVX_001000	Proteína hipotética
PVX_000810	Proteína hipotética
PVX_002510	Proteína ligadora de nucleossomo 1
PVX_002512	Proteína hipotética
PVX_002565	Proteína RAD (Pv-fam-e)
PVX_002790	Proteína hipotética
PVX_002900	Proteína hipotética
PVX_003845	SERA
PVX_003840	SERA
PVX_003830	SERA
PVX_003805	SERA
PVX_003800	SERA
PVX_003775	MSP4
PVX_003770	MSP5
PVX_003640	Proteína hipotética
PVX_003635	Proteína hipotética
PVX_003515	Proteína Phist (Pf-fam-b)
PVX_088810	TRAG (Pv-fam-a)
PVX_001670	Proteína hipotética
PVX_001685	Proteína Phist (Pf-fam-b)
PVX_099980	MSP1
PVX_091675	Antígeno de estágio do fígado
PVX_097575	TRAG (Pv-fam-a)
PVX_115055	Antígeno nuclear de proliferação celular
PVX_083550	TRAG (Pv-fam-a)
PVX_123960	Antígeno nuclear de proliferação celular

5.6 Predição de epitopos

Após a seleção das proteínas, segundo os critérios estabelecidos, uma lista de 10 proteínas foi selecionada para a predição de epitopos. Estas proteínas foram submetidas à análise de 3 programas diferentes.

A tabela 6 mostra o número de epitopos por proteína utilizando cada programa e também aqueles que eram comuns para 2 ou 3 programas. A diferença no número de epitopos nas proteínas se deve principalmente ao tamanho de cada uma, sendo que as maiores proteínas possuem maior número de epitopos preditos. Para todas proteínas o método que encontrou um maior número de possíveis epitopos foi o do IEDB.

Tabela 5: Número de epitopos preditos por proteína pelos programas utilizados.

ID	Número de epitopos preditos				
	Bepipred	**BcPreds**	**IEDB**	**2**	**3**
PVX_001000	9	13	18	11	4
PVX_002790	10	26	32	13	7
PVX_002900	7	7	12	4	5
PVX_003635	44	63	100	18	12
PVX_003805	22	29	43	15	8
PVX_003830	23	28	50	17	6
PVX_003840	18	26	42	19	6
PVX_003845	19	30	55	16	17
PVX_091675	30	36	58	12	11
PVX_000810	19	20	32	18	4

2: Epitopos que estavam na região predita por 2 dos métodos
3: Epitopos que estavam na região predita pelos 3 programas.

Posteriormente foi analisada a capacidade de predição de epitopos buscando a identificação de epitopos preditos pelos três programas (tabela 6).

Tabela 6: Proteínas de *P. vivax* selecionadas e dois epitopos preditos para cada proteína.

Identificador	Descrição	Epitopo 1	Epitopo 2
PVX_000810	Proteína hipotética	LTDPGYRGQIILMNWELS	TVTKTGPPPISAECPHNMVVLF
PVX_001000	Proteína hipotética	RDVRDVRVDQETRETLQGG	GGSVLSKEGEEATPGDFL
PVX_002790	Proteína hipotética	SNPPQGRDAGQVSDYGR	KDGDGSCAPGGNAETQE
PVX_002900	Proteína hipotética	RTRTWPAASSPEVDGDTC	DNVGCQYSPWSPWGPCVN
PVX_003635	Proteína hipotética	GNTASGSIVAKESESKEG	GLHTSAGQPAAQQGERGELG
PVX_003805	SERA	QGAVEGAKGPKPGAEEA	LGAVNPSGEEQPGPPGPP
PVX_003830	SERA	KWKVYPPKGETSSDKTL	GTGVGVAPGVGAAGRSV
PVX_003840	SERA	RPAAPQPPPTPPAEGSSS	QTAPAQPATPREPLSSLK
PVX_003845	SERA	TQPTDQPANQPVDQPTDQ	GEKPQTVAPPSASNPATP
PVX_091675	Antígeno de estágio do fígado	RGGPMQESYVGIGASSES	RAMEGSQTRAPPVMGANG

Os epitopos foram preditos pelos 3 programas: BCPreds, BEPIPRED e o preditor do IEDB.

Inicialmente foi analisada a capacidade de predição de epitopos de cada programa buscando a identificação de epitopos preditos por mais de um programa (tabela 5). A tabela final dos epitopos selecionados com maior potencial antigênico estão listados na tabela 6.

A tabela 6 mostra uma análise do número de epitopos por proteínas para cada método e também aqueles que eram comuns para 2 dos métodos e também os epitopos comuns aos 3 programas. A diferença no número de epitopos nas proteínas se deve principalmente ao tamanho de cada uma, sendo que as maiores proteínas possuem maior número de epitopos preditos. Para todas proteínas o método que encontrou um maior número de possíveis epitopos foi o do IEDB.

A tabela 7 mostra a lista das proteínas selecionadas para controle positivo e a tabela 8 a lista das proteínas selecionadas como controle negativo.

Tabela 7: Proteínas de *P. vivax* selecionadas como controle positivo e dois epitopos preditos para cada.

ID	Descrição	Epitopo 1	Epitopo 2
PVX_110810	Precussor do receptor Duffy	SNGQPAGTLDNVLEFVTGHE	DMEGIGYSKVVENNLRSIFG
PVX_099980	MSP1	SKDQIKKLTSLKNKLERRQN	YKARAKYYIGEPFPLKTLSE
PVX_119355	CS	GDRADGQPAGDRAAGQPA	GDRAAGQAAGDRAAGQAA
PVX_124060	MSP9	PAEDVSLMASIDSMI	ASIDSMIDEIDFYEK
PVX_081792	Proteína hipotética	QMEGFQKQLDRLSDSLSKIQKALGEYL	QKQLDRLSDSLSKIQKALGEY
PVX_113775	Proteína de membrana	NLGIIEVLIPSLPKKIDGC	NMHFFCACVLEEKRLVAHFEF
PVX_111175	Pvs25	YYSLFVFFLVQIALKYSKAAVTVDTI	HFKCMCNEGLVHLSEN
PVX_092275	AMA1	YRIPAGRCPVFGKGIVIENS	TTALSHPQEVDLEFPCSIYK
PVX_090250	TRAG	FNAPHAIEYHPRLLDK	FSSIVSAIMYLFSSSSVLFSNI
PVX_123575	TRAMP	FTYVAVLLLTICYQAISE	TRPCQVPLPPCNSLFEH

Tabela 8: Proteínas de *P. vivax* selecionadas como controle negativo.

ID	Descrição	Epitopo 1	Epitopo 2
PVX_123960	Antígeno nuclear de proliferação celular	NFMLKILSVIKEKSTVFLF	ASSLSDEVYISLSPNIPVSIR
PVX_080135	Metiltransferase dependete de s-adenosilmetionina	FQSCIFDGVVSISALQWLCNW	SKKYYLCLWTGSAVVSHLPEP
PVX_005570	Proteína hipotética	KECIKCYEGEFVQSHLNFD	YNIVQVSPPKVTQTVHDK
PVX_093630	1-cis peroxidoxina	DMWCILFSHPHDFTPVCTT	KGLPLTCRCVFFISPEKT
PVX_080630	Glutationa sintetase	EGIVTCMEKCHEAYLAKDQP	ISEISLFHNFVFCKNVSLLN
PVX_093655	Proteína hipotética	NKYLHHFYSKKRIKLYLCLHK	DPSVHESDASVEPVFCYPSA
PVX_114925	Proteína hipotética	SWRIFYIFKKYMYIHGTVVKG	KDVLIFSNRECFVNS
PVX_092935	Proteína hipotética	PYKHAMSAIPPAKSDHPFH	SSPFFFPTSALPDLPYSLN
PVX_059190	Proteína Vir6	RDAISKYSIIECKQYPYHA	LDSSSQLELSDVSTFSPK
PVX_092500	Proteína hipotética	KPSCHSVVPDYSNVSDLEAV	SQVSYLSIYEIFKHLDI

Parte II: Desenvolvimento do pipeline

A figura 5 mostra o desenho do *workflow* criado no Galaxy. A sequência consiste nas seguintes ferramentas: *Input dataset* é a ferramenta que carrega o conjunto de dados que vai ser submetido à análise; SignalP que foi acrescentado na versão original, assim como o buscador de motivos PEXEL e também o TMHMM. Cada um

destes programas vai gerar uma lista com os resultados das predições. A próxima etapa consiste na análise combinatória dos resultados. A ferramenta utilizada foi a *Join two datasets*, onde é possível combinar 2 conjuntos de dados tomando como parâmetro de comparação alguma coluna dos arquivos selecionados. Desta forma foi possível identificar as proteínas que possuíam as três características ao mesmo tempo. A partir desta lista, as proteínas selecionadas podem então ser submetidas à predição, que neste trabalho foi feita manualmente.

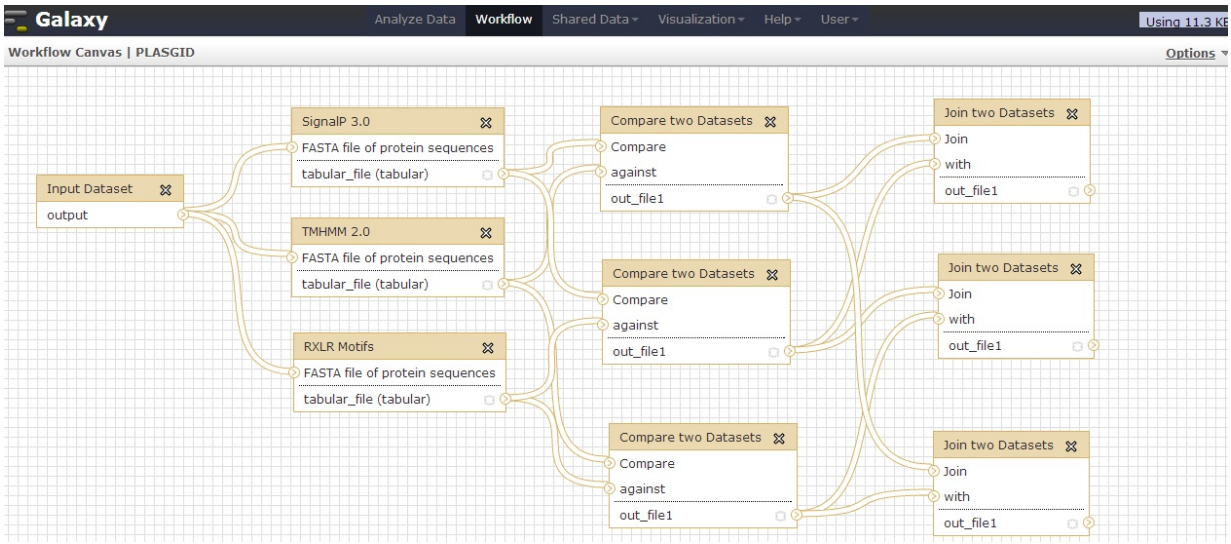

Figura 9: Esquema do pipeline desenvolvido na plataforma Galaxy. Cada caixa corresponde a uma ferramenta e as setas indicam para onde o arquivo de saída de um programa deve ser direcionado.

6. Discussão

Apesar do SignalP ter apresentado diferenças significativas em relação ao valor da mediana do *Dscore* para as diferentes espécies de *Plasmodium* em relação ao *H. sapiens*, o perfil de distribuição dos valores deste escore se mostrou semelhante para as espécies estudadas. O algoritmo que gera o *Dscore* foi treinado com sequências de diferentes organismos, porém a grande maioria dos dados é oriunda de mamíferos, principalmente *H. sapiens*, o que pode explicar esta diferença encontrada. O SignalP funciona com métodos baseados em aprendizagem de máquina, o que torna possível acrescentar informações ao conjunto de dados em que o algoritmo foi treinado. Incluir novos dados no conjunto de treinamento do algoritmo além de não ser uma tarefa simples, pois exigiria um conjunto validado de grupos controles, por exemplo, e o esforço pode não refletir em uma melhora do algoritmo. A inclusão de sequências de outras espécies que não estavam no conjunto de treinamento original poderia alterar o valor da mediana, porém como a distribuição dos valores foi semelhante, acreditamos que o número de sequências positivas não fosse alterado de forma significativa.

Na análise do motivo PEXEL, foi possível identificar algumas informações importantes. Mesmo tendo ocorrido algumas pequenas diferenças em relação ao aminoácido mais frequente em algumas posições do motivo, foi possível observar uma conservação do PEXEL nas diferentes espécies estudadas. Apesar de ser bem descrito para *P. falciparum*, este elemento ainda não havia sido bem estudado nas outras espécies do gênero. Alguns trabalhos recentes mostram que houve um grande avanço na compreensão do transporte de proteínas dentro do *P. falciparum* e no citoplasma da célula infectada (Bullen, 2012), porém mais estudos em outras espécies são necessários para ajudar na compreensão dessas vias de transporte e para traçar melhores estratégias no controle da doença. Este resultado mostra que possivelmente existe um mecanismo de exportação compartilhado entre as espécies estudadas. Outro ponto importante foi confirmado pela análise da distância do PS em relação ao motivo, que era bem conhecida para *P. falciparum* e é de aproximadamente 20 aminoácidos

(Goldberg, 2012; Hiss *et al.*, 2008) também para as outras espécies do gênero estudadas.

Com relação às proteínas com domínios transmembrana, Lingelbach e Przyborski em 2006, já haviam chamado a atenção para a necessidade de se investigar a topologia destas proteínas durante o seu transporte ao longo da via secretora. Os resultados obtidos para estas predições combinadas podem ajudar a identificar proteínas que são transportadas por esta via e que possuem um domínio que pode permitir que esta proteína se ancore em alguma membrana ao longo da via de secreção ou mesmo na membrana da célula hospedeira. Estas últimas poderiam em algum momento entrar em contato com o sistema imune do hospedeiro e desencadear uma resposta capaz de controlar a infecção. Mais estudos para caracterizar estas proteínas são necessários e nossa abordagem pode contribuir para diminuir o tempo desta etapa através da identificação das proteínas transmembrana.

Em todas as predições realizadas, a espécie que apresentou a maior porcentagem de proteínas positivas para a característica observada, foi o *Plasmodium falciparum*. Além das diferenças na biologia de cada espécie, um outro fator que pode talvez contribuir para este fato, é que esta espécie possui a melhor anotação do seu genoma. Diferentes motivos podem ser atribuídos a esta questão. Por ser a espécie responsável pela grande maioria das mortes, ela recebeu e continua a receber mais atenção do meio acadêmico e das agências de fomento. Dessa forma, um esforço maior tem sido investido para tentar controlar esta espécie. Assim, o *P. falciparum* foi a primeira espécie a ter o seu genoma seqüenciado e por esta razão o trabalho de anotação teve mais tempo de ser revisto e atualizado. No momento da busca pelas sequências no PlasmoDB, a versão que dispúnhamos contava com muitos erros. Algumas espécies, como *P. yoelli* e *P. berghei* possuíam uma anotação ruim, o que pode ter prejudicado a análise comparativa. Outra espécie, que devido a erros sistemáticos nas sequências foi excluída da nossa análise, foi *P. chabaudi*. Menezes-Neto 2013 também encontrou este problema e optou pela exclusão desta espécie na sua análise.

É válido lembrar mais uma vez a importância do *P. vivax* no contexto deste trabalho. Esta espécie é responsável por mais de 4/5 de todos os casos de malária registrados no Brasil anualmente e o laboratório no qual este projeto foi desenvolvido, já conta com muitos projetos envolvendo diferentes aspectos da doença causada por esta espécie, inclusive acompanhamento de pacientes de regiões endêmicas da doença. Por esse motivo, uma análise mais profunda destas proteínas foi feita no intuito de procurar nas proteínas selecionadas pelas características estudadas, regiões onde possivelmente poderia haver epitopos, para que pudéssemos utilizá-los como ferramenta de validação da abordagem computacional. Das proteínas identificadas como potenciais antígenos foi possível encontrar algumas que já possuíam informações confirmando sua antiginecidade. Algumas famílias de proteínas como Rifin, PfEMP1 e Stevor, que são conhecidas por serem exportadas/secretadas fizeram parte da lista final, o que corrobora a nossa abordagem para identificar esta classe de proteínas(Marti *et al.*, 2004).

As proteínas que não tinham informação sobre sua antigenicidade foram submetidas à predição de epitopos de células B. Mesmo para as proteínas menores foi possível identificar a presença de vários epitopos preditos. Ao combinar as predições e buscar epitopos que estavam em regiões preditas pelas 3 metodologias, estamos aumentando a chance de identificar um epitopo promissor e esperamos confirmar estes dados com experimentos.

As diferenças encontradas no número de proteínas presentes em cada espécie podem refletir em particularidades na biologia dos parasitos. *P. falciparum*, por exemplo, tem diferentes vias de invasão descritas, enquanto *P. vivax* tem uma única conhecida. A alta suscetibilidade que o *P. falciparum* ao sequestro nos capilares, pode sugerir um número maior de proteínas ancoradas, uma vez que algumas destas proteínas se ligam a moléculas de adesão no endotélio.

Na estratégia desenvolvida neste trabalho, foi possível identificar um grupo de proteínas de *Plasmodium* que são exportadas/secretadas. Ultimamente alguns trabalhos têm identificado uma classe de proteínas que são exportadas, porém não

possuem o motivo PEXEL em sua sequência, chamadas de Proteínas Exportadas Pexel Negativa (PNEPs). Porém, como é possível adicionar e remover ferramentas facilmente do workflow, uma análise destas proteínas pode ser feita e em breve esperamos também incluí-las no estudo. O fato de ser customizável faz com que diferentes análises possam ser feitas independentemente. Proteínas que possuem vias de exportação não clássicas, ou que possuem peptídeos sinal não típicos, dentre outras podem ser buscadas utilizando a abordagem proposta, basta retirar alguma ferramenta ou substituir uma por outra mais indicada para o tipo de análise desejado.

Este trabalho se concentrou em desenvolver uma estratégia flexível e automática para busca de proteínas com um interesse particular, no caso potenciais antígenos. Os parasitos da malária são organismos extremamente complexos em termos de biologia e vivem em um ambiente em que muitas mudanças estão acontecendo o tempo todo. Por isso é preciso pensar em múltiplos fatores influenciando eventos celulares como o transporte de proteínas, que foi abordado aqui. Dessa forma, uma metodologia que leva em consideração esta plasticidade de rotas é de extrema relevância na tarefa de identificar quais proteínas sintetizadas pelos parasitos estarão acessíveis ao sistema imune do seu hospedeiro, para que este possa responder à infecção na tentativa de eliminar este parasito.

7. Conclusão

Através da comparação das predições do SignalP de humanos e espécies do gênero *Plasmodium*, é possível afirmar que este preditor se mostrou uma boa estratégia para identificar sequências com esta característica para o gênero estudado, mesmo este não ter sido desenvolvido utilizando informações deste gênero. Nas espécies de *Plasmodium* estudadas, o peptídeo sinal mostrou um comportamento semelhante ao de *H. sapiens*, mostrando que a via clássica de exportação pode ser conservada.

O motivo PEXEL, que é importante para exportação de algumas proteínas foi identificado nas 5 diferentes espécies de plasmódios estudadas. Apesar de pequenas variações quanto a freqüência de cada aminoácido nas posições dos motivos, foi possível identificar uma conservação desta sequência, inclusive quanto a sua posição em relação ao peptídeo sinal. Este fato sugere que um mecanismo de exportação de proteínas parece ser compartilhado entre diferentes espécies do gênero. A predição dos domínios transmembrana mostrou uma certa variação entre as espécies, e não foi possível observar um padrão neste tipo de característica analisada.

Como os parasitos deste gênero são extremamente complexos e apresentam uma plasticidade nas vias de exportação, uma estratégia que levasse isso em consideração foi desenvolvida. Diferentes programas que procuram separadamente por características desejadas em proteínas foram integrados para que uma análise simultânea destas características pudesse ser feita da forma mais automática possível. Foi possível identificar proteínas que já eram conhecidas em termos de exportação, o que corrobora em parte a metodologia. Além disso, a estratégia adotada neste trabalho permitiu a identificação de novas proteínas de *Plasmodium* que provavelmente são exportadas para a célula hospedeira, o que pode ajudar no desenvolvimento de uma possível vacina contra estes parasitos. 8. Perspectivas

Sintetizar os peptídeos correspondentes aos prováveis epitopos das proteínas selecionadas segundo a técnica de *Spot synthesis* (Frank, 2002).

Validar o método de predição através da análise da reatividade imune de soros de pacientes infectados pelo *Plasmodium vivax* frente aos prováveis epitopos dos potenciais candidatos selecionados.

Disponibilizar os preditores de epitopos no pipeline e integrá-los ao galaxy para que esta etapa também possa ser automatizada.

8. Anexos

Anexo 1: Script Perl para busca do motivo pexel

```
#!/usr/bin/perl -w

use Getopt::Long;
use Bio::SeqIO;

my $sequencia;
my $motif;
my $output;

my $usage = "\n$0 -i aqruivo_entrada -m motif -o arquivo_saida\n";

GetOptions ('i=s' => \$sequencia, 'm=s' => \$motif, 'o=s' =>
\$output);
if(!$sequencia or !$motif or !$output) { die $usage;}
if (! -e $sequencia) {die "\nCannot find file: $sequencia\n\n";}

open (OUTFILE, ">$output");

my $obj_seq = Bio::SeqIO->new(-file => $sequencia);

my @matches;
my @pos;
my $seq = $obj_seq;

while ($seq = $obj_seq->next_seq()){
      @matches=();
      @pos=();
      $size = 80;
      if ($seq-> < $size){
            $size = $seq->;
      }

    my $ssubseq = $seq->subseq(1,$size);
    while($ssubseq =~ m/$motif/g) {
          push(@matches,$&);
             $position=pos($ssubseq)-($motif)+1;
             push(@pos,$position);
    }
      $num_matches=scalar(@matches);
      $id=$seq->display_id;
      $vetor_motif=" ";
      $vetor_pos=" ";
      foreach $temp (@matches){
                $vetor_motif=$vetor_motif."-".$temp;
      }
      foreach $temp2 (@pos){
                $vetor_pos=$vetor_pos."-".$temp2;
```

```
      }
      $vetor_motif=~s/ \-//g;
      $vetor_pos=~s/ \-//g;
      if($num_matches>0){

           print OUTFILE
"$id\t$num_matches\t$vetor_pos\t$vetor_motif\t".$seq->seq."\n";
      }
}
close (OUTFILE);
```

Anexo 2: Arquivo XML da ferramenta de busca por motivo PEXEL usada para carregar na versão do galaxy utilizada.

```
<tool id="motif_finder" name="MOTIFINDER" version="1.0.0"
URL_method="post">

     <description>Procura motivos em sequÃªncias
biolÃ³gicas</description>

     <command interpreter="perl">motif_finder.pl -i $input -m $motif -
o $output</command>

     <inputs>

         <param format="fasta" name="input" type="data"
label="TYPE_LABEL_HERE" help="ANY_FURTHER_INFO">
          <param name="motif" type="text" label="Motif" help="Motif
expression">

     </inputs>

     <outputs>

         <data format="fasta" name="output" />

     </outputs>

<help>
  This tool searches for a motif typed by user
     </help>
</tool>
```

Anexo 3: Lista com os identificadores das proteínas com as posições dos epitopos preditos pelo Bepipred.

ID	Posição dos epitopos
pvx000810	35-53
	87-97
	101-158
	171-181
	186-235
	258-273
	318-335
	393-424
	459-470
	484-505
	515-527
	531-539
	554-559
	574-584
	595-608
	641-657
	681-691
	702-709
	794-804
pvx001000	27-38
	46-89
	97-118
	122-165
	206-251
	297-396
	523-537
	562-568
	609-622
pvx002790	19-82
	98-343
	346-379
	388-662
	765-789
	804-813
	848-854
	861-868

	890-898
	944-971
pvx002900	32-62
	72-79
	107-113
	145-152
	168-180
	195-219
	246-274
pvx003635	24-30
	38-70
	74-85
	103-109
	117-138
	145-159
	172-185
	189-200
	208-227
	231-244
	250-261
	274-308
	316-331
	339-357
	368-414
	426-450
	460-472
	477-515
	522-561
	564-594
	663-673
	836-886
	915-975
	983-1054
	1065-1074
	1093-1107
	1181-1204
	1216-1226
	1241-1254
	1369-1378
	1576-1588

	1621-1785
	1795-1804
	1848-1856
	1876-1886
	2012-2030
	2040-2048
	2092-2128
	2153-2170
	2200-2214
	2223-2239
	2311-2320
	2332-2339
	2409-2441
pvx003805	23-184
	194-201
	218-225
	239-252
	274-286
	310-318
	327-334
	372-397
	453-460
	513-539
	546-557
	567-575
	622-637
	649-674
	684-708
	767-774
	785-804
	820-831
	852-868
	877-1051
	1090-1099
	1110-1117
pvx003830	25-150
	245-255
	349-364
	378-387
	412-425

	439-447
	478-507
	521-530
	536-544
	577-590
	616-628
	639-662
	704-712
	722-729
	740-750
	776-794
	808-823
	827-852
	854-894
	903-1001
	1024-1035
	1038-1048
	1060-1068
pvx003840	41-155
	166-172
	191-197
	248-256
	351-370
	479-509
	550-556
	590-603
	629-641
	652-675
	753-762
	765-773
	790-801
	820-927
	932-990
	1018-1025
	1030-1038
	1048-1056
pvx003845	25-272
	277-290
	329-340
	377-388

	474-493
	573-580
	614-641
	647-655
	725-736
	752-762
	768-777
	788-804
	810-877
	888-908
	956-971
	976-991
	996-1139
	1185-1193
	1205-1216
pvx091675	13-25
	44-69
	112-133
	143-211
	221-233
	244-278
	296-379
	391-478
	535-545
	548-570
	628-646
	665-675
	701-709
	719-738
	804-815
	818-826
	832-840
	924-942
	949-957
	964-981
	988-1015
	1052-1065
	1090-1099
	1113-1120
	1129-1141
	1148-1159

1301-1319
1364-1374
1464-1474
1490-1498

Anexo 4: Lista do resultados da predição de epitopos pelo método do IEDB.

PVX_008100

1	7	31	NSLLAIWCILFSICEYGYVGSLRIG	25
2	47	53	GSDVDVR	7
3	66	72	NILCNNI	7
4	80	87	ASFLSQKK	8
5	158	164	KKKVNLS	7
6	239	256	DLDVFPGLYFAGIGYDSL	18
7	273	278	RGQIIL	6
8	289	299	ANDLATLQPLN	11
9	314	323	IKECSSVSDY	10
10	330	336	EASVSGS	7
11	350	355	KKFLQE	6
12	362	383	KTYLVKSNCVKYTVGLPPYVRW	22
13	393	402	VNGLPPHFTG	10
14	406	416	DSECASDVYEQ	11
15	423	431	CETVHAWIR	9
16	437	444	GTHVIMEA	8
17	451	459	TKIIRVENS	9
18	468	485	GVSVKAQIKAQFGFASVG	18
19	507	514	EQLVVIGG	8
20	541	550	NIKLLPISTI	10
21	561	574	EKALIYYTRLYGFS	14
22	586	595	IVKILTASTT	10
23	601	622	PPPISAECPHNMVVLFGFVVKQ	22
24	640	648	ESGASSCTS	9
25	656	682	YDVSYTYIECGPQALPFTEQVVSVSGT	27
26	684	699	YNSVKCPNDYSVLFGF	16
27	709	728	QSALYSYFTPCRPGLKSCSL	20
28	736	747	KSYIYLVCVDAT	12
29	752	758	LNALSMI	7
30	761	770	DDLHSAVNRY	10
31	776	786	GELVVTCPSEG	11
32	799	823	SSPYVTVPFGKCAKSLKACSVHGSG	25

PVX_001000

1	4	19	FPLLLLLLVVFAVNQL	16
2	50	58	DDLLPSCDS	9
3	78	86	SEKVKVVDK	9
4	91	97	SMLVDVG	7

5	120	126	VRDVRVD	7
6	184	205	LLRSVIGQINFVQGSAELLKVA	22
7	214	220	GGSVLSK	7
8	257	287	YANVLLNEGKHVLVGNVRNFLSRVFNLIVRE	31
9	390	396	NCRLPKR	7
10	407	413	LYNYYSS	7
11	430	437	VSRYFTFS	8
12	455	466	FIESVRSILFDS	12
13	472	489	KAVFSSFAVVVETLFSLI	18
14	491	523	EEKVIADMYSYVKLFFQDLDILNLKVLHFLSSS	33
15	530	560	FVGPPDLSLTNFEYILAKIYSRSVLANILSP	31
16	583	608	FSFLEGVKMVHSAIPSEGVSAVVLGN	26
17	612	618	QVNVPIP	7
18	622	644	DTLCKFIPIRKKLLYERLSVTRK	23

PVX_002790

No.	Início	Fim	Peptídeo	Peptídeo
1	4	16	LFLPLLVLCVKGL	13
2	23	33	ATPITPYHLDS	11
3	63	69	AGQVSDY	7
4	77	101	STSLPSGVSFLQACTINSCLSVDND	25
5	103	110	VGGVAATD	8
6	112	121	VGGVAAVDEV	10
7	127	136	SDAVGGVAAG	10
8	175	181	NSPVDVG	7
9	188	193	GSCAPG	6
10	249	255	GAQVAGQ	7
11	340	345	KKVHLE	6
12	381	387	IGLVEKG	7
13	391	399	KGSPPQVSE	9
14	476	482	HMEVPHG	7
15	597	607	QQEVVPPSEEA	11
16	615	631	LEEVSPPPAAAVQLSQK	17
17	637	647	VKEHAQQHAHQ	11
18	675	693	LANVDVILHGLKDKLSRHK	19
19	697	704	NQELKLKF	8
20	712	732	EYKLYKDLIHKSVEILTLRLM	21
21	739	752	KKLKQSSDVALQKY	14
22	758	764	YGLVNFS	7
23	774	780	EEKVASG	7
24	791	798	KLFCPMDC	8

25	809	817	HPTQCYKLE	9
26	823	849	IQKICEPFVDLHSGTCPADFHHCAIAE	27
27	856	861	YSIFAS	6
28	868	876	PQFITIRGY	9
29	878	897	LHECLQLLVVNKGSTCSPSS	20
30	905	927	SEEILLEPVFTKLLHNEILLENI	23
31	935	946	YNICLAQFYQQP	12
32	972	980	SDIQILGID	9

PVX_002900

No.	Início	Fim	Peptídeo	Peptídeo
1	6	27	FFFVSIFLCLSQESSLELSKKL	22
2	54	59	AASSPE	6
3	62	69	GDTCFIFS	8
4	77	87	NCWCPRGYIMC	11
5	92	104	VRDVQSKLHQIKD	13
6	131	145	MSVVIDYELAVLCDD	15
7	154	162	FKIIGASGF	9
8	165	171	NEEVIQQ	7
9	177	194	TYVPRKCTVNNFYLCRKV	18
10	196	206	DDNVGCQYSPW	11
11	209	215	WGPCVNG	7
12	232	237	ELCLWN	6

PVX_003635

o.	Início	Fim	Peptídeo	Peptídeo
1	4	24	NVFYVILVFLFFLLIIKCEYV	21
2	29	36	EEVVFETD	8
3	57	62	DDVKIG	6
4	86	95	VGMALSVEGS	10
5	97	102	MSLVTM	6
6	108	117	AGSVIVVEAV	10
7	125	131	GSIVAKE	7
8	138	145	GHIVVEEA	8
9	161	173	IAMVVDPIVVDPI	13
10	186	193	GSSVVVEE	8
11	197	212	SDAVIDVATDVVPVRS	16
12	227	234	VEEAAVSN	8
13	243	251	ADAVHLLEE	9
14	263	274	HIAYDVIVVEEA	12
15	288	296	DGVVSVEEP	9
16	309	318	AYGVVAVEEA	10
17	327	339	SEPVADSLVVKEE	13

18	355	362	GAIVVEEA	8
19	388	400	DTPVVEEASVSSS	13
20	411	416	TSIVEE	6
21	421	429	EPIVSVATS	9
22	431	436	ESPVSD	6
23	451	460	IEEPIVSSFV	10
24	470	481	VAMVDEPIVSSA	12
25	497	504	ESIVSSPV	8
26	514	526	VAMVEEAAVSSLP	13
27	558	570	TAIVEESTVSSAP	13
28	598	604	PFLVEYQ	7
29	609	621	QYDQVVDEIILRS	13
30	650	666	MRYLIKVLIGTTVVTPS	17
31	685	693	RHFVTSLFR	9
32	698	704	FKLVESH	7
33	708	736	ANYVYYVGDERVYSYRYSYRLVLNLLSSE	29
34	738	756	FSFYLDLNKFTLLEILDSY	19
35	768	774	RYYPFHM	7
36	780	790	LSEFVGHYFEF	11
37	798	807	KHRHLIAEEV	10
38	819	826	LFAQLSRL	8
39	828	836	PNFVLYPFG	9
40	846	910	SSDLSSPTASSSFASSDLSSPTASSSFASSASSASPPPRVEHYLDAEVNLVFSIFEEFLVNLLND	65
41	930	944	VEAASSISAPVVEPL	15
42	962	969	EVAPLSEA	8
43	971	983	SAELSGSLGIAAG	13
44	1076	1085	YNQYILKNLN	10
45	1109	1119	KDGFIVHVATA	11
46	1121	1127	HTPLLFN	7
47	1129	1139	SKDVYAYLSQM	11
48	1150	1159	VYDYLMTIIR	10
49	1167	1175	YDGLLQSRR	9
50	1203	1215	AKRVIERVVVHDG	13
51	1232	1238	NSFLLFI	7
52	1252	1265	RAGVHSFYRSVNLS	14
53	1267	1272	FTPAAV	6
54	1279	1287	HDQLKLLKK	9
55	1296	1308	EATCVLAFLYLIG	13
56	1316	1321	LQLPYG	6
57	1327	1333	DHSVRLI	7
58	1340	1360	LCNFLSGILYHLNLPIFVNNS	21
59	1381	1388	NSFFYIYY	8

60	1416	1424	SKIVSYSLG	9
61	1452	1457	IFLRDY	6
62	1477	1498	SPFLSSCDYLLSNILGAVVDSL	22
63	1500	1507	NTSVIESG	8
64	1528	1539	NKSLFEYFLKLA	12
65	1544	1558	SYALAALGEIYYLGN	15
66	1585	1600	SALSTGYAYLDEYKKH	16
67	1603	1609	KEEVLKA	7
68	1613	1618	EDVLKM	6
69	1662	1668	GYFPFAS	7
70	1672	1684	SAGLHTSAGQPAA	13
71	1721	1727	ADFPSGL	7
72	1735	1742	PSAPPAQF	8
73	1750	1763	ASGYASGQAGVPLG	14
74	1820	1833	VEHVLAKYNIYKFG	14
75	1843	1850	AGDYLKKA	8
76	1862	1868	LGHLYSG	7
77	1872	1877	GVKVKD	6
78	1890	1898	SYKYYSLSA	9
79	1903	1920	IISLYNKSILILKGVNPN	18
80	1933	1940	KQFHFIGL	8
81	1943	1955	ERLYVLTKLLRRS	13
82	1963	1975	GSLLLSVILSELG	13
83	2001	2011	YNLVEGLLADL	11
84	2027	2038	TPPLDRIIHRLY	12
85	2046	2052	YYSLQGK	7
86	2055	2070	LKGHSILKSCAMVTRK	16
87	2074	2090	GKSVLFNVKFCRYLRDA	17
88	2122	2134	QPDVHHQDVYELT	13
89	2145	2152	RYDQIVKS	8
90	2181	2197	SEFLHCYYKPISYYQIK	17
91	2235	2241	SQQVKDL	7
92	2248	2254	RYSPLLE	7
93	2261	2267	LFYYQNT	7
94	2280	2299	SKNCDVCKQHYDIYSAYYGY	20
95	2301	2309	KSTLDLIRK	9
96	2325	2330	LQFLIR	6
97	2338	2347	EPLYYKALFL	10
98	2353	2367	DSLKNILQIYLKLAT	15
99	2370	2399	HNACNVIGVLGIFKILFKKVFFDVTVFFSR	30
100	2433	2459	SAAPVPAANSLQRYLFTDLNSCALQSN	27

PVX_003805

No.	Início	Fim	Peptídeo	Peptídeo
1	4	23	RLSLILILCVVCRDCAVRCT	20
2	29	35	QGAVEGA	7
3	131	141	PGAIPQVAPRD	11
4	184	218	VKSSLLKGHKGVKVTGPCGASFLVFFAPYLFIDVD	35
5	222	235	SNVYLGTDLSDLEV	14
6	255	280	FKFVALIGEDHLTIKWKVYDPAVKTP	26
7	299	311	EFTAVQVHTVIQQ	13
8	321	327	NYALSSG	7
9	329	358	PEKCDAVATNCFLSGSVYIEKCYRCTLKMK	30
10	360	376	VDPSDVCYNYIPKVESA	17
11	378	385	SQEAIPAK	8
12	397	409	TASIGKILQGVYK	13
13	416	421	NEVLTF	6
14	426	454	AALKAELLNYCSLMKKVDASGVLGHYQLG	29
15	471	480	SDHVLSSLQN	10
16	484	491	NPAICLKN	8
17	499	512	KTGLLLPNLFYNHL	14
18	519	527	TSNVTHVDD	9
19	537	544	YDGVIDFA	8
20	563	570	NAEYCDRA	8
21	573	580	AGSCVAKM	8
22	590	622	NSWLFASKVHLETIKCVKGYDHVGASALYVANC	33
23	629	641	DKCHSPSNPLEFL	13
24	649	670	FLPADSDLPYSYKQVGNACPEP	22
25	680	685	NVKLLG	6
26	698	704	YTAYQSD	7
27	713	720	FIKLVKSE	8
28	724	740	KGSVIAYVKVAGALSYD	17
29	743	752	GKKVLSLCGS	10
30	755	762	PDLAVNIV	8
31	770	782	AEGVKKPYWLLQN	13
32	802	820	PGCHHNFIHTAAVFNLDMP	19
33	832	840	IYNYYLKSS	9
34	868	883	HESVLHGQEVAEAAVN	16
35	890	897	APSHSGAV	8
36	911	923	EGVPPLPAKQEVT	13
37	933	938	LGAVNP	6
38	1026	1037	STPVQEAPLSKA	12
39	1042	1063	SPPVAPEAAVLGSEVTHVLKYI	22
40	1069	1076	KLNLVTYK	8

41	1080	1090	ALSSGHDCWRS	11
42	1098	1107	YEECVKLCEA	10
43	1118	1125	PGFCLYEH	8

PVX_003830

No.	Início	Fim	Peptídeo	Peptídeo
1	4	28	RLCALLILYMLLNGHSVKCTAAVGQ	25
2	37	43	DQGVSSQ	7
3	60	66	TGVPQSR	7
4	89	100	ATHVPVVPQNGH	12
5	123	132	TQNVQQANLQ	10
6	136	142	DTQVAAT	7
7	149	161	NPIQVKASLLRDQ	13
8	166	187	ITGPCKSYFQVYLVPYLYLNVN	22
9	225	232	FKLVVYMY	8
10	239	245	KWKVYPP	7
11	271	279	SMQVVVMSE	9
12	298	328	PEKCDAIANECFLSGVLDVQKCYHCTLLLQK	31
13	335	342	CFKFVSPT	8
14	360	366	PNVVELE	7
15	368	375	TIDLLLNK	8
16	387	396	VDQLAIIDSS	10
17	398	410	QSDLLKYCSLMKE	13
18	427	439	DVFANLTHLLQSN	13
19	441	450	DHDVPSLKNK	10
20	452	466	KSPAICLKNVGHWVG	15
21	468	479	KTGLVLPTLEYS	12
22	503	520	SADVHPLNVSDKLFCNDE	18
23	528	535	SSSCIAKI	8
24	545	561	TSWLFASKVHLEAIKCM	17
25	563	577	GHDHVASSALYVANC	15
26	584	596	DKCHAASNPLEFL	13
27	604	625	FLAAESDLPYSYKAVNNACPEP	22
28	635	640	NVKLLD	6
29	653	659	YTAYQSD	7
30	668	675	FIKLVKSE	8
31	681	695	SAIAYVKAQGALSYD	15
32	701	707	VQSLCGG	7
33	711	717	DLAVNIV	7
34	731	737	SYWLLQN	7
35	756	762	PDHCQNN	7
36	765	782	HTAAVFNLDVPVVAPAPS	18

37	823	830	NNSVLYGQ	8
38	836	844	SALPASVGD	9
39	847	854	GKVVAQTA	8
40	878	884	GVGVAAP	7
41	894	904	IAAAAVVVVGG	11
42	916	928	GVGVAPGVGAAGR	13
43	935	941	QLPAGAA	7
44	947	953	TQHVGGS	7
45	975	984	GQVSQTVNEA	10
46	986	992	PSTVEKP	7
47	997	1003	STGVISE	7
48	1005	1015	ITGVFHFLKNV	11
49	1034	1056	DKACSRVQSSDVDKLDDCVKFCE	23
50	1063	1074	KGKVSPGYCLTK	12

PVX_003840

No.	Início	Fim	Peptídeo	Peptídeo
1	4	38	RICALLLIVRCLLGRDHMCAYFCQNCLFTGLVFVN	35
2	52	58	AVTVQAG	7
3	95	107	AQRPAAPQPPPTP	13
4	129	164	PQNVSSVAAPVPPVEATPPPSVTNPFKVKSSLLKDQ	36
5	169	188	ITGPCESYFQVYLVPYLYMN	20
6	214	225	EKHLLHNICAAD	12
7	227	248	TFKLVLYMYDGVLTIKWKVYPL	22
8	274	281	SMQVVVMT	8
9	285	292	KTVYVESK	8
10	301	311	PEKCDAIANEC	11
11	313	332	MSGVLDVQKCYHCSLLLQKK	20
12	335	345	AQECFKFVSPK	11
13	374	381	NIFVKVYK	8
14	387	399	YKEVDQLAIIDSS	13
15	401	413	KSELLKYCSLMKE	13
16	430	438	DVFAHITTM	9
17	444	453	DLDVFSLKSK	10
18	455	462	KNAALCLK	8
19	471	479	KTGLVLPNL	9
20	506	533	DGVVDLTSLQSVDVSPFLVTDKLFCNDD	28
21	541	551	TSSCVAKIEVQ	11
22	553	590	QGDCATSWLFASKVHLETIKCVKGYDHVGASALYVANC	38
23	597	609	DKCHAASNPLEFL	13
24	617	638	FLAAESDLPYSYKAVNNACPEP	22
25	648	653	NVKLLD	6

26	666	672	YTAYQSD	7
27	681	688	FIKLVKSE	8
28	694	708	SAIAYVKAQGALSYD	15
29	711	720	GKKVLSLCGG	10
30	724	730	DLAVNIV	7
31	738	750	AEGVKKPYWLLQN	13
32	769	794	PPGCQHNFIHTAAVFNLDIPVVVPAP	26
33	869	875	STAVERS	7
34	912	918	PQAVTRP	7
35	922	935	PERVASGLASVQHS	14
36	945	958	TSSVTQNPPVATRP	14
37	966	972	TAPAQPA	7
38	975	981	REPLSSL	7
39	986	1002	GVNVTEVKEALHFLKSV	17
40	1009	1017	SNFVAYVNA	9
41	1036	1043	QTECIEFC	8
42	1051	1062	KGKVSPGYCLTK	12

PVX_003845

o.	Início	Fim	Peptídeo	Peptídeo
1	4	32	ALPFLFILSVALLDNAIKCDEEVTIPDPP	29
2	58	63	GAVPAG	6
3	76	85	EQAVDQPLTQ	10
4	100	106	DQALTQP	7
5	112	118	NQPVDQP	7
6	128	133	DQPVDQ	6
7	155	161	VDQPLTQ	7
8	200	205	DQPVDQ	6
9	244	255	DEPLPQPIVEAS	12
10	257	264	RAAAAAVK	8
11	270	277	EAKCAQLK	8
12	284	303	ITGPCGAKFQVFLIPHVTIN	20
13	310	315	AIHLGK	6
14	317	324	LDDVVITK	8
15	332	340	GKSPPLLQF	9
16	345	351	DSLLNQC	7
17	357	381	FKFVVVVKGEELILKWKVYEKVPSP	25
18	388	395	DVRTFLLK	8
19	401	409	ITAIQVHTA	9
20	415	428	SFLLESKEYILADD	14
21	430	446	PAQCDLIAANCFLSGSL	17
22	448	460	IEGCYKCALLSEN	13

23	463	474	LSSPCFDYLSPD	12
24	492	509	LKEVQLAASIGKILQGVF	18
25	517	523	NELVTFD	7
26	530	542	KEELLNYCALMKE	13
27	545	555	ASGVLDQYQLG	11
28	561	570	FANLTSILKN	10
29	585	592	NPAICLKN	8
30	601	616	KGLLLPSLSYTNVEAT	16
31	675	684	SSSCVAKIEV	10
32	687	705	QGACSNSWLFASKVHLESM	19
33	716	724	TSALYVANC	9
34	732	746	KCHVASNPLEFLDIL	15
35	748	774	ETQFLPAESDLPYSYKAVNNVCPQPKS	27
36	780	788	WADVKLLDK	9
37	792	806	PNAVSAKGYAAYQSD	15
38	815	822	FIKLVKSE	8
39	826	835	KGSVIAYVKA	10
40	845	854	GKKVLSLCGS	10
41	858	864	NLAVNIV	7
42	872	882	AEGVKKPYWLL	11
43	903	922	PPGCQHNFIHTAAVFNLHIP	20
44	930	938	KKPLLYNYY	9
45	945	954	FYNHIFYRGV	10
46	967	978	QEKVSISAVSAN	12
47	985	997	LDGVDHSSVVVEG	13
48	1009	1015	AQGVESP	7
49	1099	1107	PSQVAAPGA	9
50	1114	1123	PQTVAPPSAS	10
51	1128	1134	PPSAAPA	7
52	1139	1156	ASLIKVKQITEVIHIMKH	18
53	1178	1186	NHDCLRSYS	9
54	1192	1201	LPECIQFCYD	10
55	1211	1218	SPGYCLNQ	8

PVX_091675

.	Início	Fim	Peptídeo	Peptídeo
1	4	9	FWSFIK	6
2	18	24	KPTVAHL	7
3	83	95	RTIYAEIPGLKTI	13
4	99	110	SMQVQGGIVSPY	12
5	137	158	FIPSVKVQEKTHVGGKSPPLAK	22
6	199	205	AEALPSG	7

7	214	220	ADLLKQL	7
8	234	243	AEKSLFIPKT	10
9	273	279	ESYVGIG	7
10	286	292	PKGLELK	7
11	303	320	GVNIPPKEVRSDHTVNSI	18
12	340	368	GKGVPPLSGLPPPSGTPPLSRGPPPVAPP	29
13	375	385	GGAVKPVYFER	11
14	412	417	RAPPVM	6
15	434	446	SRSIPVRPPPPSM	13
16	472	477	GGFLPQ	6
17	482	493	LIEHILSALKKG	12
18	502	517	KGRLKLEAEKLHLAAE	16
19	521	529	ALDALYRKM	9
20	542	549	AEVVTASQ	8
21	569	579	SEVHNMLHLIN	11
22	590	599	ETFAHFCNEL	10
23	677	683	SLKVKEE	7
24	697	705	KEELKKVES	9
25	708	716	ERKVAVLEE	9
26	746	754	RNEVVAELR	9
27	771	777	NEVVTEL	7
28	782	788	KDGVITD	7
29	794	804	KEEVAAELRTQ	11
30	844	849	KNLIVE	6
31	854	862	VEALRCQLE	9
32	868	873	EICQLR	6
33	889	894	QSLLEG	6
34	898	912	HQQHLEEVKRECELE	15
35	916	923	EIEVLRGE	8
36	942	949	RLAQVEKQ	8
37	957	964	RLAQVEKQ	8
38	992	1000	EEKLAQVEN	9
39	1006	1011	LTQVEN	6
40	1015	1021	EKLAQLR	7
41	1033	1042	HLELNIALLK	10
42	1067	1075	NRRVNLLQE	9
43	1103	1109	GDLLQAE	7
44	1118	1128	EQKVHTYVKQY	11
45	1142	1148	FSDLLEK	7
46	1182	1188	EELVRKK	7
47	1208	1217	KDLLQNCMQL	10
48	1260	1271	RCELLMSQCSEL	12

49	1296	1302	LSSVNDE	7
50	1315	1321	DQLVTQN	7
51	1344	1368	EVKHELTSLQGKYEQVASQNAHLKS	25
52	1375	1383	KLSHRLSDL	9
53	1388	1394	RDLIRVT	7
54	1402	1416	IKNVSIIKSLQAQIK	15
55	1418	1424	QTSLYKE	7
56	1445	1451	NLSLSQL	7
57	1473	1489	SAALDSLTKRLSFIEAS	17
58	1496	1503	GQEVIRRA	8

Anexo 5: Lista das proteínas selecionadas com os epitopos preditos pelo BCPREDS.

PVX_000810

Posição	Epitopo	Score
184	YRFGDEEEESEEKSPGKSKS	1
105	DDDASTTTGGGSETSDDDEL	1
592	ASTTVTKTGPPPISAECPHN	1
792	FYGETHTSSPYVTVPFGKCA	0.994
485	GGSTNVSSDHSSKKNEDNYE	0.994
640	ESGASSCTSKQGNTNKYDVS	0.992
445	QLGGKITKIIRVENSSVNQM	0.989
408	ECASDVYEQKKTSEECETVH	0.988
220	DDGDEREKKQSNTRKAMKKD	0.982
512	IGGNPIKDVTKEENLYEWSK	0.96
127	PDQDEGADEGEAEDQLATQE	0.958
163	LSRHEKLIEDKKQQTENTFK	0.881
376	GLPPYVRWEQTTAFKNAVNG	0.868
243	FPGLYFAGIGYDSLFGNPLG	0.845
338	MGFGAFSASTGYKKFLQEAS	0.802
726	CSLNMNEHDDKSYIYLVCVD	0.795
619	VVKQNFWDHTNKLQSYEMEI	0.776
62	GKNANILCNNISCNSKKEAS	0.765
268	TDPGYRGQIILMNWELSNKG	0.762
661	TYIECGPQALPFTEQVVSVS	0.734

PVX_001000

317	TGERSGDPTGERSGDPTGER	1
220	KEGEEATPGDFLGGNNPNGG	1
604	VVLGNAGGQVNVPIPGADDT	0.999
145	QHGEEETGDDSKRAKQDDEA	0.995
372	PTAERSDEPTADPKGDPTNC	0.994
338	GEPTGERSGEPTGERSGEPT	0.988
54	PSCDSPGGRNDEHQVNKEVS	0.985
95	DVGESGGKSSPGVAEESGPS	0.969
296	RGGEASIERSGEPVGERSGE	0.965
521	SSSSTENTQFVGPPDLSLTN	0.953
117	GRDVRDVRVDQETRETLQGG	0.934
24	KLKHGKWDDGSYSERTRWRM	0.893
414	LEEMLKGRGIRWKTDRVSRY	0.827

PVX_002790

256	KEEEDKEEEEEKEEEEKEDE	1
277	EKEDEEEKEDDEEKEDKEDS	1
354	EAKGGSPPGGNSNGEETKAG	1
553	NVEEKTPQGGRPEGDAAANA	1
312	GENKSKSDDQGGKKKEGLKG	1
228	SEGQKGGQPASQGDGKAEGQ	1
393	SPPQVSETKGEEAKTTPPNK	1
506	EEDEKEDEETDDKTDEEMDD	1
149	EPRGDPPKEQRDIPKTNEPK	1
530	EADEEADKGGKGRKKKNEIP	1
181	GADKDGDGSCAPGGNAETQE	1
586	REDEELHGGASQQEVVPPSE	0.999
421	NGGSSRGGSPEEGKELPKSN	0.999
52	PEKSNPPQGRDAGQVSDYGR	0.996
459	GEKSEDKGGNKQQGESKHME	0.993
206	EKKGVKGATEKEVQMGGKAD	0.992
30	HLDSHEGGKDKPDDTPNGGT	0.986
626	VQLSQKQGEEQVKEHAQQHA	0.986
957	SNRGGKPSRANKARKSDIQI	0.978
482	GEEDAGEKEAKRSHGDDADG	0.961
761	VNFSDWTKNDKAQEEKVASG	0.956
882	LQLLVVNKGSTCSPSSIEKN	0.921
683	HGLKDKLSRHKELKNQELKL	0.877
9	LVLCVKGLLRNREPATPITP	0.871
800	KKNCHNKPDHPTQCYKLEQR	0.866
655	GGGKNYESNMYTLDKKMHKN	0.807

PVX_002900

246	MEQTRPCGGPSGAPARGERD	1
47	KRTRTWPAASSPEVDGDTCF	0.997
202	QYSPWSPWGPCVNGRQRRTR	0.988
224	MRSNQNNEELCLWNGKRIPR	0.975
142	LCDDGKDKGNADFKIIGASG	0.975
163	VSNEEVIQQTERDTTYVPRK	0.953
69	SAMEGDPTNCWCPRGYIMCS	0.87

PVX_003805

48	NVGEAGTGGPGGPGADGGTE	1
943	QPGPPGPPGTSGLAEQPGPE	1
89	EAEAEPARGPEPEPEAGGEG	1
900	EREGTANGQGPEGVPPLPAK	1
25	TTEAQGAVEGAKGPKPGAEE	1

144	ETSSDAADSSSPDQNPLPGA	1
969	GSTGSVGQPGQPGSSGTPRS	1
879	EAAVNGAGGEPAPSHSGAVQ	1
1005	GQSGSSGTPGSDGPSGHTGP	1
377	ASQEAIPAKASDEESSQEEL	1
1028	PVQEAPLSKAPGTDSPPVAP	1
165	NTKVGNAATPPEGAKEETQV	0.999
268	IKWKVYDPAVKTPTPNEKVE	0.999
921	EVTEASNGEATGLGAVNPSG	0.997
851	NVGGQESASSKNATEGAHES	0.996
123	EAPSDSARPGAIPQVAPRDT	0.995
510	NHLEGSTPSTSNVTHVDDSS	0.989
1112	CENDAAPGFCLYEHAKEEDC	0.966
686	PTNEPNSMSTKGYTAYQSDH	0.946
816	NLDMPVEENPQKEDAQIYNY	0.936
773	VKKPYWLLQNSWGKHWGDKG	0.935
657	PYSYKQVGNACPEPKGHWQN	0.914
620	ANCSGKEANDKCHSPSNPLE	0.907
356	KMKKVDPSDVCYNYIPKVES	0.897
304	QVHTVIQQNGSNVFESKNYA	0.894
398	ASIGKILQGVYKKGENGLNE	0.892
236	TEKMGIQDNGKNKCEDKKTF	0.886
794	FKVDMNGPPGCHHNFIHTAA	0.747
188	LLKGHKGVKVTGPCGASFLV	0.739

PVX_003635

864	SSPTASSSFASSASSASPPP	1
2012	KRGPESGANPGRNGATPPLD	1
838	GAPSSSSASSDLSSPTASSS	1
1640	KPGAAKEPGASNKPGAAKEP	1
493	STEEESIVSSPVEEGESNQD	1
1724	PSGLYGGGSAGPSAPPAQFD	1
2038	YKKGKGGKYYSLQGKAKLKG	1
66	EGGAVGMQLQGEQQGEEGAQ	1
2097	GGSSGGRSGGRSGGSSGGSN	1
335	VVKEEAASSNTAKGSEPETD	1
539	EEETTVSSASEESEPKTDDT	1
405	EPKEDDTSIVEEPIMEEPIV	1
1180	LWRGGDSPGGEVNGGASRGP	1
382	DEPKADDTPVVEEASVSSST	1
356	AIVVEEAAASNAIREVESKP	1
1238	IERGEEASGETAPPRAGVHS	0.999

139	HIVVEEAASSDVIRETESES	0.999
1676	HTSAGQPAAQQGERGELGEL	0.999
937	SAPVVEPLPGGDDLKEVPGE	0.999
2423	GKHDERGKPGSAAPVPAANS	0.999
1770	GKEENTADTSSSQSDDIIKK	0.999
1027	QRDEGEPTTDQRDEGEPTTD	0.999
204	ATDVVPVRSSTEESESKTDG	0.999
1749	AASGYASGQAGVPLGGGEQK	0.999
576	KPDDTATVEEAPPERRVANR	0.998
983	GAAGEAVGAEGSPAEGEGGE	0.998
428	TSEESPVSDATEESEPQADS	0.998
1697	ERGSSPERGSPVGAPNFMGT	0.998
2150	VKSEMEKGGTPREKIQRGEK	0.997
708	ANYVYYVGDERVYSYRYSYR	0.996
227	VEEAAVSNGAKGNEPKADAV	0.994
314	AVEEAADSNSAKESEPVADS	0.994
22	EYVNNEKEEVVFETDLLGGG	0.99
802	LIAEEVHHNGIKKSGNKLFA	0.985
1004	ITGKHGEGEPVADKPDGREA	0.982
2326	QFLIRNSDEEDHEPLYYKAL	0.982
118	ADGNTASGSIVAKESESKEG	0.973
1571	FEFWKKAADQGDTTSALSTG	0.969
450	IIEEPIVSSFVEEGNSKQNG	0.966
290	VVSVEEPKADDDVKENEHTA	0.962
259	KESEHIAYDVIVVEEAVDSN	0.958
1915	KGVNPNLKTFNEKCEKTLKQ	0.944
2401	KAQLMRAIREGGSGGSGERG	0.94
2199	EEEKRQSLKGNAQQERSNLL	0.935
1306	LIGINDNGGKLQLPYGFPRN	0.931
1973	ELGDHPNNVNASMLWDLKRR	0.929
472	MVDEPIVSSAAEESESKPDN	0.926
657	LIGTTVVTPSYKEKKKSFNY	0.915
2220	RSSNHMGYNTPGQDKSQQVK	0.913
1493	AVVDSLKNTSVIESGVYEES	0.887
165	VDPIVVDPIANDTAKGDELK	0.886
1202	PAKRVIERVVVHDGELSTGM	0.88
597	KPFLVEYQERVNQYDQVVDE	0.873
1874	KVKDYDGGNKMENLKKSYKY	0.864
1400	LSNENRIIPKAVDNMKSKIV	0.845
1092	TQKERQRMEDYKNKYELKDG	0.841
514	VAMVEEAAVSSLPKESELKP	0.839
1514	SEKHRHLINNTVDHNKSLFE	0.79

763	LNNSPRYYPFHMYTGKSLSE	0.788
1456	DYFDFTFDKKVYKNAVMKNN	0.786
2305	DLIRKYREGDEYTIKSKRKE	0.775
1803	KNAEQYFQKAIRNNEESVEH	0.774
1062	ESYKEAPKAEDLKKYNQYIL	0.735

PVX_003840

61	PGSAGGGDGGGGGGGSSAGS	1
97	RPAAPQPPPTPPAEGSSSPP	1
877	TADGVTTPPGGGGANGGRQG	1
838	STVQGPDGPAGPSGPDGNVG	1
136	AAPVPPVEATPPPSVTNPFK	1
908	DSGSPQAVTRPTPPPERVAS	1
943	STTSSVTQNPPVATRPAPQT	1
35	VFVNEGARCAPPAGEGTAVT	1
354	QAKGEDEENPNEEELQKTID	1
784	NLDIPVVVPAPNSDPEINNY	0.999
965	QTAPAQPATPREPLSSLKGS	0.998
161	LKDQKGLKITGPCESYFQVY	0.985
1049	DACKGKVSPGYCLTKKRGSN	0.974
817	LNKYEAGSDGNSDDAKSVSG	0.973
653	DPTNEPNSVSTKGYTAYQSD	0.949
260	IRKYKMKDIGQPITSMQVVV	0.942
741	VKKPYWLLQNSWGKHWGDKG	0.935
629	KAVNNACPEPKSHWKNLWEN	0.913
182	VPYLYMNVNATSSEIEMEPM	0.906
480	SPQNFEDSPTQTFDGGDEQA	0.89
762	FKVDMHGPPGCQHNFIHTAA	0.883
586	YVANCSGTEANDKCHAASNP	0.858
239	LTIKWKVYPLNGDKTADRTL	0.822
525	TDKLFCNDDYCDRAKDTSSC	0.819
330	QKKESAQECFKFVSPKIKNS	0.778
1026	CSRAFSTDADKQTECIEFCE	0.745

PVX_003830

102	QSNPPPSTSSGGPNGGGGAS	1
868	GAAAGPTGQEGVGVAAPGGR	1
912	ERGTGVGVAPGVGAAGRSVG	1
933	GQQLPAGAASPGTGTQHVGG	1
66	RPQSGGGDTGTQGGQQDRAS	1
956	GGTNSGNGATGNSGPNTQRG	1
486	ESFDEGEEDTSNSTTTQSAD	1

771	NLDVPVVAPAPSSDPEINDY	1
45	TANSAGQGIGSSTGSTGVPQ	0.998
348	EDIQTKGEDEENPNVVELEE	0.998
1016	KKGKVSSNFVTYDNTKAIGD	0.993
1059	NWNECKGKVSPGYCLTKKKG	0.99
236	LTIKWKVYPPKGETSSDKTL	0.99
980	TVNEAAPSTVEKPKSIDSTG	0.985
158	LRDQKGLKITGPCKSYFQVY	0.958
736	QNSWGKHWGDDGNFKVDMHG	0.949
640	DPTNEPNSVSTKGYTAYQSD	0.949
372	LLNKIYKNGEGESNEVDQLA	0.946
257	IRKYKMKDIGQPITSMQVVV	0.942
24	AAVGQGQGTGVSTDQGVSSQ	0.939
616	KAVNNACPEPKSHWKNLWEN	0.913
573	YVANCSGKEANDKCHAASNP	0.897
437	QSNSDHDVPSLKNKLKSPAI	0.897
818	EKGPVNNSVLYGQSGEETSA	0.89
327	QKKENAEECFKFVSPTIKNR	0.874
511	VSDKLFCNDEYCDRAKDSSS	0.853
185	NVNAKESEIEMDPMFMKVDD	0.84
533	AKIEAGDQGDCSTSWLFASK	0.739

PVX_091657

359	SRGPPPVAPPTGGPPPGGAV	1
552	EEGSATQGGEAEGEEEQSEV	1
669	RATREREASLKVKEEELERA	1
924	VSRREEELEKQLSEEHEERL	1
526	YRKMSQLKSIKEESEGAEVV	1
338	SEGKGVPPLSGLPPPSGTPP	1
625	REMEEKERRRSASFKKKEKQ	1
465	PMSGNYPGGFLPQKEKELIE	0.999
600	RRAKAKNEETIRLYKKREKK	0.998
41	IRGEEDKVESEEKVTAPPKM	0.996
425	SWTKSNDDISRSIPVRPPPP	0.996
1172	GMIKMLENQTEELVRKKIEE	0.995
1212	QNCMQLEESHKKMKQQLEEE	0.995
300	RPSGVNIPPKEVRSDHTVNS	0.992
1355	KYEQVASQNAHLKSSEKEQR	0.989
715	EEELETRRAQEMEQLGGEQR	0.989
140	SVKVQEKTHVGGKSPPLAKN	0.988
1081	EKEIKEVIENRRKESEQIRE	0.987
91	GLKTIKKKSMQVQGGIVSPY	0.986

952	EEHEERLAQVEKQLSEEHEE	0.981
1122	HTYVKQYEEMSEEYETKKKE	0.968
1415	IKEQTSLYKEYTDELKDEIE	0.966
114	SAKNELSSNTTDENRSEPFG	0.958
895	EKQHQQHLEEVKRECELEKS	0.947
1488	ASIRDQDYGQEVIRRADELH	0.94
690	REKEAALKEELKKVESEAER	0.937
161	GDAELEDHPKLGAKEEGEII	0.927
267	RGGPMQESYVGIGASSESIP	0.927
4	FWSFIKSKKNDDEPKPTVAH	0.925
488	SALKKGLTKEDEAFKGRLKL	0.925
404	RAMEGSQTRAPPVMGANGRM	0.918
997	QVENSHEERLTQVENSHEEK	0.875
862	EKMKDEEICQLREEQQRALN	0.843
1272	KMRNDEMMEGFCREKAQRER	0.831
1233	NLAMQENEQNIVRITNTYES	0.816
217	LKQLNGKDSKQESNLRSAEK	0.73

PVX_003845

1029	AESEEEGEDEPEEEGDGEQE	1
1060	EEAEEAGAEAEEGEEESEEG	1
1119	APPSASNPATPPSAAPAPTR	1
1008	ESAQGVESPPEAEEQDEEEG	1
51	PLPGEGAGAVEPAGGETDSG	1
1098	AEPSQVAAPGASPGGEKPQT	1
104	LTQPTDQPANQPVDQPTDQP	1
29	PDPPQSPDENPGGKDDPPGD	0.999
615	EATLPENAPEKEDPPKGSQK	0.999
165	DQPAGEPLTQSTDQPAGEPL	0.999
235	PTDEPLTQPTDEPLPQPIVE	0.999
277	LKDQDGVKITGPCGAKFQVF	0.998
82	QPLTQSTDQPADQPAEQPAD	0.997
141	EQPAGEPLTQSTDEPVDQPL	0.995
321	VVITKKMHKGVGGKSPPLLQ	0.994
1203	EWNNCKGAPSPGYCLNQRRR	0.993
214	QPTDQPTEQPTDQTTDQPTE	0.991
188	TDQPADQPADQSADQPVDQT	0.984
723	ANCSGKEEKDKCHVASNPLE	0.984
891	WGDDGTFKVDMHGPPGCQHN	0.969
480	YEEIKRKAQQQGDLKEVQLA	0.946
678	SCVAKIEVEDQGACSNSWLF	0.942
951	FYRGVQTGEESEMGISGQEK	0.935

643	IINFDSNEETNMQSTSFIDN	0.924
789	KQDDPNAVSAKGYAAYQSDH	0.91
298	IPHVTINVETETNAIHLGKK	0.844
371	LKWKVYEKVPSPSDNNKVDV	0.832
1143	IKVKQITEVIHIMKHIKSGK	0.83
760	PYSYKAVNNVCPQPKSHWQN	0.809
987	DGVDHSSVVVEGKEKNPAGE	0.789

9. Referências Bibliográficas

AMINO, R. *et al.* Quantitative imaging of Plasmodium transmission from mosquito to mammal. **Nature medicine**, v. 12, n. 2, p. 220-4, fev. 2006.

ARÉVALO-HERRERA, M.; CHITNIS, C.; HERRERA, S. Current status of Plasmodium vivax vaccine. v. 6, n. 1, p. 124-132, 2010.

ASANTE, K. P. *et al.* Safety and efficacy of the RTS,S/AS01E candidate malaria vaccine given with expanded-programme-on-immunisation vaccines: 19 month follow-up of a randomised, open-label, phase 2 trial. **The Lancet infectious diseases**, v. 11, n. 10, p. 741-9, out. 2011.

AURRECOECHEA, C. *et al.* PlasmoDB: a functional genomic database for malaria parasites. **Nucleic acids research**, v. 37, n. Database issue, p. D539-43, jan. 2009.

BARILLAS-MURY, C.; KUMAR, S. Plasmodium-mosquito interactions: a tale of dangerous liaisons. **Cellular microbiology**, v. 7, n. 11, p. 1539-45, nov. 2005.

BENDTSEN, J. D. *et al.* Improved prediction of signal peptides: SignalP 3. **0, J. Mol. Biol.**, v. 340, p. 783-795, 2004.

CHANG, H. H. *et al.* N-terminal processing of proteins exported by malaria parasites. **Molecular and biochemical parasitology**, v. 160, n. 2, p. 107-15, ago. 2008.

CRABB, B. S. *et al.* Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. **Cell**, v. 89, n. 2, p. 287-96, 18 abr. 1997.

CRABB, B. S.; KONING-WARD, T. F. DE; GILSON, P. R. Protein export in Plasmodium parasites: from the endoplasmic reticulum to the vacuolar export machine. **International journal for parasitology**, v. 40, n. 5, p. 509-13, abr. 2010.

CROOKS, G. E. *et al.* WebLogo□: A Sequence Logo Generator. p. 1188-1190, 2004.

DEPONTE, M. *et al.* Molecular & Biochemical Parasitology Wherever I may roam□: Protein and membrane trafficking in P . falciparum-infected red blood cells. **Molecular & Biochemical Parasitology**, v. 186, n. 2, p. 95-116, 2012.

DHANGADAMAJHI, G.; KAR, S. K.; RANJIT, M. The survival strategies of malaria parasite in the red blood cell and host cell polymorphisms. **Malaria research and treatment**, v. 2010, p. 973094, jan. 2010.

DUVAL, L. *et al.* African apes as reservoirs of Plasmodium falciparum and the origin and diversification of the Laverania subgenus. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 23, p. 10561-6, 8 jun. 2010.

EL-MANZALAWY, Y.; DOBBS, D.; HONAVAR, V. Predicting linear B-cell epitopes using string kernels. **Journal of molecular recognition : JMR**, v. 21, n. 4, p. 243-55, 2008.

FRANK, R. The SPOT-synthesis technique. Synthetic peptide arrays on membrane supports--principles and applications. **Journal of immunological methods**, v. 267, n. 1, p. 13-26, 1 set. 2002.

FREVERT, U. *et al.* Intravital observation of Plasmodium berghei sporozoite infection of the liver. **PLoS biology**, v. 3, n. 6, p. e192, jun. 2005.

GARCIA, L. S. Malaria. **Clinics in laboratory medicine**, v. 30, n. 1, p. 93-129, mar. 2010.

GIARDINE, B. *et al.* Galaxy: a platform for interactive large-scale genome analysis. **Genome research**, v. 15, n. 10, p. 1451-5, out. 2005.

GOLDBERG, D. E. Plasmodium Protein Export at Higher PEXEL Resolution. **Cell host & microbe**, v. 12, n. 5, p. 609-10, 15 nov. 2012.

GREENWOOD, B. M. *et al.* Malaria: progress, perils, and prospects for eradication. **Human Antibodies**, v. 118, n. 4, 2008.

HAASE, S.; KONING-WARD, T. F. DE. New insights into protein export in malaria parasites. **Cellular microbiology**, v. 12, n. 5, p. 580-7, 1 maio. 2010.

HANSSEN, E. *et al.* Targeted mutagenesis of the ring-exported protein-1 of Plasmodium falciparum disrupts the architecture of Maurer's cleft organelles. **Molecular microbiology**, v. 69, n. 4, p. 938-53, ago. 2008.

HILLER, N. L. *et al.* A host-targeting signal in virulence proteins reveals a secretome in malarial infection. **Science (New York, N.Y.)**, v. 306, n. 5703, p. 1934-7, 10 dez. 2004.

HISS, J. A. *et al.* The Plasmodium export element revisited. **PloS one**, v. 3, n. 2, p. e1560, jan. 2008.

KOLASKAR, A. S.; TONGAONKAR, P. C. A semi-empirical method for prediction of antigenic dete ~ inants protein antigens. v. 276, n. 1, p. 172-174, 1990.

KROGH, A *et al.* Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. **Journal of molecular biology**, v. 305, n. 3, p. 567-80, 19 jan. 2001.

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

LINGELBACH, K. Plasmodium falciparum: A Molecular view of protein transport from the parasite into the host erythrocyte. **Experimental parasitology**, 1993.

LINGELBACH, K.; PRZYBORSKI, J. M. The long and winding road: protein trafficking mechanisms in the Plasmodium falciparum infected erythrocyte. **Molecular and biochemical parasitology**, v. 147, n. 1, p. 1-8, maio. 2006.

MARTI, M. *et al.* Targeting malaria virulence and remodeling proteins to the host erythrocyte. **Science (New York, N.Y.)**, v. 306, n. 5703, p. 1930-3, 10 dez. 2004.

MENDIS, K. *et al.* The neglected burden of Plasmodium vivax malaria. **The American journal of tropical medicine and hygiene**, v. 64, n. 1-2 Suppl, p. 97-106, 2001.

MILLER, L. H. *et al.* The pathogenic basis of malaria. **Nature**, v. 415, n. 6872, p. 673-9, 7 fev. 2002.

MOTA, M. M.; HAFALLA, J. C. R.; RODRIGUEZ, A. Migration through host cells activates Plasmodium sporozoites for infection. **Nature medicine**, v. 8, n. 11, p. 1318-22, nov. 2002.

MOTA, M. M.; RODRIGUEZ, A. Migration through host cells by apicomplexan parasites. **Microbes and infection / Institut Pasteur**, v. 3, n. 13, p. 1123-8, nov. 2001.

NACER, A *et al.* Plasmodium falciparum signal sequences: simply sequences or special signals? **International journal for parasitology**, v. 31, n. 12, p. 1371-9, out. 2001.

OLIVEIRA-FERREIRA, J. *et al.* Malaria in Brazil□: an overview Review. p. 1-15, 2010.

OOIJ, C. VAN *et al.* The malaria secretome: from algorithms to essential function in blood stage infection. **PLoS pathogens**, v. 4, n. 6, p. e1000084, jun. 2008.

PARKER, J. M.; GUO, D.; HODGES, R. S. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. **Biochemistry**, v. 25, n. 19, p. 5425-32, 23 set. 1986.

PRICE, R. N. *et al.* Vivax malaria: neglected and not benign. **The American journal of tropical medicine and hygiene**, v. 77, n. 6 Suppl, p. 79-87, dez. 2007.

SCHWARTZ, L. *et al.* A review of malaria vaccine clinical projects based on the WHO rainbow table. **Malaria journal**, v. 11, p. 11, jan. 2012.

STURM, A. *et al.* Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. **Science (New York, N.Y.)**, v. 313, n. 5791, p. 1287-90, 1 set. 2006.

TALMAN, A. M. *et al.* Gametocytogenesis: the puberty of Plasmodium falciparum. **Malaria journal**, v. 3, p. 24, 14 jul. 2004.

VITA, R. *et al.* The immune epitope database 2.0. **Nucleic acids research**, v. 38, n. Database issue, p. D854-62, jan. 2010.