Ministério da Saúde

Fundação Oswaldo Cruz – FIOCRUZ

Centro de Pesquisa René Rachou

Programa de Pós-graduação em Ciências da Saúde

Aprimoramento da anotação N-terminal de proteínas através da predição de peptídeo sinal em proteínas ortólogas e desenvolvimento de uma ferramenta automática para a identificação de grupos ortólogos contendo erros de anotação

por

Armando de Menezes Neto

Belo Horizonte

Novembro/2012

Ministério da Saúde

Fundação Oswaldo Cruz – FIOCRUZ

Centro de Pesquisa René Rachou

Programa de Pós-graduação em Ciências da Saúde

Aprimoramento da anotação N-terminal de proteínas através da predição de peptídeo sinal em proteínas ortólogas e desenvolvimento de uma ferramenta automática para a identificação de grupos ortólogos contendo erros de anotação

por

Armando de Menezes Neto

Tese apresentada com vistas à obtenção do Título de Doutor em Ciências na área de concentração em Biologia Molecular e Celular

Orientação: Dra. Cristiana Ferreira Alves de Brito

Belo Horizonte

Novembro/2012

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

M541a 2012

Menezes Neto, Armando de.

Aprimoramento da anotação N-terminal de proteínas através da predição de peptídeo sinal em proteínas ortólogas e desenvolvimento de uma ferramenta automática para a identificação de grupos ortólogos contendo erros de anotação/ Armando de Menezes Neto. – Belo Horizonte, 2012.

XV, 92 f: il.; 210 x 297mm.

Bibliografia: f. 100 – 107.

Tese (doutorado) - Tese para obtenção do título de Doutor 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/genética 2. *Plasmodium*/imunologia 3. Peptídeos/imunologia I. Título. II. Brito, Cristiana Ferreira Alves de (Orientação).

CDD – 22. ed. – 616.936 2

Ministério da Saúde

Fundação Oswaldo Cruz – FIOCRUZ

Centro de Pesquisa René Rachou

Programa de Pós-graduação em Ciências da Saúde

Aprimoramento da anotação N-terminal de proteínas através da predição de peptídeo sinal em proteínas ortólogas e desenvolvimento de uma ferramenta automática para a identificação de grupos ortólogos contendo erros de anotação

por

Armando de Menezes Neto

Foi avaliada pela banca examinadora composta pelos seguintes membros:

Profa. Dra. Cristiana Ferreira Alves de Brito (Presidente)

Prof. Dr. Jerônimo Conceição Ruiz

Prof. Dr. Fabiano Sviatopolk Mirsky Pais

Prof. Dr. Gustavo Fioravanti Vieira

Prof. Dr. Francisco Pereira Lobo

Suplentes: Prof. Dr. Carlos Eduardo Calzavara Silva

Tese defendida e aprovada em: 30/11/2012

AGRADECIMENTOS

Gostaria muito de agradecer,

À minha orientadora e amiga, Dra. Cristiana Ferreira Alves de Brito, pelo seu papel fundamental no meu crescimento profissional, por em muitos momentos acreditar na minha capacidade mais que eu mesmo, por seu carinho e paciência, por sempre me estimular a exercer o pensamento crítico, porestar sempre disponível para discussões e conversas, por me dar uma “liberdade criativa” e muitas vezes me “salvar” quando eu me perdia nesta liberdade. Enfim, pelo exemplo tanto profissional quanto humano, que ficará como um modelo na minha vida.

Aos amigos do Laboratório de Malária do Centro de Pesquisas Rene Rachou (LAMAL), pesquisadores, funcionários, técnicos e estudantes. Por criarem o ambiente de união, confiança, companheirismo e prosperidade científica no qual este trabalho foi desenvolvido. Em especial,

À Dra. Luzia Helena Carvalho, chefe do laboratório de Malária, por compartilhar a sua paixão pela ciência, por sempre acreditar no meu potencial, por ficar tão feliz com o meu sucesso e por me ajudar tanto a alcançá-lo. Obrigado pelo voto de confiança!

Ao amigo Geraldo Felício, técnico do LAMAL, por todos estes anos de convivência, por sua paciência com todas as brincadeiras, por todas as conversas nas horas de folga e por sua disponibilidade em ajudar a todos e fazê-lo com a maior competência. Uma pessoa especial!

À Alice de Sabatino, pela sua extrema competência que torna o nosso trabalho muito mais fácil! Por sempre me receber bem humorada e disposta a ajudar e resolver todo tipo de problema e pelas conversas na hora do café!

À Denise Anete, à Sarah Resende e ao Ricardo Ribeiro, que colaboraram diretamente para a realização deste trabalho e durante este percurso se tornaram meus grandes amigos.

A todos os bons amigos que fiz no laboratório de Malária do CPqRR.

Ao grande amigo Antônio Mauro, pela ajuda fundamental na realização deste trabalho, não só ao me ensinar os fundamentos da programação, mas pelas discussões científicas, sempre bem regadas!

Ao Luke Baton, pela paciência com que escutava às minhas inúmeras “consultas científicas informais” sobre os mais variados temas e pela dedicação ao responder cada uma de maneira excepcional.

À equipe da biblioteca do Centro de Pesquisas René Rachou, pelo suporte para o acesso a diversos artigos científicos que enriqueceram, em muito, este trabalho.

À todos os participantes e organizadores do XIV Seminário Laveran & Deane sobre Malária (Ilha de Itacuruçá*/RJ, 2009).

Aos Doutores Jerônimo C. Ruiz, Gustavo F. Vieira, Fabiano S. Pais, Francisco P. Lobo e Carlos Eduardo C. Silva, membros da banca examinadora, por aceitar o convite e pelas sugestões e contribuições que ajudaram a enriquecer este trabalho.

À FAPEMIG e ao Centro de Pesquisas René Rachou – CPqRR/FIOCRUZ, pelos recursos financeiros que viabilizaram o desenvolvimento deste Projeto. Ao CNPq pela bolsa de estudos.

À CAPEs pela bolsa de estudos.

Aos meus grandes amigos (que não cito nominalmente por medo de cometer injustiças, mas eles sabem quem são!), sem os quais todo o resto perde a graça.

À minha família. Às minhas queridas irmãs Paula, Tânia, Moema e Laura por todo o carinho e por uma convivência que a cada dia se torna mais prazerosa e engraçada. Aos meus sobrinhos e sobrinhas, pelas alegrias que trazem. Às minhas queridas primas Alice e Marina, pelo companheirismo fraterno. Ao tio Luiz e à tia Mônica pela atenção especial que sempre tiveram comigo, me tratando como um verdadeiro filho. Ao meu pai pelo exemplo e por todo o seu esforço em garantir uma educação de qualidade aos filhos. E um obrigado especial à minha mãe, por sua dedicação supra-humana, por todo o seu esforço, sua doação e sua dedicação, em todos os momentos!

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

O peptídeo sinal é um motivo encontrado, geralmente, na extremidade N-terminal de proteínas e a sua presença determina a entrada na via clássica de transporte intracelular, após a translocação da proteína para o lúmen do retículo endoplasmático. Portanto, a presença ou ausência do peptídeo sinal influencia a função biológica de uma proteína ao ser um fator determinante da sua localização subcelular. Como a conservação de função entre proteínas ortólogas é esperada, foi hipotetizado que a localização subcelular e, consequentemente, a presença do peptídeo sinal deveriam, também, se apresentar conservadas. Partindo desta premissa, as predições de peptídeo sinal em proteínas ortólogas de cinco espécies de *Plasmodium* foram analisadas.

Predições de peptídeo sinal (SignalP) e informações de ortologia (OrthoMCL-DB) para proteínas de cinco espécies do gênero *Plasmodium* (*Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium knowlesi*, *Plasmodium berguei* e *Plasmodium yoelii*) foram combinadas em uma estratégia inovadora, visando a identificação de grupos de proteínas ortólogas que apresentam predições de peptídeo sinal divergentes (grupos Mistos). As proteínas pertencentes a estes grupos foram submetidas a uma análise comparativa baseada na inspeção visual de alinhamentos múltiplos e de modelos gênicos e regiões genômicas flanqueadoras da extremidade N-terminal. Novos modelos gênicos foram sugeridos para aquelas proteínas que apresentavam prováveis erros de anotação de sequência, especialmente na região N-terminal. Alguns dos novos modelos gênicos foram validados por RT-PCR. Os resultados da inspeção visual foram usados para treinar uma Máquina de Suporte de Vetores (Support Vector Machine) com o objetivo de classificar grupos Mistos em: **(1)** Com erros de anotação ou **(2)** Sem erros de anotação. O SVM foi aplicado para classificar os grupos Mistos de cinco bancos de dados, montados a partir de vinte e duas espécies.

Os grupos contendo proteínas com predições de peptídeo sinal divergentes apresentaram uma alta taxa de erros de anotação. Um total de 478 proteínas de *Plasmodium* foram reanotadas sendo que a maioria apresentou inversões das suas predições de peptídeo sinal originais, representando um impacto significativo no conjunto final de proteínas destinadas à via clássica de transporte intracelular, principalmente para *Plasmodium vivax* e *Plasmodium yoelii*. O classificador baseado nos dados da inspeção visual se mostrou bastante flexível e robusto, apresentando uma performance boa e consistente mesmo frente a cenários variados de agrupamento de espécies.

A metodologia proposta introduz uma abordagem simples, porém promissora, para a realização de tarefas de curadoria e controle de qualidade dos dados de anotação de sequências proteicas em uma escala genômica. Os resultados do classificador definem a

base para seu desenvolvimento em uma ferramenta computacional e os resultados das reanotações em *Plasmodium* impactarão a busca por novos alvos vacinais e quimioterápicos.

ABSTRACT

Signal peptide is a motif usually found in the N-terminal end of proteins and its presence directs proteins to enter the classical intracellular transport pathway, after their co-translational translocation to the endoplasmic reticulum lumen. Therefore, the presence or absence of a signal peptide plays an indirect role in defining the biological function of a protein, as it is a determinant of subcellular localization. Since function is usually conserved among orthologous proteins, it has been hypothesized that subcellular localization and, consequently, signal peptide status are expected to behave accordingly. Based on this premise, signal peptide predictions among orthologous proteins from five *Plasmodium* species were analyzed.

Signal peptide predictions (SignalP) and orthology information (OrthoMCL-DB) for proteins from five *Plasmodium* species (*Plasmodium falciparum*, *Plasmodium vivax*, *Plasmodium knowlesi*, *Plasmodium berguei* and *Plasmodium yoelii*) were combined into an innovative strategy, intending the identification of groups of orthologous proteins showing diverging signal peptide predictions (Mixed groups). The proteins belonging to these groups were submitted to a comparative analysis based on visual inspection of multiple alignments and of gene models and their upstream flanking regions. New gene models were proposed for those proteins presenting putative sequence misannotations, especially in their N-terminal region. Some of the new gene models were validated through RT-PCR. Results from the visual inspection were used to train a Support Vector Machine to be able to classify Mixed groups into: **(1)** With misannotations and **(2)** Without misannotations. The SVM was applied in the classification of Mixed groups from five datasets, built from twenty-two species.

Groups featruing proteins with diverging signal peptide predictions showed an elevated rate of misannotations. A total of 478 *Plasmodium* proteins were reannotated, and most had their original signal peptide predictions inverted, representing a significant impact in the final set of proteins destined to the classical intracellular transport pathway, especially for *Plasmodium vivax* and *Plasmodium yoelii*. The classifier based on the visual inspection data was shown to be flexible and robust, performing well and consistently even when dealing with highly ecletic species clusterings.

The proposed methodology introduces a simple yet promising approach to the tasks of curation and quality control of annotation data from proteins sequences in genomic scale. The classifier's results define the groundwork for its development into a computational tool and the reannotations results for *Plasmodium* proteins shall impact the search for new vaccine and drug targes.

1. INTRODUÇÃO

1.1. Predição gênica, ortologia e localização subcelular de proteínas

O sequenciamento de ácidos nucléicos revolucionou a pesquisa biológica. Nas últimas três décadas registrou-se, em paralelo com o avanço tecnológico das metodologias de sequenciamento, o crescimento exponencial do número de sequências depositadas em bancos de acesso público como o GenBank, que atualmente hospeda mais de 150 milhões de sequências individuais. Além disso, o sequenciamento de genomas completos, impulsionado pelo advento das técnicas de sequenciamento em larga escala que produzem um enorme volume de dados que aumentam a cobertura dos genomas sequenciados e reduzem drasticamente o tempo necessário para a montagem e anotação, tem sido o principal responsável por esta expansão. Atualmente, mais de 3.400 projetos genoma, de espécies e isolados, já estão completos e mais de 13.000 estão em andamento (http://www.genomesonline.org).

Obter a informação de sequencias é apenas a primeira etapa de um projeto genoma. Um dos grandes desafios da era genômica é predizer, a partir da sequência nucleotídica, a função biológica que um fragmento de ácido nucleico ou a proteína que ele codifica irá exercer no contexto biológico do organismo em estudo. Portanto, após a montagem do genoma, uma das principais etapas da anotação do genoma é a predição dos genes. Considerando-se a velocidade crescente com que novas sequências são geradas e a complexidade dos genomas, a anotação manual de cada gene se torna inviável e a anotação automatizada se torna a única escolha possível.

Vários programas preditores têm sido desenvolvidos para realizar a identificação de genes a partir de sequências genômicas. Estes programas se dividem em duas grandes categorias, os métodos intrínsecos (ou *ab initio*) que se restringem à informação contida na sequência fornecida e procuram por padrões estatísticos que definem regiões internas externas e os limites de um gene, e os métodos extrínsecos ou baseados em similaridade que utilizam análises comparativas e fazem uso de informações complementares como dados de expressão. Geralmente, os dois métodos são combinados para que se alcance um alto desempenho na anotação em larga escala (DO; CHOI, 2006; WINDSOR; MITCHELL-OLDS, 2006).

Porém, a identificação gênica por métodos computacionais, principalmente para organismos eucariotos, ainda apresenta muitos desafios. A baixa densidade gênica em genomas eucariotos e a estrutura complexa dos genes, que podem apresentar múltiplos exons e splicing alternativos, aumentam consideravelmente a dificuldade em se identificar modelos gênicos precisos (DO; CHOI, 2006). Além disso, as predições de exons iniciais e terminais são ainda mais desafiadoras quando comparadas à identificação de exons internos, pois os sítios de início da transcrição e os sinais de parada são menos

conservados que os sítios de splicing e, portanto, mais difíceis de serem preditos corretamente (BERNAL et al., 2007).

Com o número de genomas disponíveis crescendo exponencialmente, a análise de novos genomas tem sido cada vez mais baseada em estudos comparativos e, apesar de se tratar de uma abordagem eficaz, um dos seus pontos negativos é a propagação de erros ou informações imprecisas entre genomas e bancos de dados (LINIAL, 2003). Portanto, concomitantemente com o avanço das tecnologias de sequenciamento, das metodologias de predição e anotação, da implementação dos bancos de dados para armazenamento e acesso à informação biológica, deve haver um esforço para garantir a qualidade e a consistência desta informação, através do emprego constante de mecanismos que sejam capazes de identificar e corrigir erros pontuais ou sistemáticos nos processos de anotação.

A base das análises genômicas comparativas é a ancestralidade comum entre os organismos, mais especificamente as relações de homologia entre genes ou demais elementos dos genomas. Portanto, é extremamente importante que os diferentes tipos de relação intergênicas sejam claramente conceitualizados. Quando ocorre um evento de especiação, os genes originários de um mesmo gene ancestral são denominados ortólogos. Quando o processo de divergência entre os genes é iniciado após um evento de duplicação gênica, os genes resultantes são denominados parálogos (FITCH, 1970; KOONIN, 2005).

Por divergirem a partir de um evento de especiação, a filogenia de genes ortólogos representa exatamente a filogenia de seus organismos (FITCH, 2000), e existe um paradigma dominante de que ortólogos geralmente apresentam uma maior conservação funcional quando comparados a parálogos (PETERSON et al., 2009). Devido a estas características, genes ortólogos são componentes essenciais de diversas análises e aplicações no campo da genômica comparativa como, por exemplo, a reconstrução das relações evolutivas entre espécies, os testes de modelos evolutivos para genomas, as inferências sobre propriedades funcionais de genes, a identificação de genes taxonomicamente restritos e a anotação de genomas. Portanto, determinar qual a verdadeira relação de homologia entre duas ou mais sequências é reconhecidamente um dos maiores e mais importantes desafios da biologia computacional (DESSIMOZ et al., 2012), a ponto de ter sido formado um consórcio internacional (http://www.questfororthologs.org) para organizar e estruturar este campo de pesquisa (GABALDÓN et al., 2009).

Atualmente, existem dezenas de bancos de dados especializados em informações sobre a ortologia entre genes de dois ou mais genomas. Cada banco utiliza seus próprios métodos inferenciais para a montagem de grupos de ortólogos. Figurando entre os principais recursos disponíveis encontra-se o OrthoMCL-DB (LI; STOECKERT; ROOS, 2003), que utiliza um algoritmo de agrupamento de Markov (*Markov Clustering Algorithm*-MCL) para realizar o agrupamento de genes de múltiplos organismos, a partir de uma matriz

dos resultados de analises de similaridade de sequência por BLAST. O programa OrthoMCL define genes ortólogos como os melhores resultados recíprocos do BLAST entre dois genomas e também agrupa parálogos “recentes” (genes de um mesmo genoma cuja similaridade recíproca é maior do que a similaridade com qualquer gene em um segundo genoma). Em sua última versão (versão 5.0), o OrthoMCL-DB realizou o agrupamento das proteínas de 150 espécies criando mais de 120.000 grupos ortólogos com 1.398.546 sequências de proteínas.

Como dito anteriormente, o conceito de conservação funcional entre proteínas ortólogas é amplamente utilizado para a anotação de genes em análises de genômica comparativa. Apesar de ainda haver controvérsia em relação ao quão abrangente é esta conservação e se realmente existem diferenças entre ortólogos e parálogos, este paradigma ainda é considerado um dos pilares das metodologias de anotação automatizada (DESSIMOZ et al., 2012). Portanto, quando o ortólogo de um gene, cujo papel biológico já foi descrito, é identificado em um novo genoma, a sua anotação funcional é, geralmente, transferida para o novo gene.

A definição de papel biológico ou função biológica de um gene que foi empregada neste trabalho é uma definição ampla, que engloba não somente as interações moleculares diretas nas quais o seu produto proteico se envolve ou reações que catalisa (no caso de enzimas), mas considera também o momento e, principalmente, o microambiente subcelular nos quais estas interações ocorrem. Portanto, a conservação funcional de um gene está diretamente relacionada com a localização final do seu produto proteico (MAK; WANG; KUNG, 2011; NAIR; ROST, 2002).

O direcionamento de proteínas para os diversos compartimentos subcelulares e também para o meio extracelular é essencial a todos os organismos (KOCH; MOSER; MÜLLER, 2003). Geralmente, este endereçamento de proteínas é dependente de sinais intrínsecos encontrados nas sequências proteicas (BLOBEL; DOBBERSTEIN, 1975). Dentre estes sinais, o mais conhecido e bem estudado é o peptídeo sinal, responsável pela translocação de proteínas através da membrana plasmática (em procariotos) ou através da membrana do Retículo Endoplasmático Rugoso (em eucariotos). O peptídeo sinal ou sequência sinal é formado por uma porção central hidrofóbica flanqueada na sua extremidade C-terminal por uma região polar contendo, na maioria das vezes, resíduos de prolina e glicina e resíduos sem carga que determinam o seu sítio de clivagem, e na sua extremidade N-terminal por uma região polar com carga positiva (**Figura 1**) (VON HEIJNE, 1990). Geralmente, o peptídeo sinal se localiza na extremidade N-terminal de proteínas e em muitos casos inicia a translocação através de membrana do Retículo Endoplasmático Rugoso (RER) ao mesmo tempo em que a proteína é traduzida por ribossomos (SHAN; WALTER, 2005), entretanto há casos em que o peptídeo sinal é encontrado em regiões mais

centrais da proteína ou mesmo em sua extremidade C-terminal (KUTAY; AHNERT-HILGERL; HARTMANN, 1995).

Figura 1: Estrutura típica do peptídeo sinal. O peptídeo sinal geralmente se localiza na extremidade N-terminal de proteínas pré-processadas, sendo formado por uma região polar com carga positiva na sua extremidade N-terminal, uma região central hidrofóbica e uma região polar que determina o seu sítio de clivagem.

Células eucarióticas apresentam diferentes compartimentos envolvidos por membranas que exercem funções especializadas. Porém, a maioria das proteínas é sintetizada no citoplasma da célula e precisam ser direcionadas ao compartimento correto para que cumpram o seu papel biológico (**Figura 2**). A presença de um peptídeo sinal em uma proteína determinará a sua translocação para o lúmen do RER, de onde ela poderá seguir para compartimentos como, por exemplo, o complexo de golgi, retículo endoplasmático, lisossomos, vacúolo digestivo, vesículas secretórias (proteínas que são destinadas ao meio extracelular) e diferentes membranas plasmáticas (TASHIRO, 1983). A ausência de um peptídeo sinal, por sua vez, indica uma localização citoplasmática, nuclear ou, em casos mais raros, que o transporte desta proteína para outros compartimentos é feito através de um via independente de peptídeo sinal (CLEVES, 1997; NICKEL, 2003).

Organelas como a mitocôndria e cloroplastos, mesmo apresentando uma maquinaria própria para a síntese das proteínas que são codificadas em seus genomas, precisam importar a maioria das suas proteínas, que são codificadas no genoma nuclear e sintetizadas no citoplasma. O transporte de proteínas para a mitocôndria também é dependente de um sinal encontrado na sequência proteica, porém a sua estrutura é diferente (GAVEL; NILSSON; VON HEIJNE, 1988). O transporte para o lúmen dos tilacóides em cloroplastos é mediado por um motivo bipartido, formado por um peptídeo de trânsito seguido por um sinal que compartilha características estruturais com um peptídeo sinal clássico (VON HEIJNE, 1990).

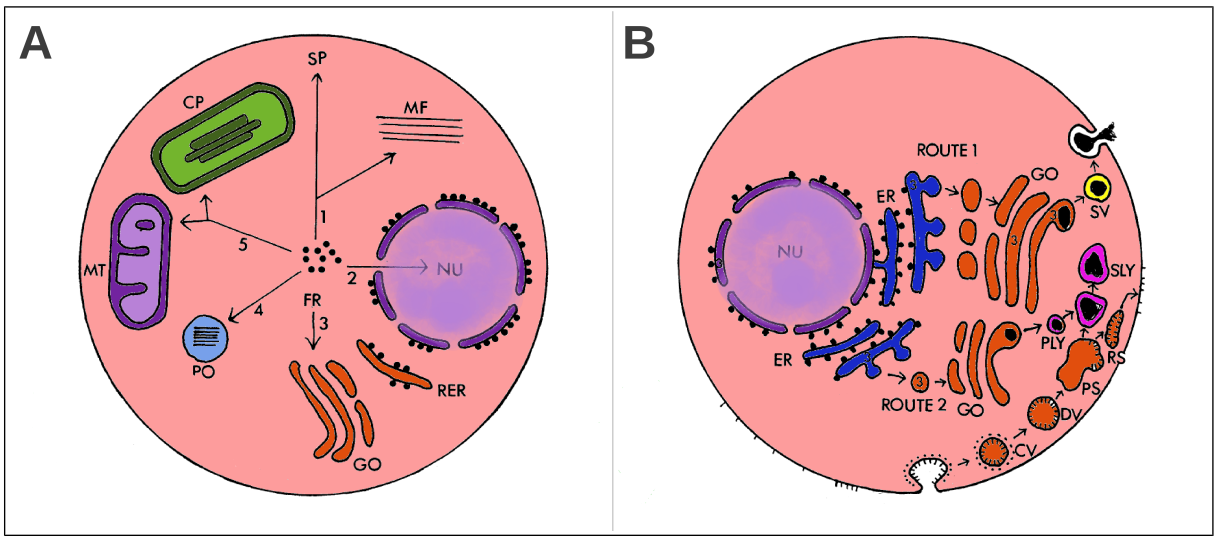


Figura 2: Transporte intracelular de proteínas em uma célula eucariótica. (A) Proteínas que não apresentam peptídeo sinal são sintetizadas em ribossomos livres no citoplasma e têm como destinos mais comuns o próprio citoplasma, na forma de proteínas solúveis (SP) ou insolúveis, estruturas do citoesqueleto (MF), o núcleo (NU), o lado citoplasmático de membranas, peroxissomos (PO), mitocôndrias (MT) e cloroplastos (CP). (B) Proteínas que apresentam um peptídeo sinal são sintetizadas por ribossomos ligados à membrana do retículo endoplasmático rugoso e são translocadas através desta membrana para o lúmen do RER em um processo co-traducional. Estas proteínas são destinadas a organelas envoltas por membranas, tais como o golgi (GO) e lisossomos (PLY), para vesículas secretórias (SV) e para a superfície celular. Adaptada de TASHIRO, 1983.

Portanto, a definição sobre a presença/ausência de peptídeo sinal em uma proteína é uma informação importante para auxiliar a identificação de sua função. E este conhecimento tem aplicações práticas, uma vez que proteínas que são secretadas ou estão expostas na superfície celular são especialmente interessantes para a indústria farmacêutica, pois são alvos mais acessíveis para estratégias de intervenção molecular como o desenvolvimento de vacinas e drogas (GOODSWEN; KENNEDY; ELLIS, 2012).

Demonstrações experimentais da localização subcelular de proteínas exigem metodologias caras e que demandam tempo, não permitindo que as validações acompanhem a velocidade na qual novos genomas e novas proteínas são descritos. Portanto, mais uma vez metodologias computacionais de predição de propriedades proteicas são alternativas para viabilizar o estudo funcional de genomas em larga escala. A predição de peptídeo sinal, assim como a predição mais refinada de diferentes localizações subcelulares, são tópicos bastante explorados pela biologia computacional, justamente por causa da relevância que o tema apresenta nos contextos de anotação funcional de genomas e de busca por novos alvos moleculares (SCHNEIDER; FECHNER, 2004). Existe uma grande variedade de métodos preditivos disponíveis, entretanto, estas diferentes abordagens podem ser divididas em quatro grandes classes metodológicas (MAK; WANG; KUNG, 2011): predições baseadas (i) exclusivamente na análise da sequência de aminoácidos; (ii) em propriedades globais da sequência; (iii) em homologia; e (iv) em informações adicionais à sequência proteica. Vale ressaltar que existem métodos que combinam mais de uma estratégia.

O programa SignalP (NIELSEN et al., 1997) é um exemplo de preditor baseado na análise da estrutura primária em busca de sinais de transporte. O SignalP foi um dos primeiros preditores a usar metodologias de aprendizado de máquinas, empregando Redes Neurais (*Neural Networks*) em sua versão inicial. Posteriormente, foi adicionado um Modelo Oculto de Markov (*Hidden Markov Model*) para melhorar a predição de proteínas cujo peptídeo sinal não era clivado e que, portanto, permaneciam ancoradas à membrana. Em sua terceira versão (BENDTSEN et al., 2004), foi incluída a análise da composição de aminoácidos da sequência completa, uma propriedade global das proteínas. O dado de entrada para a predição com o SignalP é a sequência de aminoácidos de um proteína, sendo que, por padrão, somente os 70 primeiros aminoácidos são considerados, o que torna o resultado da análise extremamente sensível à predição do modelo gênico e à anotação da sequência proteica.

1.2. Máquinas de Vetores de Suporte (SVMs) e suas aplicações biológicas

A crescente acumulação de dados biológicos provenientes das novas tecnologias de alto rendimento (sequenciamento de ácidos nucleicos ou de proteínas, microarranjos, interações de proteínas por duplo híbrido, etc) cria possibilidades inéditas de investigações dos fenômenos biológicos, principalmente através de abordagens computacionais. Porém, explorar estes dados requer metodologias analíticas sofisticadas, capazes de relacionar e processar grandes volumes de dados.

Uma tarefa frequente em questões biológicas é a classificação de entes em classes ou categorias através do reconhecimento de padrões distintos. As técnicas de aprendizado de máquina são ferramentas que realizam esta tarefa, e entre elas destacam-se as Máquinas de Vetores de Suporte (*Support Vector Machines*), pois apresentam alta precisão, processam dados multidimensionais, realizam classificações não lineares e aceitam diversos tipos de dados como entrada (BEN-HUR et al., 2008; BOSER; GUYON; VAPNIK, 1992). Uma Máquina de Vetores de Suporte é treinada a partir de um conjunto e dados conhecidos contendo exemplos positivos e negativos (no caso de um classificador binário) e a classificação é alcançada com a definição de uma função que descreve um hiperplano dividindo o espaço amostral em duas áreas. Os dados mais próximos do hiperplano, e que definem a margem de separação, são deniminados vetores de suporte.

Considerando-se especificamente problemas de teor biológico, SVMs já foram empregados, por exemplo, para predições de epitopos de células B ou T (GOODSWEN; KENNEDY; ELLIS, 2012), sítios de clivagem de peptídeo sinal (BENDTSEN et al., 2004), localização subcelular de proteínas (BLUM; BRIESEMEISTER; KOHLBACHER, 2009; CHERIAN; NAIR, 2010; SHATKAY et al., 2007) e para mineração de dados da literatura em busca de interações entre proteínas (DONALDSON et al., 2003). Estas aplicações

demonstram a versatilidade dá técnica de Máquinas de Vetores de Suporte e justificam o seu amplo uso na biologia computacional.

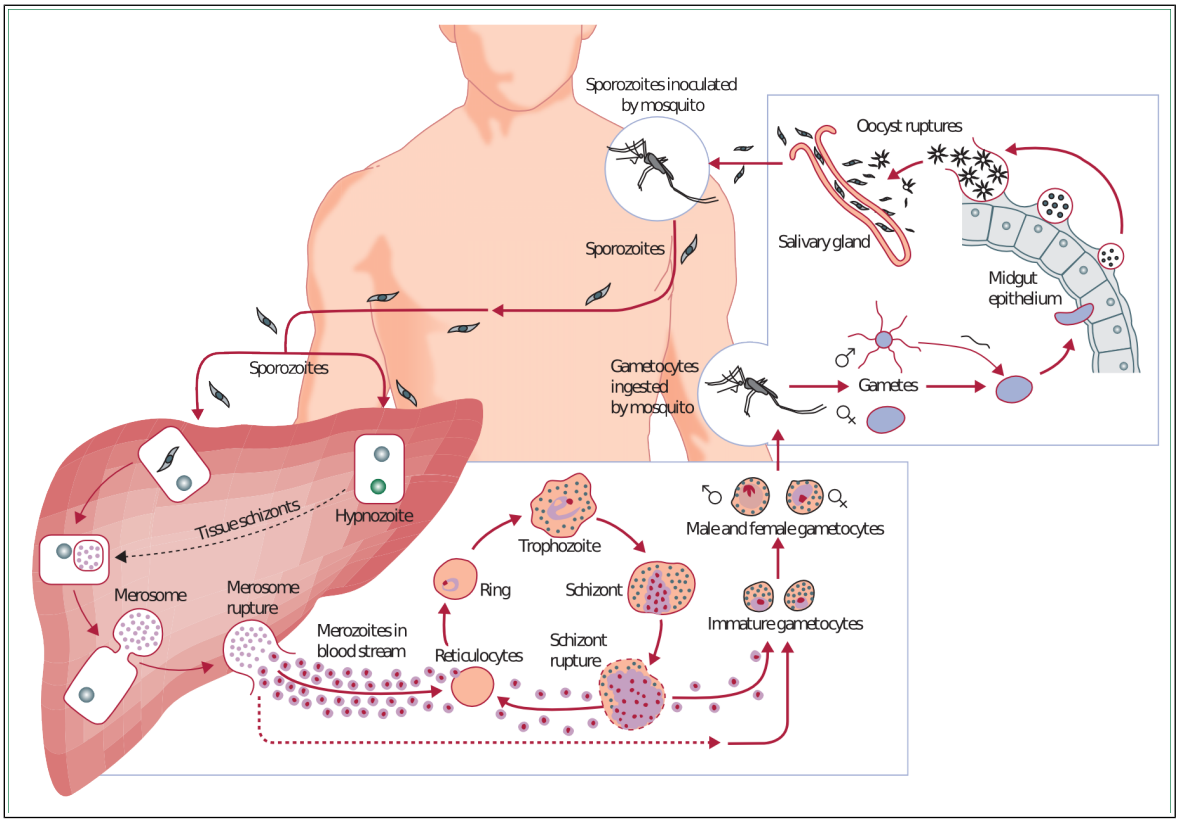
1.3. Malária

A malária é causada por protozoários do filo Apicomplexa, família Plasmodiidae e gênero *Plasmodium.* Atualmente, são conhecidas aproximadamente 150 espécies do parasito que infectam diferentes hospedeiros vertebrados, sendo quatro espécies as principais causadoras da malária no homem: *Plasmodium falciparum, Plasmodium vivax, Plasmodium malariae* e *Plasmodium ovale*. Recentemente foi demonstrado que uma quinta espécie, o *Plasmodium knowlesi*, que naturalmente infecta primatas não humanos, tem causado infecções graves em humanos, principalmente, em algumas áreas específicas do sudeste asiático e do continente africano (COX-SINGH et al., 2008; ONG et al., 2009; WHITE, 2008).

As relações evolutivas entre as diferentes espécies do gênero *Plasmodium* têm sido alvos de vários estudos ao longo de décadas, pórem, apesar de vários avanços, o tema ainda guarda algumas controvérsias, principalmente acerca da origem e da idade de *Plasmodium falciparum*. Alguns clados monofiléticos dentro do gênero como, por exemplo, *P. falciparum/Plasmodium reichenowi* (espécie que parasita chimpanzés), *P. vivax/P. Knowlesi* e o dos parasitos de roedores *P. berghei/P. yoelli/P. Chabaudi* têm suporte em vários estudos (ESCALANTE; AYALA, 1994; LECLERC et al., 2004; OLLOMO et al., 2009; PERKINS; SCHALL, 2002; SILVA; EGAN, 2010; WATERS; HIGGINS; MCCUTCHAN, 1991). Estudos também apoiam as hipóteses da origem africana do *Plasmodium falciparum* e dos parasitos de roedores, assim como a origem asiática das demais espécies que parasitam o homem (CORNEJO; ESCALANTE, 2006; ESCALANTE; AYALA, 1994; ESCALANTE et al., 2005; LECLERC et al., 2004; WATERS; HIGGINS; MCCUTCHAN, 1991). A maior controvérsia sobre a história evolutiva em *Plasmodium* é exatamente a origem de *P. falciparum* e o período do início do seu parasitismo em humanos. Enquanto alguns estudos defendem a hipótese de um parasitismo recente (centenas ou dezenas de milhares de anos), após um evento de transferência lateral entre hospedeiros (KRIEF; ESCALANTE; PACHECO, 2010; RICH et al., 2009; WATERS; HIGGINS; MCCUTCHAN, 1991), outros suportam a hipótese de co-especiação entre parasitos (*P. falciparum/P. reichenowi*) e hospedeiros vertebrados no momento da separação entre homens e chimpanzés (entre 5-7 milhões de anos atrás) (ESCALANTE; AYALA, 1994). Um trabalho recente (SILVA; EGAN, 2010), que defende a hipótese de co-especiação, usou o tempo de divergência de 5-7 milhões de anos entre *P. falciparum/P. reichenowi* como calibração e calculou as seguintes estimativas para o tempo de divergência entre: *P. vivax/P.knowlesi* em 15-46 milhões de anos, correspondendo com a expansão da linhagem de primatas do Velho Mundo; os

parasitos de roedores em 25-33 milhões de anos, em concordância com a origem da família Muridae, os hospedeiros naturais destas espécies.

A malária humana é uma das mais importantes causas de mortalidade e morbidade no mundo. A doença é endêmica na África subsaariana, no sudeste Asiático e em áreas da América do Sul, provocando mais de 240 milhões de casos e causando cerca de 800.000 mortes anuais (WORLD HEALTH ORGANIZATION, 2011). A África concentra 90% dos casos de malária do mundo sendo que a mortalidade é maior do que em outras regiões devido, principalmente, ao limitado acesso ao tratamento nas vilas. Nos últimos anos, a doença vem se propagando para outras áreas devido ao intenso e não planejado processo de urbanização, e estima-se que este processo na África subsaariana e em outros continentes, esteja alterando profundamente a epidemiologia da malaria nestas áreas (SIRI et al., 2008; TATEM; SMITH, 2010). O *P. falciparum* é a espécie mais patogênica e o principal responsável pela maioria das causas de morte e morbidade neste continente (HAY et al., 2010; SNOW et al., 2005).



O ciclo biológico dos parasitos da malária humana compreende uma fase de reprodução sexuada, que ocorre dentro do hospedeiro invertebrado, e outra de reprodução assexuada, que se desenvolve no hospedeiro vertebrado (**Figura 3**). **Figura 3: Ciclo de vida de parasitos do gênero *Plasmodium***. A inoculação de esporozoítos através da picada do mosquito fêmea inicia a infecção do hospedeiro vertebrado. Os esporozoítos invadem células hepáticas e se transformam em formas replicativas (esquizontes) ou, especificamente no caso de *Plasmodium vivax* (e outras espécies relacionadas), podem se transformar em formas dormentes (hipnozoítos). Os

merozoítos provenientes dos esquizontes hepáticos invadem os eritrócitos e dão início à fase eritrocítica da doença, responsável pelas manifestações clínicas. Os parasitos se reproduzem, dentro das hemácias, em um ciclo assexuado de aproximadamente 48 horas de duração e os novos merozoítos produzidos reiniciam o ciclo de invasão de hemácias. Uma parcela dos parasitos se diferencia em formas sexuais, os gametócitos (macro e micro). Quando os gametócitos são ingeridos durante o repasto sanguíneo dão início à fase sexuada do ciclo de vida, que ocorre no hospedeiro invertebrado. Os gametócitos se desenvolvem em gametas, ocorre a fecundação e a colonização do mosquito pelo oocineto. O oocineto se aloja na membrana basal do epitélio intestinal e se transforma em um oocisto. Dentro do oocisto ocorre o processo de esporogonia que dará origem a centenas de esporozoítos que invadirão a glândula salivar do mosquito. Fonte: (MUELLER et al., 2009).

Durante o repasto sanguíneo no hospedeiro vertebrado, a fêmea de mosquito infectada deposita, em média, 100 esporozoítos por picada (JIN; KEBAIER; VANDERBERG, 2007), que permanecem no sítio da picada por aproximadamente 1 a 3 horas até alcançarem os vasos sanguíneos (AMINO et al., 2006; YAMAUCHI et al., 2007). Alguns autores descreveram uma rota de migração dos esporozoítos, em que os mesmos, após atravessarem o epitélio do hospedeiro vertebrado, podem também atingir o sistema linfático (YAMAUCHI et al., 2007). Entretanto, os parasitos não parecem atingir o fígado por esta via (AMINO et al., 2006). Recentemente, um estudo demonstrou que uma fração dos esporozoítos inoculados podem se diferenciar em merozoítos na epiderme, na derme e também nos folículos pilosos, entretanto merozoítos derivados da pele, em condições normais, não contribuem significativamente para a infecção dos eritrócitos (GUEIRARD et al., 2010).

Os esporozoítos que chegam ao sistema circulatório atingem o fígado, onde infectam hepatócitos. Na malária de mamíferos, não está claro o mecanismo pelo qual os esporozoítos passam dos capilares sinusóides do fígado até os hepatócitos: se é através das células de Kupffer ou através das células do endotélio dos vasos sanguíneos (MOTA; HAFALLA; RODRIGUEZ, 2002; PRADEL; FREVERT, 2001). Mais recentemente, utilizando o modelo de malária murina, foi demonstrado que uma parte dos esporozoítos podem atingir o fígado sem passar pelas células de Kupffer (GUEIRARD et al., 2010). O processo de invasão de hepatócitos é complexo e depende de várias interações do tipo ligante-receptor. Recentemente foi demonstrado que os esporozoítos invadem vários hepatócitos, migrando através deles, antes de se desenvolverem dentro de um hepatócito (MOTA; RODRIGUEZ, 2004). Nas infecções por *P. vivax* e *P. ovale* alguns parasitos se desenvolvem rapidamente nos hepatócitos, enquanto outros permanecem em estado de latência no fígado, estes são denominados hipnozoítos e são as responsáveis pelos casos de recaídas (KROTOSKI et al., 1982). Uma vez dentro dos hepatócitos os esporozoítos se diferenciam em trofozoítos que, após sofrerem várias divisões por esquizogonia, formam os esquizontes. Os esquizontes maduros liberam os merozoítos teciduais através de um processo de brotamento de

vesículas, os merosomas, que após atingirem a corrente sanguínea, repletos de parasitos, liberam os merozoítos (STURM et al., 2006). Os merozoítos liberados invadem as hemácias iniciando assim a fase eritrocítica. Para que o merozoíto invada o eritrócito é necessário que haja também o reconhecimento de receptores específicos (HADLEY, 1986; HOWARD; MILLER, 1981). Após varias gerações de merozoítos sanguíneos, alguns se diferenciam dando origem às formas sexuadas, os gametócitos masculinos e femininos, que amadurecem sem sofrer divisão celular. Ao serem ingeridos pelo mosquito susceptível durante o repasto sanguíneo, inicia-se a jornada do parasita no hospedeiro invertebrado onde ocorrerá o ciclo sexuado ou esporogônico.

Dentro do estômago do mosquito, os gametócitos masculinos e femininos se diferenciam transformando-se em gametas, sob influência das condições do ambiente em que se encontra no hospedeiro invertebrado, bem como fatores internos do mosquito (BILLKER et al., 1998, 2004). Logo após o repasto sanguíneo, macrogametócitos e microgametócitos escapam das células vermelhas dando origem aos macrogametas e microgametas, respectivamente. Dentro dos próximos trinta minutos ocorre a fecundação dos gametas e a formação do zigoto, com morfologia esférica. Nas vinte e quatro horas seguintes o zigoto se transformará no oocineto, uma célula alongada, com diferenciação antero-posterior e capaz de realizar movimentos amebóides (SINDEN, 1999). O oocineto por sua vez atravessa a matriz peritrófica (membrana que envolve o bolo alimentar) e, por um mecanismo transcelular (BATON; RANFORD-CARTWRIGHT, 2004; VLACHOU et al., 2004; ZIELER; DVORAK, 2000), atinge as células do intestino médio, onde se aloja entre o epitélio e a membrana basal. Então o parasito se encista, e passa a denominar-se oocisto. Inicia-se o processo de multiplicação esporogônica, e em aproximadamente duas semanas, a parede do mesmo se rompe liberando esporozoítos que invadem a hemolinfa do inseto. Assim, muitos parasitos migrarão até atingir as glândulas salivares, sendo inoculados em outro hospedeiro vertebrado, completando o ciclo evolutivo dos plasmódios.

1.4. Transporte de proteínas em *Plasmodium*

O transporte intracelular de proteínas nas espécies do Filo Apicomplexa apresenta características peculiares e é essencial ao estilo de vida adotado por estes organismos. Os organismos deste grupo são, em sua maioria, parasitos intracelulares obrigatórios que ao invadirem a célula hospedeira formam um vacúolo parasitóforo, onde permanecerão durante o seu desenvolvimento (LINGELBACH; JOINER, 1998). A invasão e o estabelecimento do vacúolo parasitóforo são mediados por organelas especializadas encontradas em um complexo no polo apical das formas invasoras, caracterizando a estrutura que dá nome ao Filo. Este complexo apical é formado por três tipos de organelas, as roptrias, os micronemas e os grânulos densos, que abrigam proteínas diretamente envolvidas tanto nas várias etapas de invasão quanto na constituição e manutenção do vacúolo parasitóforo. Justamente devido ao seu envolvimento com as etapas de invasão e seus mecanismos de ação, as proteínas liberadas por estas organelas são consideradas alvos prioritários para o desenvolvimento de vacinas e drogas que bloqueiem a invasão (KATS et al., 2006).

A maior parte das proteínas destinadas a estas organelas especializadas apresentam peptídeos sinal (KATS et al., 2008) (**Figura 4**). Experimentos de imunofluorescência demonstraram que uma proteína destinada às roptrias trafega pelos compartimentos da via secretória antes de alcançar seu destino, e o transporte desta proteína foi bloqueado após tratamento com brefeldina A, um inibidor clássico do transporte de proteínas pelo RER e golgi (TOPOLSKA et al., 2004).

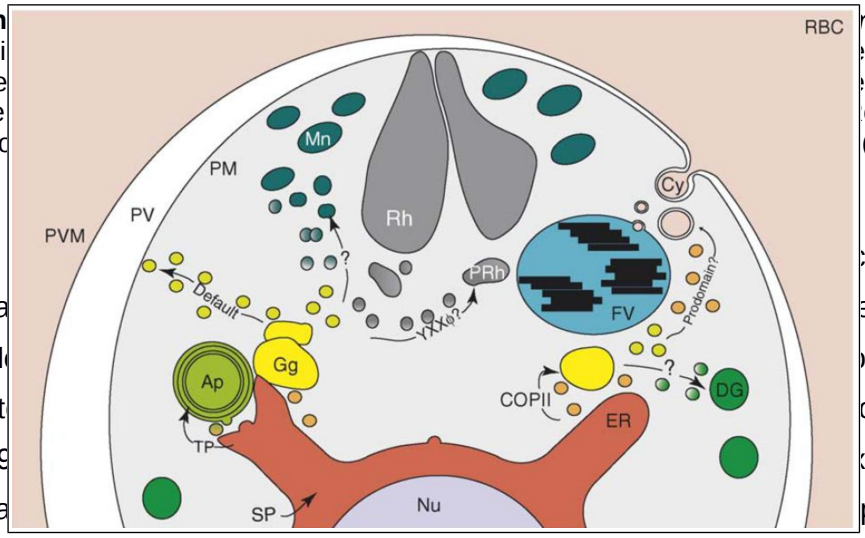


Figura 4: Tran ***n.*** Esquema de um parasito (ci ermelha (RBC). Proteínas apre ndoplasmático (ER) antes de optrias (Rh) e granulos denso (Ap) e vacúolo digestivo (FV).

Além cessária para viabilizar o tra extracelulares durante o cicl de proteínas ao apicoplast ossintetizante (KOHLER, 19 erce funções essenciais pa presenta vias metabólicas exclusivas. O genoma circular do apicoplasto tem apenas 35 kilobases, sendo o menor genoma de plastídeos conhecido, e codifica somente 30 genes (WILSON et al., 1996). Entretanto, calcula-se que aproximadamente 500 proteínas (10% das proteínas totais) codificadas nos genomas nucleares de espécies de *Plasmodium* têm como destino o apicoplasto. O direcionamento ao apicoplasto é mediado por um sinal bipartido formado por um peptídeo sinal clássico, que permite a translocação para o lúmen do RER, seguido por um peptídeo de trânsito, que é responsável pelo reconhecimento e interiorização da proteína pela organela (WALLER et al., 2000).

Em *Plasmodium*, peptídeos sinal também estão presentes em todas as proteases destinadas ao vacúolo digestivo (TONKIN et al., 2006), a organela responsável pela proteólise da hemoglobina e a subsequente detoxificação dos grupos heme resultantes por cristalização e formação da hemozoína, uma função metabólica especializada que é alvo de intervenções quimioterápicas.

Outra característica típica relacionada ao ciclo biológico de espécies de *Plasmodium* é a remodelagem de eritrócitos parasitados (GOLDBERG; COWMAN, 2010). Na fase eritrocítica de seu desenvolvimento, as espécies deste gênero evitam alguns mecanismos da resposta imune do hospedeiro ao se refugiarem, envoltas por seus vacúolos parasitóforos, dentro dos glóbulos vermelhos, porém, ao mesmo tempo, se isolam das fontes de nutrientes essenciais e ficam vulneráveis à filtragem de hemácias no baço. Para contornar estes problemas estes organismos desenvolveram um intricado sistema de exportação de proteínas tanto para o citoplasma quanto para a membrana plasmática do eritrócito infectado. As proteínas exportadas modificam drasticamente a estrutura da célula infectada, tornando-a ideal para sustentar o desenvolvimento e crescimento do parasito, através de mecanismos de obtenção de nutrientes e da expressão de proteínas do parasito na superfície da célula hospedeira que permitem a adesão destas células ao endotélio de vasos capilares, evitando a sua detecção e destruição pelo sistema imune (CRAIG; SCHERF, 2001). Foi descoberto um motivo, na extremidade N-terminal de proteínas, formado por cinco aminoácidos e denominado PEXEL (*Plasmodium EXport ELement*) (MARTI et al., 2004) ou VTS (*Vacuolar Targeting Signal*) (HILLER et al., 2004) que é responsável pela sinalização do transporte de proteínas entre o parasito e a célula hospedeira. Porém, para a grande maioria das proteínas que são enviadas através da membrana do vacúolo parasitóforo, a exportação é iniciada com a translocação da proteína para o lúmen do RE, em um processo mediado por um peptídeo sinal localizado anteriormente ao motivo PEXEL/VTS (CRABB et al., 2010). O papel dos peptídeos sinal nestas proteínas é direcioná-las ao lúmen do vacúolo parasitóforo, que é o destino padrão da via secretória em *Plasmodium* (PRZYBORSKI; LANZER, 2005).

Portanto, considerando-se esta especialização funcional e estrutural de organelas e compartimentos celulares, fica claro que a presença de peptídeos sinal em proteínas de organismos do Filo Apicomplexa, especialmente parasitos do gênero *Plasmodium*, têm implicações biológicas muito importantes para a progressão do ciclo de vida e para as interações parasito-hospedeiro.

A hipótese deste trabalho é que, considerando-se a esperada conservação funcional entre proteínas ortólogas e a relação direta observada entre a localização subcelular de uma proteína e o seu papel biológico, a presença/ausência de peptídeos sinal seja conservada dentro de grupos de proteínas ortólogas.

Três explicações foram antecipadas para os casos onde fossem observadas divergências entre as predições de peptídeo sinal de proteínas ortólogas: **(1)** Erros de anotação da extremidade N-terminal em sequências de aminoácidos, que inviabilizam a predição correta; **(2)** Limitações ou imprecisões das metodologias empregadas para a predição de peptídeos sinal ou para a definição de relações de homologia; **(3)** Diferenças biológicas reais, resultantes da evolução divergente de uma ou mais proteínas ortólogas.

Para explorar a nossa hipótese, analisamos os resultados da predição de peptídeo sinal em grupos de proteínas ortólogas. Esta estratégia simples de combinar a ortologia com a predição de peptídeo sinal se mostrou uma ferramenta em potencial para a revisão de anotações em uma escala genômica, além de viabilizar a rápida identificação de grupos ortólogos que podem ajudar a explicar as diferenças biológicas observadas entre espécies.

Em um primeiro momento, empregamos a nossa estratégia para revisar modelos gênicos em cinco espécies do gênero *Plasmodium*, visando melhorar a qualidade da anotação das sequências proteicas, especialmente do conjunto de proteínas que apresentam peptídeos sinal, uma vez que muitas destas são reconhecidamente importantes alvos para o desenvolvimento de vacinas e drogas. Posteriormente, a etapa de identificação de grupos ortólogos contendo proteínas mal anotadas foi automatizada e implementada em um preditor. Este preditor foi então testado em vários grupos de organismos para demonstrar a aplicabilidade da estratégia de combinação nas tarefas de revisão de anotações em escala genômica e curadoria de bancos de dados genômicos.

2. JUSTIFICATIVA

A pesquisa biológica é cada vez mais dependente dos dados provenientes do sequenciamento de genes e genomas. Além disso, devido ao grande volume e à velocidade com que estes dados são gerados, os processos de identificação e anotação de genes só podem ser realizados, em larga escala, por metodologias computacionais.

Apesar da constante evolução das metodologias de predição gênica, encontrar as coordenadas corretas de modelos gênicos ainda é um grande desafio e, muitas vezes, os modelos preditos apresentam erros, principalmente na extremidade N-terminal da região codificante, justamente pela dificuldade de se identificar o correto códon de iniciação da transcrição (geralmente uma Metionina). A imprecisão destes modelos impactará nas análises subsequentes que forem dependentes da estrutura primária de proteínas como, por exemplo, a predição de peptídeos sinal e consequentemente a sua localização subcelular. Em uma escala genômica, a acumulação de erros da anotação pode ter consequências negativas significativas ao provocar a recuperação de informações equivocadas para dezenas ou mesmo centenas de genes. Além disso, informações eradas podem ser propagadas através das estratégias de genômica comparativa.

Portanto, assegurar a qualidade da anotação gênica, em todos os níveis, é essencial para garantir resultados confiáveis a partir de abordagens que utilizam direta ou indiretamente estes dados. A revisão, reanotação e curadoria das informações disponíveis em bancos de dados devem ser realizadas constantemente.

A reanotação de proteínas de espécies do gênero *Plasmodium*, em especial, tem implicações práticas para pesquisas voltadas à identificação de novos alvos moleculares de intervenção. Atualmente, muitas estratégias para a seleção de potenciais alvos para vacinas ou drogas apresentam etapas de mineração de dados que analisam, primariamente, as sequências proteicas preditas a partir dos genomas sequenciados, evidenciando a importância de se corrigir eventuais erros de anotação. Assim, o peptídeo sinal tem papel de destaque nestas estratégias, pois proteínas que são secretadas ou estão expostas na superfície celular são, geralmente, consideradas alvos promissores. Porém a sua predição é particularmente afetada pela anotação errada das proteínas, uma vez que a extremidade N-terminal, onde geralmente se localiza o peptídeo sinal, é uma região mais propensa a erros de anotação.

Ainda, a estratégia proposta para identificar prováveis erros de anotação não se restringe somente ao gênero *Plasmodium*, pois é baseada em princípios universais: a relação de ortologia entre genes e a presença/ausência de peptídeo sinal; tornando-a, teoricamente, uma ferramenta com aplicabilidade extensível a qualquer grupo de espécies. Portanto, em um contexto onde a quantidade de informação biológica acumula-se rapidamente e a minimização de erros é um desafio, torna-se importante que existam

ferramentas que se proponham a auxiliar na tarefa de identificar ou revisar sequências mal anotadas.

3. OBJETIVOS

3.1. GERAL

Aplicar a metodologia de análise de predição de peptídeos sinal em proteínas ortólogas visando aprimorar a anotação N-terminal das proteínas e desenvolver um método automático para identificação de grupos ortólogos contendo proteínas com erros de anotação na extremidade N-terminal

3.2. ESPECÍFICOS

- Combinar a informação sobre a ortologia de proteínas com a predição de peptídeo sinal para selecionar grupos de ortólogos contendo predições de peptídeo sinal divergentes;
- Identificar, após inspeção visual de alinhamentos e dos modelos gênicos, proteínas ortólogas que apresentem prováveis erros de anotação em sua sequência N-terminal;
- Propor modelos gênicos alternativos para as proteínas mal anotadas;
- Validar experimentalmente alguns dos novos modelos gênicos propostos;
- Avaliar o impacto dos novos modelos gênicos na classificação de grupos ortólogos;
- Automatizar a identificação de grupos ortólogos que contenham proteínas mal anotadas na região N-terminal;
- Testar a metodologia de identificação em diferentes conjuntos de organismos.

4. MATERIAL E MÉTODOS

4.1. Montagem dos bancos de dados

Foram criados seis bancos de dados locais com o programa MySQL 5.5.28 contendo informações sobre todos os genes preditos de vinte e duas espécies de parasitos. As denominações dos bancos e as distribuições de espécies por banco foram as seguintes: **1- Plasmodium** (cinco espécies do mesmo gênero): *Plasmodium falciparum, Plasmodium vivax, Plasmodium knowlesi, Plasmodium berguei, Plasmodium yoelii*; **2- Toxoplasma** (três cepas da uma única espécie): cepas ME49, GT1 e VEG de T*oxoplasma gondii*; **3- Cryptosporidium** (Três espécies do mesmo gênero): *Cryptosporidium parvum, Cryptosporidium hominis* e *Cryptosporidium muris*; **4- Trypanosoma** (quatro espécies do mesmo gênero): *Trypanosoma brucei, Trypanosoma cruzi, Trypanosoma vivax, Trypanosoma congolense*; **5- Leishmania** (cinco espécies do mesmo gênero): *Leishmania major, Leishmania infatum, Leishmania braziliensis, Leishmania mexicana, Leishmania tarentolae*; **6- Apicomplexa** (cinco espécies de cinco gêneros distintos): *Eimeria tenella, Neorospora caninum, Theileria annulata, Babesia bovis, Toxoplasma gondii*.

Os bancos foram criados com os seguintes dados iniciais: o identificador gênico (ID do gene), o grupo de ortólogos ao qual pertence o gene segundo dados do PlasmoDB obtidos do OrthoMCL-DB (LI; STOECKERT; ROOS, 2003), a descrição do seu produto, o organismo ao qual pertence, o número de exons, a sequência proteica predita, a sequência codificadora estendida 2000 (ou mais) bases *upstream* na extremidade N-terminal e 30 bases *downstream* na extremidade C-terminal, as coordenadas do modelo gênico original. Estas informações foram recuperadas do Banco de dados EuPathDB versão 2.14 (AURRECOECHEA et al., 2010), com exceção dos dados de *Plasmodium*, que foram recuperados do PlasmoDB versão 7.1 (AURRECOECHEA et al., 2009). Todas as atividades de alimentação, manutenção e atualização dos bancos foram realizadas através de *scripts* na linguagem perl.

Os seis bancos de dados apresentam uma estrutura organizacional idêntica, com 10 tabelas (**Figura 5**), sendo 5 com informações sobre genes individuais (as informações originais retiradas dos bancos do EuPathDB se encontram na tabela *main_table*) e 5 com informações sobre grupos de ortólogos. Os resultados de predição de peptídeo sinal para cada proteína estão na tabela *signalp_BEFORE* que, juntamente com a *main_table*, alimenta a tabela *orthogroups_default*. Estas 3 tabelas contêm as informações iniciais essenciais. As tabelas *revised_genes*, *gene_categor*y, *signalp_AFTER*, *category_by_group* e *orthogroups_AFTER* contêm informações sobre os processos de inspeção e reanotação de proteínas e as informações nas tabelas *metrics_SD_before* e *metrics_SD_after* são empregadas para a construção do classificador.

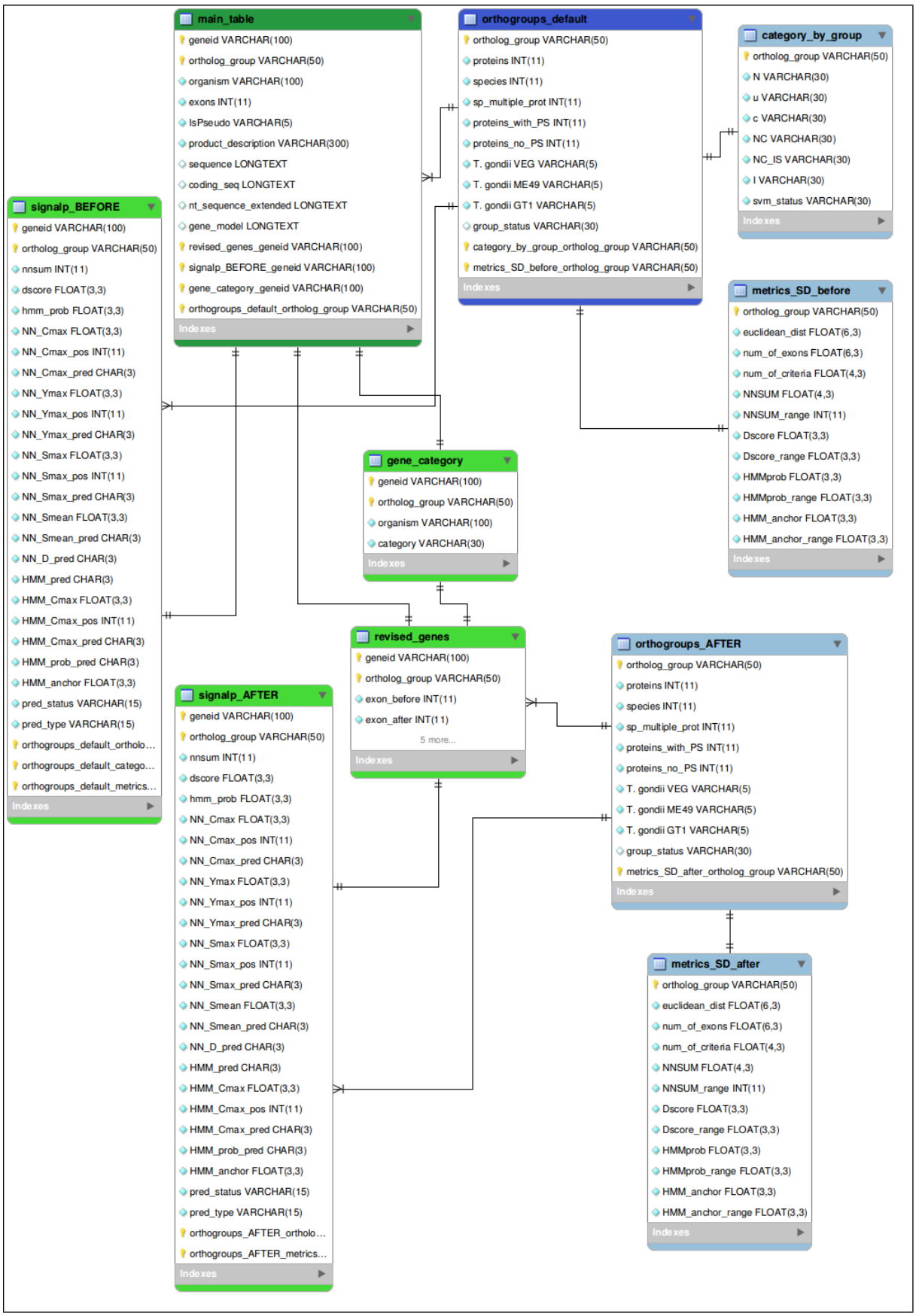


Figura 5: Diagrama da estrutura organizacional dos bancos de dados locais. As tabelas contendo informações sobre genes individuais estão destacadas em verde (em verde escuro a tabela com os dados retirados dos bancos do EuPathDB). As tabelas contendo informações sobre grupos de ortólogos estão destacadas em azul (em azul escuro a tabela resultante da classificação de grupos ortólogos através das predições de peptídeo sinal, item 4.3). Destacadas em amarelo estão as

entradas-chave de cada tabela e as linhas escuras representam os tipos de relação existentes entre as tabelas: um-pra-um ou vários-pra-um.

4.2. Predição de peptídeo sinal

O programa SignalP versão 3.0 (BENDTSEN et al., 2004) foi utilizado para as predições de peptídeo sinal, porém os parâmetros de predição foram alterados. O EuPathDB e os seus bancos afiliados permitem a predição de peptídeos sinal através de uma interface hospedada nos próprios bancos, que executa o SignalP 3.0. No entanto, os parâmetros de predição são diferentes daqueles originalmente empregados pelo SignalP 3.0 no seu próprio servidor ou em sua versão de distribuição.

O SignalP 3.0 realiza a predição de peptídeos sinal a partir de sequências de proteínas através de duas metodologias, SignalP-NN e SignalP-HMM. A primeira destas metodologias, SignalP-NN, utiliza duas Redes Neurais (*Neural Networks*) e gera como resultado cinco escores: S-escore, C-escore, Y-escore, S-médio e D-escore, que variam em uma escala gradual de 0 a 1 e apresentam limiares específicos (0,87; 0,33; 0,32; 0,48 e 0,43, respectivamente) que indicam o valor a partir do qual são considerados positivos. A segunda metodologia, SignalP-HMM, utiliza um Modelo Oculto de Markov (*Hidden Markov Model*) e gera três escores: Probabilidade de peptídeo sinal, Probabilidade de clivagem e Probabilidade de âncora, que também variam de 0 a 1 e apresentam limiares de positividade de 0,5. A Probabilidade de âncora é usada para a predição de peptídeos sinal que não sofrem clivagem após a internalização da proteína nascente que é, portanto, retida como uma proteína integral de membrana. Em sua versão independente, SignalP-NN e SignalP-HMM podem ser usados separada ou concomitantemente, sendo que, por padrão ambas as metodologias são executadas e considera-se uma predição positiva (para uma proteína cujo peptídeo sinal é clivado) quando o D-escore (NN) ou a Probabilidade de peptídeo sinal (HMM) forem iguais ou superiores aos seus respectivos limiares.

A predição realizada através dos bancos de dados do EuPathDB é baseada em três escores, o D-escore (NN) com um limiar de 0,5; a Probabilidade de peptídeo sinal (HMM) também com um limiar de 0,5 e um terceiro escore chamado NN-Sum. O NN-Sum é derivado da combinação dos demais quatro escores gerados pelo SignalP-NN (S-escore, C-escore, Y-escore e S-médio). A positividade de cada escore é avaliada (de acordo com o seu limiar padrão específico) e o NN-Sum representa a contagem de escores positivos, variando portanto, em uma escala unitária de 0 a 4. O limiar padrão a partir do qual se define a positividade do NN-Sum é 3. Uma predição positiva é alcançada se qualquer um dos três escores atingir ou ultrapassar os seus respectivos limiares. Neste trabalho, foram empregadas as condições de predição de peptídeo sinal definidas pelo EuPathDB.

4.3. Classificação dos grupos de ortólogos através das predições de peptídeo sinal

Para cada um dos seis bancos de dados, os genes codificadores de proteínas foram agrupados de acordo com sua afiliação a um grupo ortólogo. Os dados de ortologia recuperados do EuPathDB são gerados pelo programa OrthoMCL e se encontram depositados no OrthoMCL-DB. O objetivo do OrthoMCL é determinar as possíveis relações evolutivas entre proteínas de espécies distintas e entre proteínas de um mesmo genoma. Baseando-se em melhores resultados recíprocos de BLAST, o OrthoMCL cria grupos de proteínas que são ortólogas e/ou parálogas entre si, sendo que somente parálogos recentes são incluídos nos grupos. A informação sobre a qual grupo ortólogo pertencem as proteínas do banco de *Plasmodium* foram geradas com a versão 4.0, enquanto para as proteínas dos demais bancos a informação foi proveniente da versão 5.0 do OrthoMCL-DB.

Os resultados de predição de peptídeo sinal para cada proteína foram associados à informação sobre ortologia (grupo ao qual a proteína pertence), permitindo a separação dos grupos ortólogos em três classes: **I- Grupos Positivos:** nos quais todas as proteínas apresentam uma predição positiva; **II- Grupos Negativos:** nos quais todas as proteínas apresentam predições negativas; **III- Grupos Mistos:** contendo proteínas ortólogas tanto com predições positivas quanto negativas de peptídeo sinal, formando um mosaico.

4.4. Seleção e Inspeção de grupos ortólogos e Reanotação de modelos gênicos

Proteínas que não pertencem a grupos ortólogos (singletons) e os grupos contendo mais de uma proteína da mesma espécie foram excluídos das análises para a reanotação de proteínas. No banco de dados de *Plasmodium* todos os demais grupos (Positivos, Negativos e Mistos), contendo no máximo uma proteína por espécie, foram submetidos ao alinhamento múltiplo global de suas proteínas com o programa MAFFT (versão 6.717b) (KATOH et al., 2002) usando os parâmetros padrão (método rápido e progressivo FFT-NS-2, matriz de escores de aminoácidos BLOSUM62 e penalidade por abertura de gap = 1,53). Para os demais cinco bancos, somente os grupos Mistos foram submetidos ao alinhamento múltiplo global.

Para cada proteína alinhada foi criado um arquivo do tipo "*.embl*" contendo a representação da estrutura do seu modelo gênico original para visualização no programa de anotação Artemis (versões 12.0 e 13.0) (RUTHERFORD et al., 2000). Para a criação do arquivo "*.embl*", as coordenadas do modelo gênico foram projetadas na sequência codificadora extendida (2000 bases *upstream* na extremidade N-terminal e 30 bases *downstream* na extremidade C-terminal).

Alinhamentos foram visualmente inspecionados para a detecção de proteínas com possíveis erros de anotação de sequências. Além disso, também foram inspecionados os modelos gênicos originais da proteína selecionada e de seus ortólogos utilizando o programa Artemis e submetidos a uma análise comparativa considerando-se diversas

caraterísticas, tais como: posicionamento relativo das prováveis metioninas iniciais em relação a marcadores próximos conservados; número e tamanho de exons e introns; conservação das junções de exons e introns e seus sinais de splicing. Proteínas foram preliminarmente consideradas contendo prováveis erros de anotação quando, após a inspeção do alinhamento e a análise comparativa de modelos gênicos, eram constatadas inconsistências no alinhamento das sequências N-terminais (sequências de aminoácidos demasiadamente longas ou curtas) ou uma clara falta de conservação restrita a uma ou poucas proteínas. Sempre que possível, foi proposto um novo modelo gênico representando a sequência reanotada e apresentando os limites dos exons da proteína reanotada através da criação de um arquivo "*.embl"* no software Artemis. Em alguns casos foi necessário estender a região flanqueadora *upstream* e recuperar até 5000 bases (ao invés das 2000 recuperadas a princípio). Algumas proteínas selecionadas não puderam ser reanotadas, devido à falta da sequência flanqueadora completa na extremidade N-terminal (região não coberta pelo sequenciamento) ou a possíveis erros de sequenciamento ou sequenciamentos de baixa qualidade, e foram marcadas para tentativas futuras. É importante salientar que a predição de peptídeos sinal foi usada somente para a classificação de grupos ortólogos e não influenciou o processo de seleção de proteínas para reanotação, sendo este baseado inteiramente na inspeção visual dos alinhamentos e na análise comparativa dos modelos gênicos.

Concluídas as etapas de inspeção e reanotação, os grupos ortólogos foram separados em três categorias: **1- Sem erros de anotação**, onde a inspeção visual não detectou proteínas possivelmente mal anotadas; **2- Contendo prováveis erros de anotação**, onde pelo menos uma proteína provavelmente mal anotada foi identificada; **3- Inconclusivos**, onde a inspeção visual foi insuficiente para detectar se havia necessidade de reanotação e qual proteína deveria ser reanotada. A segunda categoria (Contendo prováveis erros de reanotação) foi ainda dividida em duas subcategorias: **1- Reanotados**, onde todas as proteínas selecionadas como mal anotadas foram reanotadas; **2- Parcialmente reanotados**, onde uma ou mais das proteínas selecionadas não puderam ser reanotadas.

No banco de dados de *Plasmodium* foram inspecionados 169 grupos Positivos, 291 grupos Negativos e todos os 541 grupos Mistos, porém, somente os grupos Mistos tiveram suas proteínas reanotadas. Apesar de não ter havido a proposição de novos modelos gênicos para proteínas mal anotadas provenientes de grupos Positivos e Negativos, a inspeção destes grupos seguiu os mesmos critérios aplicados à inspeção dos grupos Mistos e proteínas somente foram consideradas mal anotadas após a identificação de metioninas iniciais alternativas. Ou seja, a possibilidade de modelos gênicos alternativos foi confirmada, porém os modelos não foram criados.

Nos demais bancos, subconjuntos de grupo Mistos foram selecionados aleatoriamente para inspeção: 50 grupos de **Toxoplasma**; 80 grupos de **Cryptosporidium**; 100 grupos de **Trypanosoma**; 150 grupos de **Leishmania**; 100 grupos de **Apicomplexa**. Não houve criação de novos modelos gênicos para as proteínas mal anotadas identificadas nestes subconjuntos.

4.5. Cálculo das taxas de grupos com proteínas mal anotadas

Para o banco de dados de *Plasmodium*, calculou-se a frequência de grupos contendo pelo menos uma proteína mal anotada em cada uma das três classes (Positivos, Mistos e Negativos). As frequências para cada classe foram obtidas através do cálculo: Grupos com erros de anotação / (Grupos contendo prováveis erros de anotação + Grupos com erros de anotação). As diferenças entre as proporções obtidas foram comparadas através do Teste do Qui-quadrado, seguido do teste de Marascuilo (StatTools, http://www.stattools.net/index.php) para comparações entre pares de proporções. Intervalos de Confiança (95%) foram calculados considerando-se o número total de grupos Positivos, Mistos e Negativos antes das reanotações, 398, 541 e 3380, respectivamente.

4.6. RT-PCR de modelos gênicos alternativos propostos

4.6.1. Amostra para extração de RNA

Uma amostra de sangue de 5 mL foi coletada após a confirmação por microscopia de infecção aguda por *Plasmodium vivax* em um paciente atendido no Hospital Universitário Júlio Muller em Cuiabá, MT. O paciente foi tratado para malária de acordo com as instruções do Ministério da Saúde (MINISTÉRIO DA SAÚDE, 2010) e a amostra foi armazenada em RNALater (Invitrogen) a 4°C e transportada em nitrogênio líquido.

4.6.2. Extração de RNA total

O RNALater foi removido por centrifugação a 16.000 x g a 4°C por 10 minutos e a extração do RNA total da amostra foi realizada por TRIZOL (Invitrogen), seguindo as instruções do fabricante. Resumidamente, a amostra foi incubada com TRIZOL a temperatura ambiente por 5 minutos para a completa dissociação dos complexos nucleoprotéicos. Foram adicionados 200 µL de clorofórmio para cada 1 mL de TRIZOL usado na homogeneização das amostras, os tubos foram agitados vigorosamente e incubados a temperatura ambiente por 3 minutos antes de serem centrifugados a 12.000 x g por 10 minutos a 4ºC. Após a centrifugação a fase aquosa, contendo o RNA, foi recuperada e transferida para um novo tubo estéril. O RNA foi precipitado pela adição de 500 µL (para cada 1mL de TRIZOL) de isopropanol à amostra, que foi incubada a temperatura ambiente por 10 minutos e centrifugada a 12.000 x g a 4ºC por 10 minutos. O sobrenadante foi descartado. O sedimento contendo o RNA foi lavado com 1 mL de etanol 75%, agitado em um vortex e centrifugado a 7.500 x g por 5 minutos a 4ºC. O sobrenadante foi descartado e o sedimento foi seco ao ar. O RNA foi dissolvido em 20 µL de água livre de RNAses, aquecido a 60ºC por 10 minutos para evaporação do etanol residual e armazenado a -70ºC.

4.6.3. Dosagem do RNA total

Logo após a extração do RNA total, uma alíquota de 2 µL foi retirada e diretamente dosada em um leitor NanoDrop ND-1000. O leitor foi zerado com água proveniente da mesma alíquota usada para ressuspender e dissolver o RNA. Foram obtidas as leituras nos comprimentos de onda de 230, 260 e 280 nm.

4.6.4. Tratamento com DNAse

Após a extração, 5,0 µg de RNA total foram tratados com DNAse RQ1 (Promega) em uma reação contendo 6 µL de tampão 10 x (Tris-HCl 400 mM, $MgSO_4$ 100 mM, $CaCl_2$ 10 mM,pH 8,0), 6 µL de RQ1 DNAse e água livre de RNAses para um volume final de 60 µL. As reações foram realizadas a 37ºC por 40 minutos, ao final, 1 µL de solução de parada (EDTA 20 mM, pH 8,0) foi acrescentado e a temperatura foi elevada a 65ºC por mais 10 minutos para a inativação da enzima.

4.6.5. Síntese de DNA complementar (cDNA)

O RNA tratado foi submetido à síntese de cDNA através do sistema de transcrição reversa com a enzima ImProm-II (Promega), que permite a síntese eficiente de cDNAs com sequência completa, uma vez que as nossas amplificações são realizadas na extremidade 5' de RNAs mensageiros. O RNA total foi incubado com 6 µL de iniciadores hexameros aleatórios a 70°C por 5 minutos e resfriado rapidamente em gelo. Foram adicionados tampão para a enzima ImProm-II 5X, $MgCl_2$ [1,5mM], dNTPs [0,5mM], a enzima ImProm-II e água livre de RNAses. A mistura foi incubada sequencialmente a 25°C por 5 minutos, 42°C por 60 minutos e 70°C por 15 minutos. O cDNA sintetizado foi armazenado a -20°C. Uma segunda reação sem a presença da enzima transcriptase reversa ImProm-II foi realizada a partir da mesma amostra de RNA para servir como controle negativo.

4.6.6. Amplificações por RT-PCR

Sete proteínas reanotadas de *Plasmodium vivax*, que tiveram o seu número de exons alterado após a reanotação, foram selecionadas para validação experimental, por RT-PCR, dos novos modelos gênicos propostos. Três iniciadores senso foram desenhados para cada um dos genes, sendo que o iniciador antisenso foi o mesmo nos três pares (**Tabela 1**). Sempre que possível, os iniciadores foram desenhados em junções de exons, para evitar amplificações a partir de DNA genômico, ou em regiões intrônicas, para gerar fragmentos de tamanhos distintos quando amplificados a partir de cDNA ou DNA genômico, evitando erros de interpretação dos resultados devido à contaminação indesejada das amostras de RNA por DNA. Um dos pares foi desenhado para servir como controle positivo da reação, amplificando uma região comum aos dois modelos (original e novo), os outros dois pares foram desenhados para permitir a identificação de qual modelo gênico seria mais coerente com a sequência do RNA mensageiro, com amplificações mutuamente exclusivas. Nas reações de amplificação a concentração de Magnésio usada foi de 1,5 mM, os iniciadores foram usados a 333 nM e dNTPs a 0,2 mM. As temperaturas de anelamento e o número de ciclos para cada reação estão descritos na **Tabela 1**. As amplificações foram visualizadas em gel de agarose 1,5% corado com GelRed (Biotium) diluído 1:10.000.

Tabela 1: Sequência dos iniciadores e condições das reações de amplificação

ID do gene	Par*	Senso (5' → 3')	Antisenso (5' → 3')	Tamanho do fragmento (pb)	Temperatura de anelamento (°C)	Número de ciclos
	Ctr**	CCACTGCGAAGGAAGCACATTCG		338	60	
PVX_081500	Ori	CCTCCATCGCGTACAGAGACCTC	GCTTCCCATTTGCCAGGCGATG	798	60	40
	Alt	TGGCTAGCGAAGGAGCTGTCAAAC		453	60	
	Ctr	CCAGACAGGAAGTTGCCATTTAAAG		915	59	
PVX_083205	Ori	CTTGCCATTTGTTATCCGCTTCTC	CTGCCATTGTCCCAAAATATTCAC	1011	59	35
	Alt	ATGGTGAGATTCCTCAATTTGATC		975	59	
	Ctr	AGCTACGATGCGGAAGAAAAGCTG		421	58	
PVX_083025	Ori	GCAAAATTGTACGCAAAACTATGCGC	TCCCCAAATCGGCGGAAACTTC	483	58	30
	Alt	TGTGCGCTGTCATTCGAAGAAAG		670	62	
	Ctr	AAACCCGGGCACACACATCGATG		657	64	
PVX_002580	Ori	CTTAAAGGAGTGCACCATCGCTGC	TTCCGTCCCTCACGTACGCTG	801	64	35
	Alt	CCCGCCGTTCCCACGTTTG		961	62	
	Ctr	GTGTTTAAGTACAGCCATATTCAC		958	60	
PVX_118150	Ori	ACCCTCCACGATGTTTTTATG	GCTTTTGATGAGCCTGATTTG	1027	60	33
	Alt	AAAAATATAAGGCTGTTTAGGAAGAC		1014	60	
	Ctr	ACGAGTTTAAGAACAACGTGGAGG		297	61	
PVX_100770	Ori	TCTCACCACCCACACTACTTCC	TGTAACGAATGTACTTGCTGATCCC	423	61	30
	Alt	TTTTGCTTCCTCTTCCTGCCCC		423	61	
	Ctr	CTAGCCACGACGGAGAGAGGG		210	62	
PVX_116975	Ori	AAAGATAGCACGGGTGATTGCACA	TGTCGTCCCTAAAGCCGAAGGTG	369	62	29
	Alt	GCCTCCTGTACCAGAAAAATGAACTACC		514	62	

* Iniciadores senso: Ctr: Controle; Ori: Modelo original; Alt: Modelo alternativo

** $[Mg^{2+}]$ = 2,5mM, enquanto as demais reações foram feitas a 1,5mM

4.7. Predição de peptídeo sinal a partir de fragmentos genômicos

Os genomas das espécies *P. falciparum*, *P. berghei* e *P. knowlesi* (PlasmoDB 7.1) foram traduzidos em suas seis fases de leitura e todas as sequências com 40 ou mais aminoácidos que eram iniciadas por metioninas foram recuperadas e submetidas à predição de peptídeos sinal segundo as configurações descritas no item 4.2, e predições positivas e negativas foram contabilizadas.

4.8. Cálculo das métricas de variância

Um conjunto de onze métricas foi desenvolvido para se calcular a variância de grupos ortólogos. Dez destas métricas são derivadas da predição de peptídeos sinal com o SignalP 3.0, mesclando-se parâmetros do programa com parâmetros incorporados na predição realizada através dos bancos do EuPathDB. Oito destas métricas foram obtidas a partir dos seguintes escores: NN-Sum (parâmetro do EuPathDB), D-escore, Probabilidade de peptídeo sinal (HMM) e Probabilidade de âncora (HMM), calculando-se a amplitude e o desvio padrão para cada grupo ortólogo de cada um dos seis bancos de dados. A nona métrica é o desvio padrão do número de escores que foram positivos segundo os critérios de predição do EuPathDB. Como o EuPathDB utiliza três escores na sua predição e a positividade de um deles é suficiente para se alcançar uma predição positiva, uma dada

proteína pode apresentar de zero (predição negativa) a três escores positivos. A décima métrica é baseada no resultado completo de uma predição pelo SignalP 3.0 mais o resultado do NN-Sum. O resultado completo do SignalP 3.0 para uma dada proteína apresenta 20 entradas individuais, somando-se aos oito escores descritos anteriormente (os 5 escores das Redes Neurais e os três escores do HMM) sete predições categóricas (uma para cada escore das Redes Neurais, uma para Probabilidade sinal e uma para a Probabilidade de clivagem) e predições de posições de aminoácidos para alguns destes escores (S-escore, C-escore, Y-escore e Probabilidade de clivagem do HMM) (**Figura 6**). Os 21 elementos resultantes da predição de cada proteína (predição completa + NN-Sum) foram transformados em um vetor representativo de cada proteína, com as variáveis categóricas assumindo valores discretos (0, 1, 2), e foram calculadas as distâncias euclidianas entre vetores de proteínas ortólogas. Posteriormente, calculou-se a média e o desvio padrão das distâncias encontradas. O desvio padrão é a décima métrica. A última métrica é a única não relacionada à predição de peptídeo sinal, sendo calculada como o desvio padrão do número de exons observados para proteínas ortólogas.

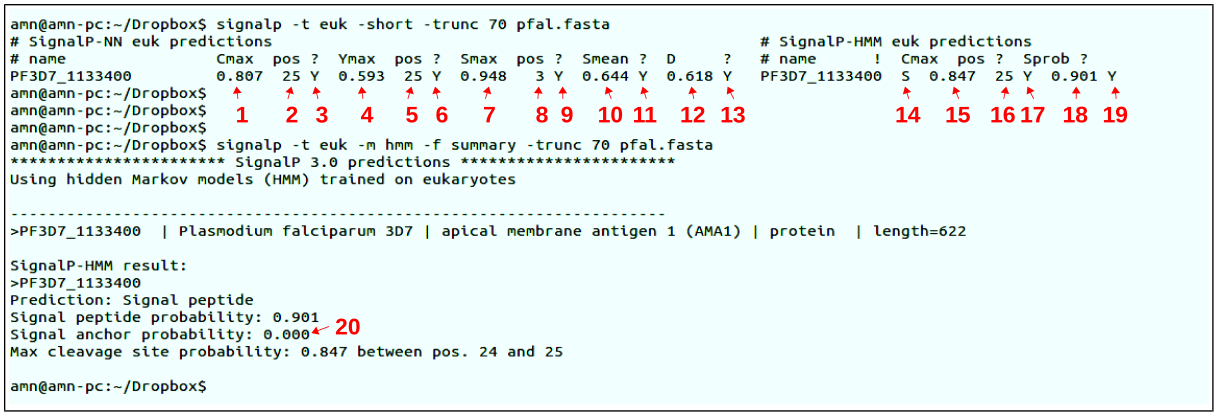


Figura 6: Resultado completo do SignalP 3.0. A predição de peptídeo sinal para um proteína gera, ao todo, 20 entradas individuais, enumeradas e apontadas pelas setas vermelhas. **1:** C-escore (valor máximo); **2:** posição do valor máximo de C-escore; **3:** variável categórica para o resultado do C-escore; **4:** Y-escore (valor máximo); **5:** posição do valor máximo de Y-escore; **6:** variável categórica para o resultado do Y-escore; **7:** S-escore (valor máximo); **8:** posição do valor máximo de S-escore; **9:** variável categórica para o resultado do S-escore; **10:** S-médio; **11:** variável categórica para o resultado do S-médio; **12:** D-escore; **13:** variável categórica para o resultado do D-escore; **14:** variável categórica para o resultado do Modelo Oculto de Markov; **15:** Probabilidade de clivagem do HMM; **16:** posição do valor máximo da Probabilidade de clivagem do HMM; **17:** variável categórica para o resultado da Probabilidade de clivagem do HMM; **18:** Probabilidade de sinal do HMM; **19:** variável categórica para o resultado da Probabilidade de sinal do HMM; **20:** Probabilidade de âncora do HMM. Para as variáveis categóricas representadas por (?), Y = predição positiva (acima do valor de corte). Para a variável categórica (!), S = predição positiva para peptídeo sinal.

4.9. Sequências reanotadas e novas predições de peptídeo sinal

As novas sequências proteicas e suas novas contagens de exons foram recuperadas a partir dos arquivos "*.embl*" dos novos modelos gênicos das proteínas reanotadas, e estes dados foram inseridos no banco de dados **Plasmodium**. As sequências reanotadas foram submetidas à predição de peptídeo sinal, seguindo as mesmas configurações de predição

aplicadas anteriormente (configurações padrão do EuPathDB, item 4.2), e as predições revisadas foram inseridas no banco de dados **Plasmodium**.

4.10. Reclassificação de grupos ortólogos após reanotações

Os grupos do banco de dados **Plasmodium** que apresentaram proteínas reanotadas foram reclassificados de acordo com as novas predições de peptídeo sinal de suas proteínas. A combinação de ortologia com predição de peptídeo sinal seguiu os mesmo parâmetros aplicados na etapa de classificação inicial dos grupos (item 4.3). Após a reclassificação, a proporção que cada classe representava em relação ao total de grupos ortólogos foi recalculada.

4.11. Cálculo das métricas de grupos após reanotações

As novas predições também foram empregadas no cálculo das onze métricas para os grupos Reanotados, seguindo os mesmos procedimentos descritos no item 4.8.

4.12. Atualizações dos bancos de dados

Todas as etapas de atualização dos bancos de dados foram executadas através de *scripts* na linguagem perl.

4.13. Obtenção de evidências experimentais através do ApiLoc

Evidências experimentais encontradas na literatura descrevendo a localização subcelular para proteínas de *Plasmodium* reanotadas ou para suas proteínas ortólogas (em espécies do gênero *Plasmodium* somente) foram recuperadas do ApiLoc (http://apiloc.biochem.unimelb.edu.au/apiloc/apiloc). O ApiLoc é um repositório curado manualmente e que utiliza um vocabulário estruturado para descrever a localização subcelular de proteínas de organismos do Filo Apicomplexa que tiveram a sua localização demonstrada experimentalmente. As informações recuperadas para cada proteína foram analisadas para se identificar qual seria a predição de peptídeo sinal (positiva ou negativa) esperada para cada uma delas.

4.14. Descrição de proteínas pelo BDA (*Blast Description Annotator*)

Além de evidências experimentais, a anotação funcional das proteínas de *Plasmodium* reanotadas foi revisada com a aplicação do algoritmo *Blast Description Annotator*, disponível no pacote do programa Blast2go (CONESA et al., 2005) e que recupera a partir de resultados do BLAST a melhor descrição possível para o produto da expressão de um gene. O BDA funciona em dois estágios, primeiramente excluindo termos pré-definidos e, posteriormente, transferindo as descrições mais frequentes encontradas

entre os resultados. O programa BLASTP foi executado a partir do pacote Blast2go, aplicando-se os parâmetros padrão (banco de dados nr, valor de corte do *e-value* = $1{,}0E^{-3}$) e o BDA foi executado concomitantemente com o BLAST.

4.15. Otimização dos valores de predição de peptídeo sinal

Os três parâmetros empregados para a predição de peptídeos sinal pelo PlasmoDB: NN-Sum; D-escore; e Probabilidade de peptídeo sinal (HMM), foram testados com diferentes combinações de limiares de positividade e o número de grupos Mistos da cada combinação foi registrado. Primeiramente, para realizar uma varredura menos precisa, porém mais rápida, D-escore e Probabilidade de peptídeo sinal (HMM) variaram de 0,05 a 1,0 (valor máximo) em incrementos de 0,05 enquanto o NN-Sum foi testado em todas as suas possibilidades (0 a 4). Posteriormente, após a identificação de uma região onde as contagens de números Mistos foram mais baixas, o NN-Sum foi fixado em seu limiar ótimo e D-escore e Probabilidade de peptídeo sinal (HMM) foram testados em uma faixa mais restrita, porém com uma variação de 0,01, realizando assim uma varredura mais minuciosa. É importante salientar que para a seleção de limiares ótimos os grupos ortólogos contendo mais de uma proteína por espécie também foram considerados na contagem de grupos Mistos.

Uma vez selecionados os limiares ótimos para NN-Sum, D-escore e Probabilidade de peptídeo sinal (HMM), estes valores foram aplicados aos grupos ortólogos para sua reclassificação. A reclassificação por otimização foi realizada após a incorporação das reanotações e, consequentemente, de suas reclassificações resultantes.

4.16. Treinamento da Máquina de Vetores de Suporte (SVM)

O programa LIBSVM (CHANG; LIN, 2011) juntamente com algumas de suas ferramentas associadas (LIBSVM *Tools*) foi utilizado para a construção de um classificador baseado em uma Máquina de Vetores de Suporte, SVM (*Support Vector Machine*), capaz de identificar, dentro de um conjunto de grupos Mistos, quais apresentariam proteínas mal anotadas. O classificador foi construído a partir dos dados da inspeção manual do banco de dados **Plasmodium**, onde 111 grupos Mistos não apresentaram proteínas mal anotadas (exemplos negativos) e 352 grupos apresentaram pelo menos uma proteína cuja sequência N-terminal estaria possivelmente errada (exemplos positivos), formando os 463 grupos Mistos que formam o conjunto de treinamento. As onze métricas calculadas, para cada um destes 463 grupos ortólogos, foram normalizadas e fornecidas como atributos ao classificador. Anteriormente ao treinamento do classificador, foi realizada a seleção de atributos (CHEN; LIN, 2005) utlizando-se a ferramenta '*Feature selection tool*' (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#feature_selection_tool) que calcula o F-escore

de cada atributo e determina, através de validação cruzada em 5 vezes, qual a combinação de atributos (métricas) retorna a maior acurácia.

Para o treinamento do classificador, os atributos selecionados foram normalizados e fornecidos ao LIBSVM e os parâmetros C e γ da função kernel de base radial, RBF (R*adial Basis Function*), foram otimizados através de uma varredura de diferentes combinações par a par (C x γ), com os valores crescendo exponencialmente. O par que apresentou a maior acurácia, segundo os resultados da validação cruzada em 5 vezes, foi selecionado para o treinamento (HSU; CHANG; LIN, 2010).

A curva ROC (*Receiver Operating Characteristic*) e a sua respectiva Área Sob a Curva, AUC (*Area Under the Curve*), referentes à validação cruzada em 5 vezes do treinamento do classificador, foram calculadas com a ferramenta '*ROC curve for binary SVM*' (http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#roc_curve_for_binary_svm) disponível no LIBSVM *tools*. Foram calculadas também as seguntes medidas de performance do classificador: Exatidão global, Taxa de retorno (*recall*) e precisão referentes aos grupos **Com erros de anotação** e o coeficiente de correlação de Matthews. O treinamento apresenta como resultado um modelo contendo os parâmetros da Máquina de Vetores de Suporte.

4.17. Classificação de grupos Mistos com a Máquina de Vetores de Suporte (SVM)

O modelo gerado após o treinamento do classificador com os dados de **Plasmodium** foi testado em problemas de classificação de grupos dos cinco bancos de dados restantes (**Cryptosporidium**, **Toxoplasma**, **Trypanososoma**, **Leishmania** e **Apicomplexa**), novamente usando o LIBSVM. Para cada banco, os conjuntos de teste foram formados por todos os seus grupos ortólogos Mistos, para os quais foram recuperadas as mesmas métricas empregadas na construção e treinamento do modelo. Os conjuntos de teste foram normalizados na mesma escala usada para o conjunto de treino e, então, submetidos à predição automática pelo classificador, e cada grupo foi colocado em uma das duas classes: **(1)** Sem erros de anotação ou **(2)** Contendo pelo menos uma proteína mal anotada. Como descrito anteriormente, subconjuntos aleatórios de grupos Mistos dos cinco bancos de dados usados para testar o classificador foram selecionados e inspecionados visualmente em busca de proteínas mal anotadas. Os resultados da inspeção manual foram comparados às predições automáticas obtidas pelo classificador, para se calcular qual a acurácia do classificador nos diferentes cenários representados por cada conjunto de testes.

4.18. Análises estatísticas

Diferenças entre duas ou mais proporções foram calculadas com o teste do Qui-quadrado seguido pelo teste *post hoc* de Marascuilo para comparações par a par. O teste não paramétrico de Mann-Whitney foi usado na comparação de medianas de amostras

independentes e o teste pareado de Wilcoxon foi usado nas comparações de amostras dependentes. Todos os testes consideraram o nível de significância estatística de 5% ($p < 0,05$).

5. RESULTADOS

PARTE I – Predição de peptídeos sinal e erros de anotação em proteínas de espécies do gênero *Plasmodium*

Todo o processo, desde o agrupamento de proteínas em grupos ortólogos, passando pela classificação, seleção e inspeção manual destes grupos, seguida pela reanotação de proteínas do gênero *Plasmodium,* encontra-se ilustrado esquematicamente na **Figura 7**.

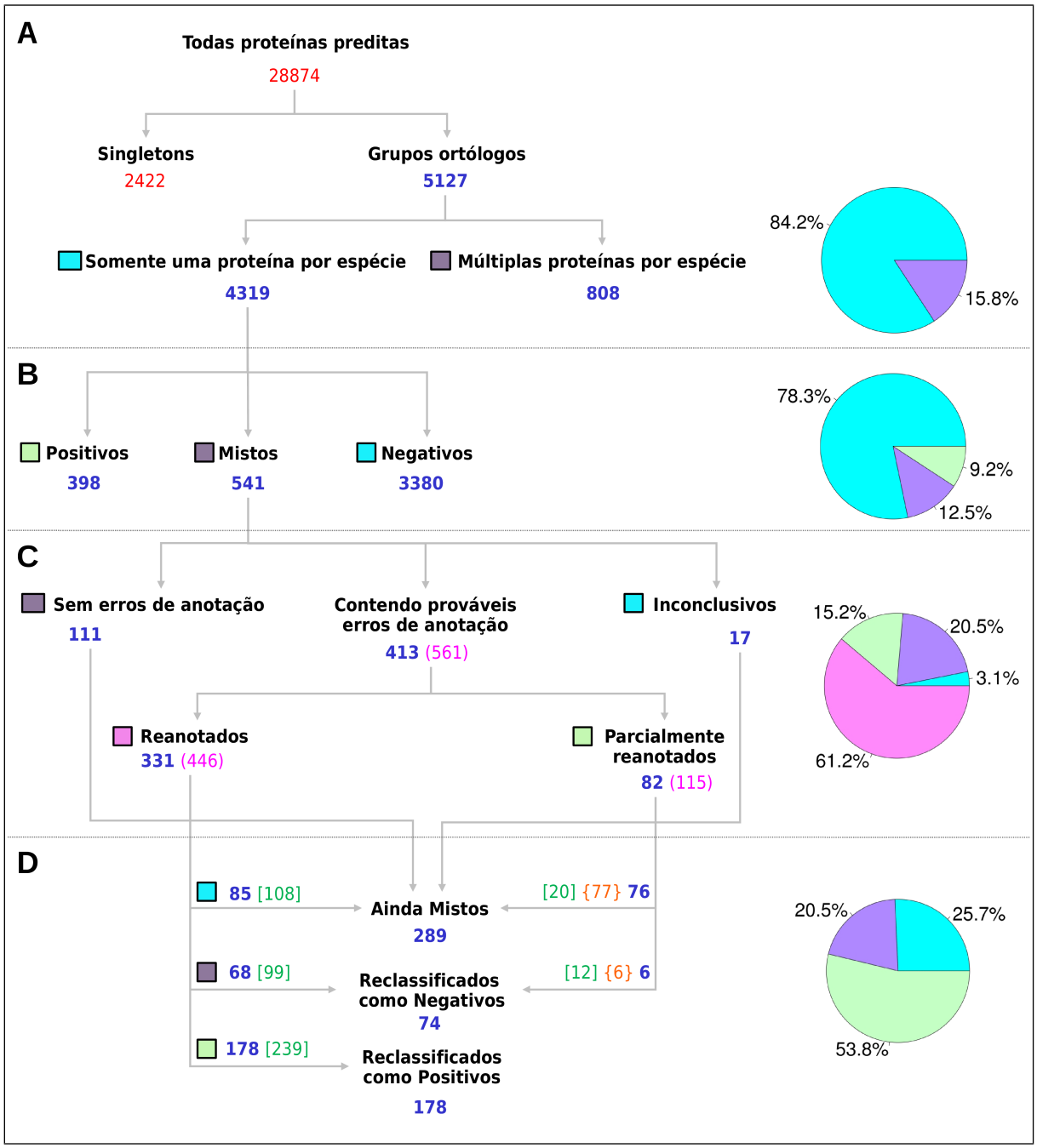


Figura 7: Desenho esquemático dos processos de seleção, classificação, reanotação e reclassificação dos grupos ortólogos. (A) Seleção dos grupos ortólogos. Agrupamento das proteínas preditas de cinco espécies de plasmódios (*P. vivax*, *P. knowlesi*, *P. falciparum*, *P. berghei*, *P. yoelii*) de acordo com os grupos ortólogos definidos pelo OrthoMCL-DB (versão 4). Numeros em azul

representam os grupos de ortólogos e em vermelho o total de proteínas. **(B)** Classificação dos grupos ortólogos de acordo com a predição do peptideo sinal: grupos **Positivos**, todas as proteínas do grupo possuem predição positiva de peptideo sinal; grupos **Negativos**, todas as proteínas do grupo possuem predição negativa de peptideo sinal; e grupos **Mistos**, algumas proteínas ortólogas do grupo possuem predição positiva e outras negativa de peptideo sinal. Os números em azul representam o número de grupos por categoria. **(C)** Classificação dos grupos **Mistos** , após inspeção manual dos alinhamento e dos modelos gênicos, em três categorias: (*i*) *Sem erros de anotação*; (*ii*) *Contendo prováveis erros de anotação*; e (*iii*) *Inconclusivos (informações de sequencia insuficientes para classificação em outra categoria)*. A categoria Contendo prováveis erros de anotação foi dividida em duas subcategorias: (*i*) *Reannotados (todas as proteínas identificadas como prováveis mal anotadas foram reanotadas)*; e (*ii*) *Parcialmente reanotados (uma ou mais proteínas identificadas como potencialmente mal anotadas não puderam ser reanotadas por falta de informações disponíveis)*. Números em azul representam os grupos de ortólogps em cada categoria e em rosa o número de proteínas provavelmente mal anotadas em cada categoria ou subcategoria. **(D)** Reclassificação dos grupos de acordo com a predição de peptídeo sinal de suas proteínas após a reanotação das mesmas. Números em azul representam os grupos de ortólogos, em verde o número de proteinas reanotadas e em laranja as proteínas provavelmente mal anotadas que não puderam ser reanotadas por falta de informações disponíveis. Do lado direito do esquema os gráficos representam as porcentagens dos grupos de ortólogos em cada painel de acordo com as cores indicativas.

5.1. Classificação de grupos ortólogos

As cinco espécies de *Plasmodium* estudadas reuniam, ao todo, 28.874 proteínas preditas, sendo que, 2.422 (Singletons) não apresentavam relação de ortologia com nenhuma outra proteína neste conjunto (**Figura 7A**). As demais proteínas, quando agrupadas de acordo com sua ortologia, foram distribuídas em 5.127 grupos ortólogos definidos pelo banco de dados OrthoMCL-DB 4.0 (**Figura 7A**). Em 808 (15,8%) destes grupos, observou-se o agrupamento de múltiplas proteínas provenientes de uma única espécie e estes grupos foram excluídos das análises de reanotação de proteínas (**Figura 7A**). Os 4.319 grupos restantes apresentam, no máximo, uma proteína por espécie e, portanto, variam entre duas a cinco proteínas. A grande maioria destes grupos (77%) apresenta proteínas de todas as cinco espécies de *Plasmodium* estudadas, refletindo a proximidade evolutiva destas espécies (**Figura 8**).

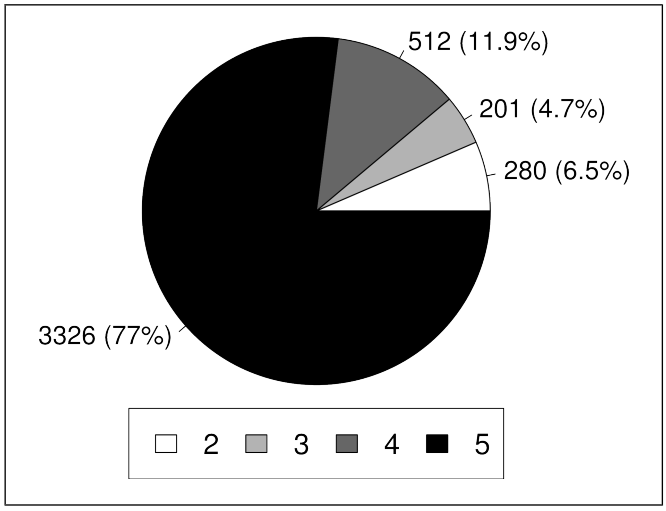


Figura 8: Total de proteínas compondo os grupos ortólogos. Os 4319 grupos ortólogos com no máximo uma proteína por espécie podem apresentar de duas a cinco proteínas. Foi feito o levantamento do número de proteínas em cada grupo.

Os 4.319 grupos foram divididos em três classes: 9,2% **Positivos**, 12,5% **Mistos** e 78,3% **Negativos**, de acordo com a concordância entre as predições de peptídeos sinal de suas proteínas ortólogas (**Figura 7B**, **Figura 9**). Analisando a **Figura 9**, é possível observar que *P. yoelii* se destaca das demais espécies por apresentar uma maior proporção de proteínas com predição negativa pertencentes a grupos Mistos. Além disso, esta mesma espécie não está representada em uma grande parte dos grupos ortólogos, como evidenciado pelos espaços em branco na coluna que representa a espécie.

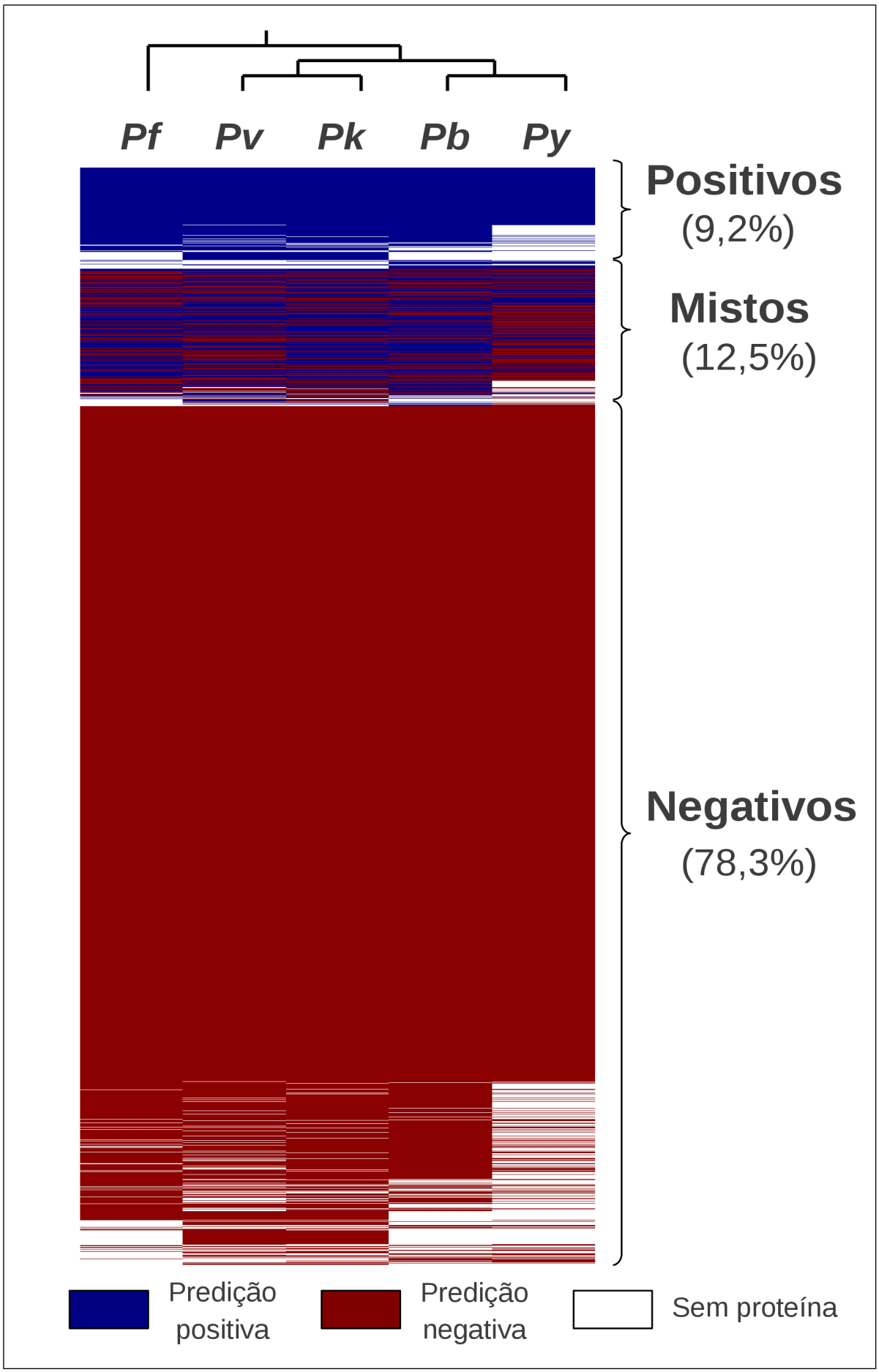


Figura 9: Distribuição dos grupos ortólogos classificados de acordo com a predição de peptídeo sinal de suas proteínas. Composição dos 4319 grupos por proteínas das diferentes

espécies de *Plasmodium*: Pv (*P. vivax*); Pk (*P. knowlesi*); Pf (*P. falciparum*); Pb (*P. berghei*); Py (*P. yoelii*). Linhas horizontais representam cada grupo ortólogo e as cores representam as predições de peptídeo sinal positivas (azul), negativas (vermelho) ou a ausência (branco) da proteína da espécie no grupo. Foram utilizadas as configurações padrão do SignalP implementadas no PlasmoDB e nos demais bancos de dados do EuPathDB para definir predições positivas: NN-Sum ≥ 3 ou D-escore ≥ 0,5 ou Probabilidade de peptídeo sinal por HMM ≥ 0,5.

5.2. A maioria dos grupos Mistos apresenta proteínas mal anotadas

A decisão sobre a necessidade de reanotação de uma proteína era resultado tanto da inspeção minuciosa do alinhamento múltiplo das sequências proteicas do grupo Misto ao qual ela pertencia quanto da análise comparativa realizada entre as regiões genômicas codificadoras (até 2000 bases, ou mais, na sua extremidade 5') dos modelos gênicos destas mesmas proteínas. Todos os 541 grupos Mistos foram inspecionados e divididos em três categorias (como descrito na Metodologia): **(1)** 17 grupos foram considerados Inconclusivos; **(2)** 111 grupos não apresentaram proteínas mal anotadas; e **(3)** 413 grupos apresentaram pelo menos uma proteína contendo provável erro de anotação em sua sequência N-terminal (**Figura 7C**).

Ao todo foram identificadas 561 proteínas mal anotadas (número rosa entre parênteses na **Figura 7C**), porém não foi possível realizar a reanotação de todas. Portanto, os 413 grupos apresentando proteínas mal anotadas foram separados em duas subcategorias, que descrevem o seu estado de reanotação: **(1)** nos 331 grupos Reanotados (**Figura 7C**), todas as 446 (**Figura 7C**) proteínas identificadas como mal anotadas foram revisadas e novos modelos gênicos foram propostos; **(2)** nos 82 grupos Parcialmente reanotados (**Figura 7C**), 83 proteínas (números em laranja dentro de chaves na **Figura 7D**), que haviam sido selecionadas para reanotação, não puderam ser corrigidas. A não correção destas proteínas deve-se, principalmente, a truncamentos da região 5' (não foi possível recuperar as 2000 bases) que impediram a eventual identificação de uma metionina inicial alternativa e, em um caso, uma mudança da janela de leitura no meio de um exon, que foi interpretada como um possível erro de sequenciamento, impediu a revisão da anotação. Ainda nos 82 grupos Parcialmente reanotados, 32 proteínas foram reanotadas e, juntamente com as 446 proteínas citadas anteriormente, totalizaram as 478 proteínas (números em verde dentro de colchetes na **Figura 7D**) para as quais novos modelos gênicos, apresentando sequências alternativas, foram propostos.

Para exemplificar como o processo de inspeção de alinhamentos resulta na identificação de proteínas mal anotadas, e como a reanotação de proteínas pode alterar tanto a predição de peptídeo sinal quanto a classificação de um grupo de ortólogos, cinco alinhamentos (somente a região N-terminal das proteínas) são demonstrados na **Figura 10**. Na **Figura 10A**, três exemplos de grupos onde a inspeção não encontrou proteínas mal anotadas (Sem erros de anotação). É interessante ressaltar que mesmo proteínas ortólogoas com extremidades N-terminais muito conservadas podem apresentar resultados

de predição de peptídeo sinal diferentes. Na **Figura10B**, dois grupos que apresentaram proteínas mal anotadas que foram reanotadas (Reanotados) e que após a reanotação foram reclassificados em consequência da alteração das predições de peptídeo sinal. No grupo OG4_54958, as proteínas PVX_116975 e PY01697 foram reanotadas, sendo que somente a predição de PVX_116975 foi alterada. Este grupo foi reclassificado de Misto para Negativo. No grupo OG4_54960, as proteínas PVX_083025 e PY07307 foram reanotadas e as predições de peptídeo sinal para ambas foram alteradas de negativas para positivas. O grupo foi reclassificado de Misto para Positivo.

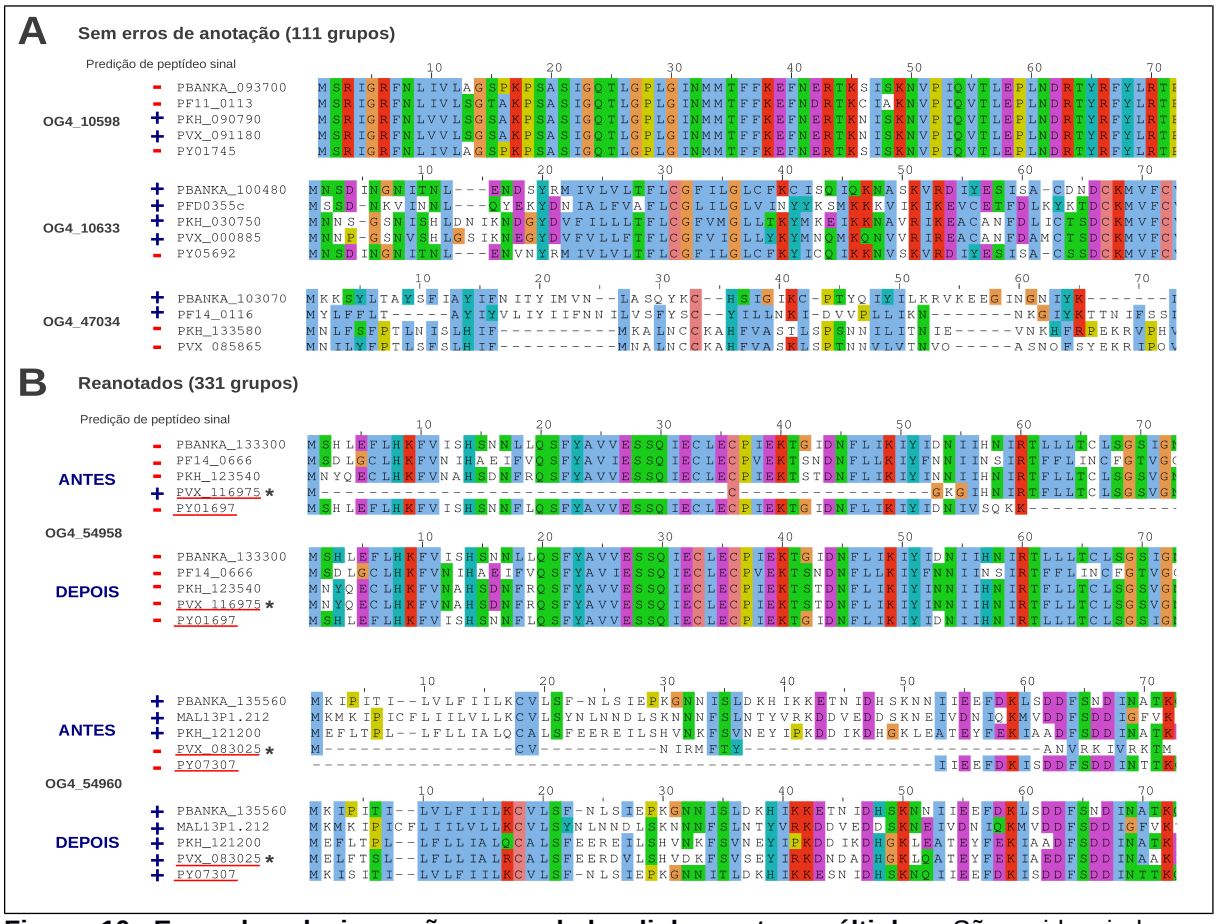


Figura 10: Exemplos da inspeção manual de alinhamentos múltiplos. São evidenciadas as porções N-terminais dos alinhamentos múltiplos de proteínas ortólogas. Em **(A)**, três grupos Mistos onde não foram encontrados proteínas mal anotadas (Sem erros de anotação). Em **(B)**, dois grupos contendo proteínas mal anotadas que foram reanotadas (Reanotados), em uma comparação Antes e Depois das reanotações. Entre parênteses o total de grupos em cada categoria. Proteínas destacadas com uma linha vermelha em B foram reanotadas. Proteínas destacadas com um * foram validadas experimentalmente (ver seção 5.9).

5.3. Grupos ortólogos Mistos concentram erros de anotação na extremidade N-terminal de proteínas

Além da inspeção realizada em todos os grupos Mistos, subconjuntos de grupos Positivos (N = 169) e Negativos (N = 291) também foram inspecionados para a identificação de proteínas com indícios de erros de anotação N-terminal. A inspeção das três classes

revelou que a proporção de grupos apresentando proteínas com prováveis erros de anotação N-terminal é significativamente maior entre os grupos Mistos, que apresentaram uma taxa de erros de 78,8 ± 0,6 %, enquanto Positivos e Negativos apresentaram taxas de 14,2 ± 4,0 % e 33,7 ± 5,2 %, respectivamente (**Figura 11**). Apesar de não terem sido propostos novos modelos gênicos para as proteínas mal anotadas de grupos Positivos e Negativos, a inspeção destes grupos seguiu rigorosamente os mesmos critérios adotados para os grupos Mistos, e proteínas somente foram consideradas mal anotadas após a constatação da possibilidade de modelos gênicos alternativos.

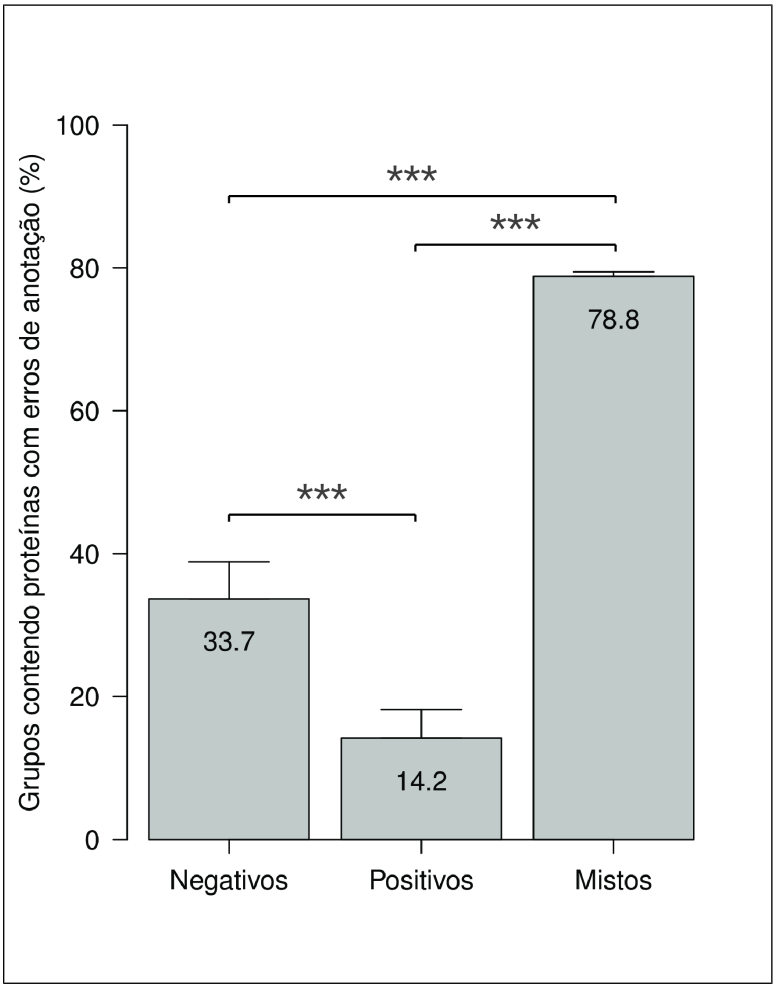


Figura 11: Porcentagem de grupos contendo pelo menos uma proteína mal anotada. Nos 291 grupos Negativos analisados foram identificados 98 grupos com pelo menos uma proteína mal anotada; nos grupos Positivos foram identificados 24 grupos em 169 analisados e nos grupos Mistos, 330 grupos em 442 analisados. Diferenças entre as porcentagens foram analisadas pelo teste do Qui-quadrado e pelo teste de Marascuilo post-hoc para identificar diferenças significativas entre os pares de grupos (*** p<0.0001).

5.4. Reanotações geralmente promovem mudanças de predição de peptídeo sinal

Após a reanotação, as novas sequências das 478 proteínas revisadas foram submetidas à predição de peptídeo sinal, seguindo as mesmas condições de predição aplicadas anteriormente. Mudanças de predição de peptídeo sinal ocorreram em 364

(76,2%) proteínas reanotadas, com 279 (58,4%) proteínas passando a apresentar predições positivas e 85 (17,8%) apresentando predições negativas (**Figura 12A**). Um total de 114 proteínas (23,8%) manteve o resultado da predição de peptídeo sinal original, mesmo tendo sido reanotadas (**Figura 12A**). Estas proteínas, cujas predições se mantiveram, foram encontradas com maior frequência (teste de proporções Qui-quadrado, p = 0.0002286) entre grupos Mistos que apresentavam múltiplas proteínas reanotadas **(Figura 12B**), indicando que a seleção destas proteínas para serem reanotadas pode ter sido um resultado secundário do próprio processo de reanotação, que durante a inspeção considerava todas as proteínas do grupo. Ainda, em 14 destas proteínas, as reanotações foram direcionadas para regiões posteriores ao sítio de clivagem do peptídeo sinal predito e, portanto, não afetariam a predição. Estas reanotações que não afetam propriamente o peptídeo sinal são, também, consequências da dinâmica do processo de reanotação.

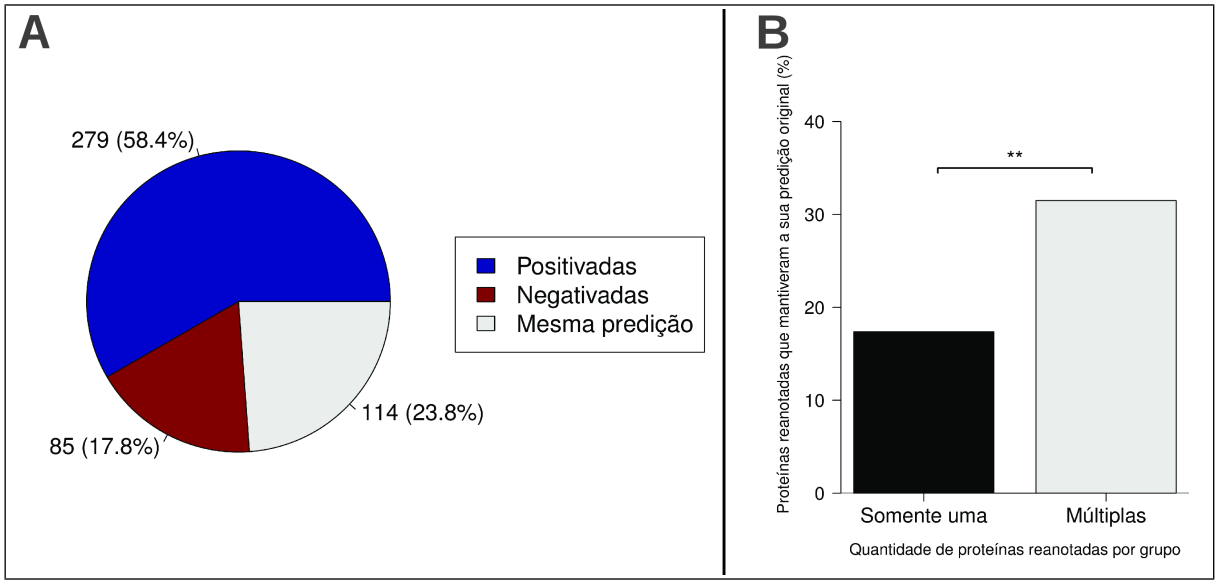
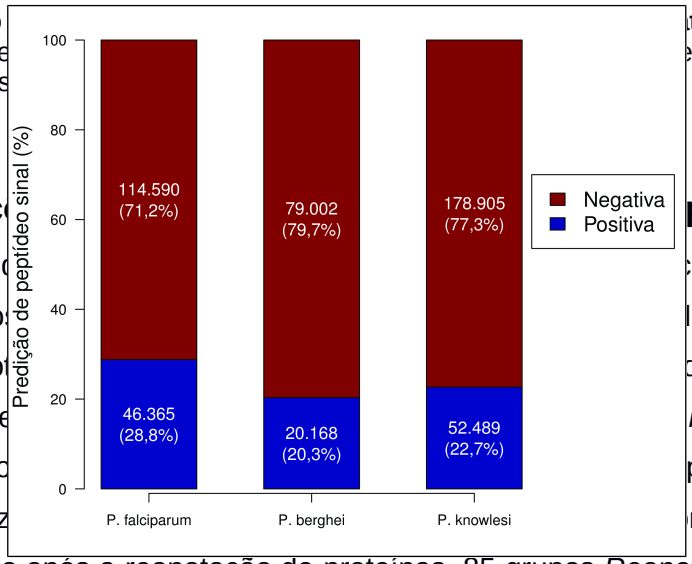


Figura 12: Impacto da reanotação das proteínas na predição do peptídeo sinal. (A) Número e porcentagem de proteínas que tiveram a predição de peptídeo sinal alterada para positiva (azul), para negativa (vermelho) ou se mantiveram inalteradas (cinza claro). (B) Porcentagem de proteínas que mantiveram sua predição de peptídeo sinal inalterada após a reanotação, distribuídas pelo número de proteínas reanotadas por grupo: somente uma proteína (preto) reanotada ou múltiplas proteínas reanotadas (cinza). O teste do Qui-quadrado foi utilizado para calcular significância estatística entre as diferentes proporções (** p<0.001).

5.5. A probabilidade da predição de peptídeo sinal em sequências aleatórias ser negativa é mais alta

Todas as possíveis sequências com 40 ou mais aminoácidos e iniciadas por metioninas que podem ser codificadas nos genomas de *P. falciparum*, *P. berghei* e *P. knowlesi* foram recuperadas e tiveram a sua predição de peptídeo sinal determinada (**Figura 13**). A proporção de predições negativas foi sempre superior à de predições positivas, demonstrando que existe uma maior probabilidade que uma proteína que se inicia em uma metionina aleatória apresente predição negativa.

Figura 13: Predição [illegible] tórias. Proporção de predições positivas e[illegible]eis com 40 ou mais aminoácidos, iniciadas[illegible] *falciparum*, *P. berghei* e *P. knowlesi*.



5.6. Reanotaç[illegible]pos Mistos

A mudança [illegible]classificação original dos grupos ortólogos[illegible]le ser Mistos após a revisão de suas pro[illegible]dança mais comum) e 68 tornaram-se Ne[illegible] *reanotados* também foram reclassificado[illegible]proteínas revisadas (**Figura 7D**), totaliz[illegible]r consequência de reanotações. Mesmo após a reanotação de proteínas, 85 grupos *Reanotados* mantiveram a sua classificação como Mistos (**Figura 7D**).

Com estes resultados, a proporção de grupos Mistos caiu de 12,6% para somente 6,7% (**Figura 14**), representando uma redução significativa e corroborando a hipótese inicial de que a presença de sequências mal anotadas é uma das possíveis razões da divergência entre predições de peptídeos sinal entre proteínas ortólogas. Os 289 grupos que mantiveram a classificação de Mistos são formados pelos 111 que não apresentaram erros de anotação, 85 *Reanotados*, 76 *Parcialmente reanotados* (contendo proteínas que não podem ser reanotadas nas condições atuais) e os 17 grupos *Inconclusivos* (**Figura 7D**).

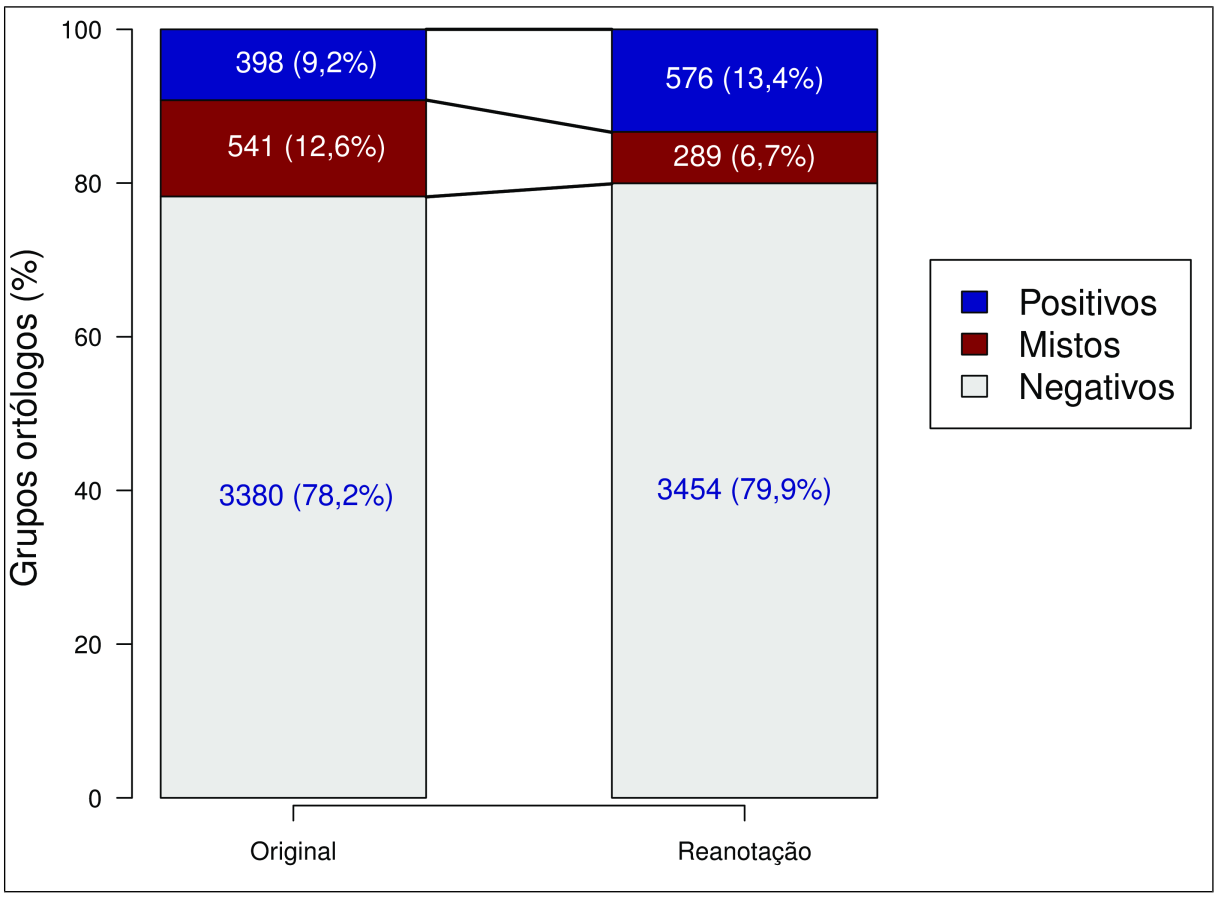


Figura 14: Impacto da reanotação das proteínas na classificação dos grupos quanto à predição dos peptídeos sinal. Reclassificação dos grupos Mistos após a reanotação das proteínas. Configurações padrão do SignalP utilizadas pelo PlasmoDB foram usadas para definir predições positivas de peptídeo sinal: NN-Sum ≥ 3 ou D-escore ≥ 0,5 ou Probabilidade de peptídeo sinal por HMM ≥ 0,5.

5.7. Reanotações refletem os estados de montagem dos genomas de *Plasmodium*

As espécies mais afetadas pela reanotação foram *Plasmodium yoelii* seguida por *P. vivax*, apresentando 208 e 158 proteínas corrigidas, respectivamente (**Tabela 2**). Esta reanotação extensiva reflete o nível de montagem dos genomas destas duas espécies, que entre as espécies estudadas são as que apresentam os genomas mais fragmentados. Segundo dados da versão 7.1 do PlasmoDB, o genoma de *P. yoelii* está fragmentado em 5617 contigs, não organizados em cromossomos e o genoma de *P. vivax*, apesar de apresentar os 14 cromossomos típicos das espécies de *Plasmodium*, ainda apresenta boa parte de seu genoma distribuído em 2764 contigs. Em contrapartida, os genomas de espécies mais amplamente estudadas como *P. falciparum* e *P. berghei*, encontram-se melhor consolidados, com todas as sequências atribuídas a um dos 14 cromossomos (**Tabela 2**).

Os números totais de proteínas com predições positivas por espécie foram comparados antes e depois das reanotações e, com exceção de *P. berghei*, as reanotações resultaram no aumento do número total de proteínas positivamente preditas, especialmente

para *P. vivax* e *P. yoelii*, que apresentaram acréscimos de 89 e 84 proteínas positivas, respectivamente (**Tabela 2**). Proporcionalmente, o impacto das reanotações foi mais alto para *P. vivax*, representando um incremento acima de 10% no número final de proteínas com predições de peptídeo sinal positivas, após a incorporação das novas sequências de proteínas (**Tabela 2**).

Tabela 2: Classificação por espécie das proteínas reanotadas, reclassificação da predição de peptídeo sinal após a reanotação e corelação com o status de montagem do genoma

Espécie	Predição de peptídeo sinal após reanotações				Predições positivas de peptídeo sinal				Montagem do genoma	
	Negativada	Positivada	Mesma predição	**TOTAL**	Antes	Depois	Δ (Depois - Antes)	Aumento (%)	Cromossomos + contigs	Mediana (contigs)
P. vivax	16	105	37	158	802	911	89	10,8%	14[a] + 2764[b]	987,5
P. knowlesi	15	29	21	65	840	854	14	1,7%	14[a] + 67[b]	2112
P. falciparum	5	15	8	28	1057	1067	10	0,9%	14[a]	-
P. berghei	8	5	6	19	807	804	-3	-0,4%	14[a]	-
P. yoelii	41	125	42	208	983	1067	84	8,5%	5617[b]	2142
TOTAL	85	279	114	478	4509	4703	194	4,3%	-	

[a] Número de cromossomos resultantes da montagem dos genomas de cada espécies segundo dados do PlasmoDB versão 7.1

[b] Numero de contigs resultantes da montagem dos genomas de cada espécies segundo dados do PlasmoDB versão 7.1

5.8. Alterações curtas foram mais comuns durante as reanotações

Durante o processo de reanotação, a remoção de sequências de aminoácidos ocorreu em 346 proteínas e foi mais comum que a adição de sequências de aminoácidos (131 proteínas) (**Figura 15**). Uma única proteína manteve o seu comprimento inalterado mesmo após sua sequência ter sido revisada. As distribuições dos comprimentos dos fragmentos de aminoácidos, tanto os adicionados quanto os removidos, revelam que predominaram modificações curtas, sendo adições de até 10 aminoácidos (detalhe **Figura 15A**) e remoções entre 20-25 aminoácidos (detalhe **Figura 15B**) as alterações mais comuns.

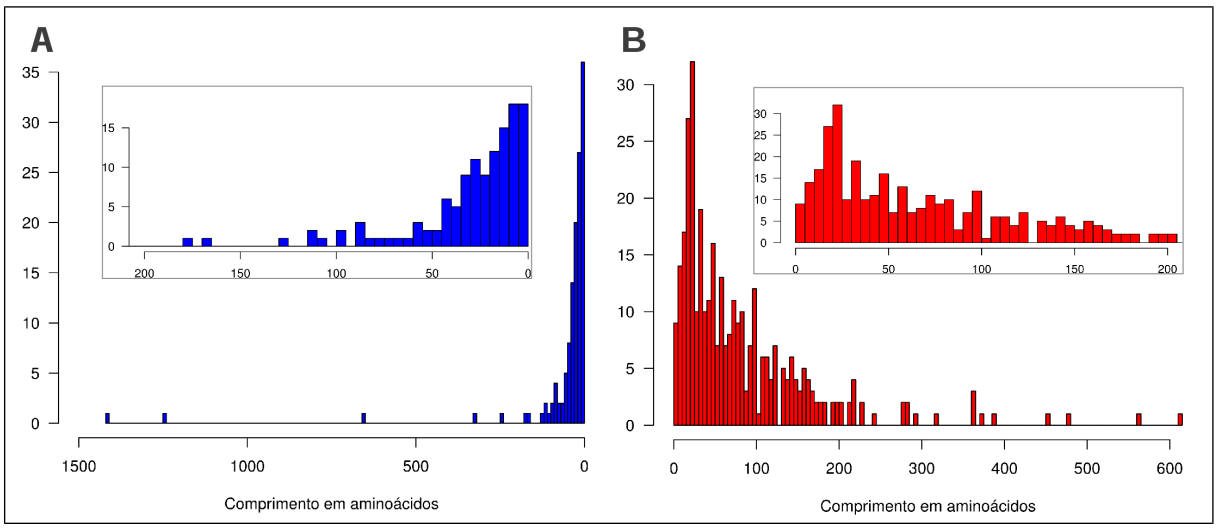


Figura 15: Tamanho dos segmentos de aminoácidos inseridos ou retirados nos processos de reanotação das proteínas. (A) Tamanhos dos segmentos adicionados nas reanotações das proteínas, em detalhe a faixa de maior concentração de alterações (até 200 aminoácidos). **(B)** Tamanho dos segmentos removidos durante os processos de reanotação das proteínas, em detalhe a faixa de maior concentração de alterações (até 200 aminoácidos).

A **Figura 16** mostra as distâncias entre as posições das novas metioninas iniciais em relação à metionina originalmente proposta. Para as novas metioninas que se encontravam em coordenadas anteriores no genoma (**Figura 16A**), a grande maioria foi localizada a uma distância inferior a 1000 bases, demonstrando que a escolha de 2000 bases extras na extremidade 5' para cada gene foi satisfatória para permitir as reanotações, sendo que somente em quatro casos houve a necessidade de extensão da região flanqueadora além das 2000 bases (**Figura 16A**). Nos casos onde a metionina alternativa foi localizada em região posterior à metionina original (**Figura 16B**), quanto menores as distâncias maiores foram suas frequências, com um pico entre 50 e 100 bases. Estes resultados revelam quais as falhas mais comuns do processo de predição gênica automatizada e podem, portanto, auxiliar no aperfeiçoamento destas metodologias, ao indicar regiões onde a probabilidade de se encontrar metioninas iniciais alternativas seria maior.

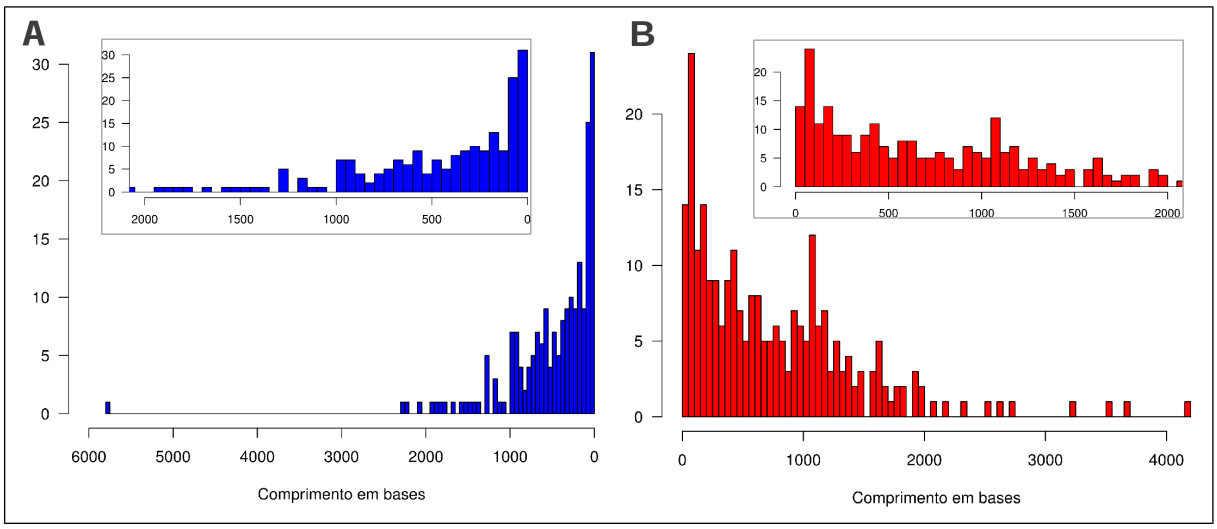


Figura 16: Distâncias entre a localização dos códons da metionina inicial do modelo gênico original e da metionina inicial proposta no novo modelo gênico. As reanotações resultaram no deslocamento da posição da metionina inicial para posições anteriores (*upstream*) (A) ou posteriores (*downstream*) (B) da metionina inicial original. Em detalhe um aumento na escala até 2.000 bases.

5.9. Os novos modelos gênicos têm suporte experimental

Mudanças no número de exons decorrentes das reanotações possibilitaram testar, através da amplificação por RT-PCR, se os modelos gênicos propostos estariam em concordância com as sequências dos RNAs mensageiros sendo expressos. Ao todo, considerando as cinco espécies de *Plasmodium*, 218 proteínas reanotadas eram elegíveis para esta validação. Particularmente em *P. vivax* (a espécie mais prevalente no Brasil), havia 60 proteínas elegíveis e destas, 7 foram selecionadas para a validação, sendo que três tiveram suas predições alteradas de positivas para negativas (PVX_081500, PVX_118150 e PVX_116975), três foram alteradas de negativas para positivas (PVX_083025, PVX_002580 e PVX_100770) e uma manteve sua predição negativa, mesmo após ter sido reanotada (PVX_083205) (**Tabela 3**).

Tabela 3: Proteínas selecionadas para validação experimental por RT-PCR dos novos modelos gênicos propostos

ID do gene	Descrição do produto (BDA*)	Alteração[a]	Grupo	Reclassificação do grupo[b]
PVX_081500	Proteína associada a adenilil ciclase	+ → -	OG4_11290	Negativo
PVX_083205	Proteína de trasnporte Sec61 subunidade alfa	-	OG4_10575	Negativo
PVX_083025	Proteína do micronema de esporozoíto	- → +	OG4_54960	Positivo
PVX_002580	Pseudouridina sintetase	- → +	OG4_11573	Misto
PVX_100770	Proteína conservada de *Plasmodium*	- → +	OG4_44949	Positivo
PVX_116975	Protéina hipotética, conservada em espécies de *Plasmodium*	+ → -	OG4_54958	Negativo
PVX_118150	Glutamina ciclotransferase	+ → -	OG4_16178	Negativo

** BLAST Description Annotator*

[a] Predição de peptídeo sinal pelo SignalP (+, predição positiva, -, predição negativa)

[b] Classificação dos grupos de acordo com as predições das proteínas ortólogas

Para todas as sete proteínas, a validação demonstrou que os novos modelos gênicos encontram suporte nos transcritos expressos por *P. vivax*, enquanto nenhum dos modelos originais foi capaz de gerar amplificações (**Figura 17**). A validação por RT-PCR não permite a identificação exata da metionina inicial, mas indica de forma inequívoca que os modelos gênicos propostos são alternativas mais plausíveis que os modelos originais.

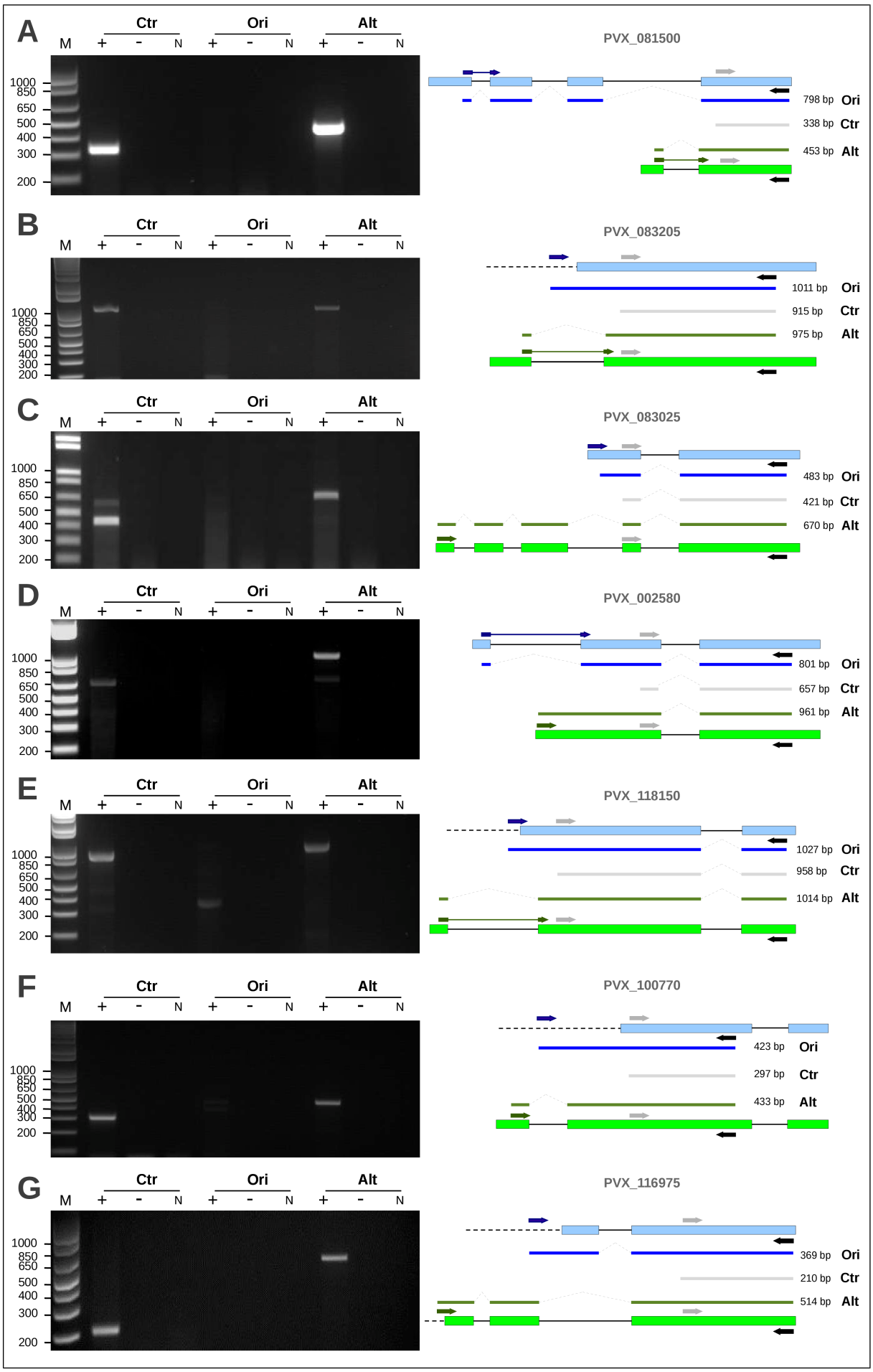

A
Ctr
Ori
Alt
M + - N + - N + - N
1000
850
650
500
400
300
200
PVX_081500
798 bp Ori
338 bp Ctr
453 bp Alt
B
PVX_083205
1011 bp Ori
915 bp Ctr
975 bp Alt
C
PVX_083025
483 bp Ori
421 bp Ctr
670 bp Alt
D
PVX_002580
801 bp Ori
657 bp Ctr
961 bp Alt
E
PVX_118150
1027 bp Ori
958 bp Ctr
1014 bp Alt
F
PVX_100770
423 bp Ori
297 bp Ctr
433 bp Alt
G
PVX_116975
369 bp Ori
210 bp Ctr
514 bp Alt

Figura 17: Validação através de RT-PCR de novos modelos gênicos de sete proteínas de *Plasmodium vivax* reanotadas. Nos painéis da esquerda, géis de agarose corados com GelRed apresentando os resultados das amplificações utilizando iniciadores controle (Ctr), modelo original (ori) e do novo modelo gênico (Alt), na presença (+) ou ausência (-) de transcriptase reverse na preparação do cDNA. Nos painéis da direita estão representados os modelos gênicos originais (azul claro) e propostos após a reanotação (verde claro). Os tamanhos dos fragmentos esperados nas amplificações estão indicados à direita das representações dos fragmentos em azul (modelo original) cinza (controle) e verde (novo modelo). M – marcador de peso molecular, 1 Kb plus (Invitrogen). N – controle negativo da amplificação na ausência de cDNA. Os genes submetidos à validação foram (A) PVX_081500, (B) PVX_083205, (C) PVX_083025, (D) PVX_002580, (E) PVX_118150, (F) PVX_100770, (G) PVX_116975.

5.10. Novas predições têm suporte de anotações funcionais

Como a presença ou ausência do peptídeo sinal influencia o papel biológico exercido por uma proteína, foi feita a análise das informações sobre a anotação funcional das proteínas reanotadas. Primeiramente, foram analisadas evidências experimentais de localização subcelular para as proteínas reanotadas. A recuperação de dados de localização foi feita através do banco de dados público ApiLoc, que apresenta uma compilação de dados retirados da literatura sobre experimentos de localização de proteínas dos organismos do Filo Apicomplexa. Somente 8 das 478 proteínas reanotadas já haviam sido previamente submetidas a experimentos que confirmassem a sua localização subcelular. Todas as 8 proteínas apresentam predições positivas para peptídeo sinal, sendo que 6 se tornaram positivas somente após as reanotações (**Tabela 4**). A descrição da localização subcelular destas proteínas sugere que a presença de peptídeo sinal seria necessária para sete delas, sendo que a descrição de PF14_0517 também sugere uma localização no citosol (**Tabela 4**), que seria independente de peptídeo sinal. Uma destas proteínas, PVX_090075, localiza-se nas roptrias e foi caracterizada como um promissor candidato vacinal, capaz de estimular uma resposta imune humoral e a proliferação de linfócitos em pacientes humanos (MONGUI et al., 2009). A única descrição aparentemente conflitante com a predição de peptídeo sinal foi a de PFB0400w (*Citoplasma durante gametócito estágio V*), que, segundo o ApiLoc, se localizaria no citoplasma de gametócitos maduros, porém esta proteína já apresentava uma predição positiva mesmo antes da sua reanotação e apesar de se encontrar no citoplasma, o seu padrão de localização é focal, sugerindo que esta proteína se localize dentro de vesículas citoplasmáticas (EKSI; WILLIAMSON, 2002).

Tabela 4: Confirmação experimental da localização subcelular de proteínas após reanotação, segundo o banco de dados ApiLoc

ID do Gene	Espécie	Predição de peptídeo sinal*		Localização segundo ApiLoc	Referência
		Antes	Depois		
PF14_0517	*Pf*	-	+	*Citosol e vacúolo digestivo*	(OLIVIERI et al., 2009)
PFB0400w	*Pf*	+	+	*Citoplasma durante gametócito estágio V*	(EKSI; WILLIAMSON, 2002)
PVX_090075	*Pv*	-	+	*Roptria durante esquizonte*	(MONGUI et al., 2009)
PY03011	*Py*	-	+	*Apical e basal e não núcleo durante esporozoíto de glândula salivar*	(KAISER; MATUSCHEWSKI; et al., 2004)
PY00454	*Py*	+	+	*Micronema durante esporozoíto*	(KAISER; CAMARGO; et al., 2004)
PY00819	*Py*	-	+	*Apicoplasto durante esquizonte hepático e esporozoíto de glândula salivar*	(PEI et al., 2010)
PY07092	*Py*	-	+	*Apical e não superfície durante esporozoíto de glândula salivar*	(KAISER; MATUSCHEWSKI; et al., 2004)
PY04986	*Py*	-	+	*Apical durante esporozoíto oocisto*	(MIKOLAJCZAK et al., 2008)

* Predição de peptídeo sinal usando o programa SignalP indicando a presença (+) ou asência (-) do peptídeo sinal antes e depois da reanotação

A mesma análise de evidências experimentais sobre a localização subcelular, através do ApiLoc, foi realizada para os ortólogos das proteínas reanotadas. Nesta análise, as predições das proteínas reanotadas foram comparadas às descrições das localizações de seus órtologos, desta forma, mais uma vez, recorreu-se à ortologia para auxiliar na definição de propriedades das proteínas reanotadas. Ao todo, 57 proteínas reanotadas apresentaram pelo menos um ortólogo cuja localização pôde ser recuperada do ApiLoc (**Tabela 5**).

Quarenta destas proteínas apresentaram, após as reanotações, predições positivas e para 39 (98%) delas a localização descrita para os seus ortólogos no ApiLoc sugere a presença de peptídeos sinal. PVX_117660, uma serina hidroximetiltransferase, foi a única proteína, dentre estas 40, cuja predição claramente não concordou com a descrição obtida para sua ortóloga, PF14_0534, encontrada na mitocôndria e que apresenta uma predição negativa.

As demais 17 proteínas reanotadas apresentaram predições negativas (**Tabela 5**). Para 8 destas proteínas, houve a concordância com a localização experimentalmente validada para suas ortólogas (**Tabela 5**), enquanto para 9 proteínas não houve concordância entre predição e localização da sua ortóloga. Porém, 5 destas 9 proteínas discordantes já apresentavam predições negativas mesmo antes das reanotações, e a predição de peptídeo sinal para as ortólogas de 8 destas 9 proteínas também é negativa, mesmo suas descrições sugerindo localizações dependentes da presença de peptídeos sinal. Esta menor concordância observada entre proteínas com predição negativa pode ser devido à existência de vias de transporte não dependentes de peptídeos sinal, porém vale ressaltar que mesmo a concordância tendo sido menor, os grupos ortólogos aos quais pertencem estas proteínas ainda foram, em sua maioria, reclassificados como negativos (**Tabela 5**), demonstrando a coerência entre predições de proteínas ortólogas.

Aproximadamente 60% das proteínas reanotadas estavam descritas no PlasmoDB

(7.1) como desconhecidas ou hipotéticas. Após a submissão das sequências revisadas ao programa *Blast Description Annotator* do pacote Blast2go, a taxa de proteínas desconhecidas/hipotéticas caiu para aproximadamente 35%. Após esta revisão das descrições de produtos proteicos, foi feita uma análise mais detalhada das proteínas reanotadas de *P. vivax* em relação à concordância de suas descrições com as novas predições de peptídeo sinal. Do total de 158 proteínas reanotadas para a espécie, 53 (33,5%) ainda foram descritas como hipotéticas ou desconhecidas. Entre as demais 105 proteínas, 78 (49,3%) apresentaram descrições de produtos gênicos condizentes com suas predições, 15 (9,4%) apresentaram descrições conflitantes e 12 (7,6%) foram consideradas inconclusivas, pois suas descrições não puderam ser atribuídas a nenhuma localização subcelular específica.

Tabela 5: Proteínas reanotadas que apresentaram ortólogas com validação experimental segundo o ApiLoc

Grupo	Classificação após reanotação	Proteínas reanotadas				Proteínas ortólogas			Localização segundo o ApiLoc
		ID do gene	Espécie	Predição		IDs dos genes	Espécies	Predições	
				Antes	Depois				
OG4_10640	Misto	PBANKA_040570	*Pb*	+	+	PFC0310c	*Pf*	+	*apicoplast and not mitochondrion during early schizont and trophozoite, apicoplast lumen*
OG4_19172	Positivo	PKH_126600	*Pk*	-	+	PF14_0382	*Pf*	+	*apicoplast and not mitochondrion during schizont and trophozoite*
OG4_22398	Positivo	PKH_124170	*Pk*	+	+	PF14_0607	*Pf*	+	*erythrocyte cytosol during 18-36 hours after merozoite invasion, parasitophorous vacuole*
OG4_26765	Positivo	PKH_131750	*Pk*	-	+	PFB0100c	*Pf*	+	*erythrocyte cytosol during early schizont, erythrocyte plasma membrane during ring and schizont and trophozoite, cytoplasmic side of erythrocyte membrane and maurers cleft, erythrocyte cytoplasm and food vacuole during late trophozoite, parasitophorous vacuole during early trophozoite and late trophozoite and ring and trophozoite, not erythrocyte cytoplasm during early trophozoite and ring, parasite plasma membrane during schizont and trophozoite*
OG4_48202	Positivo	PKH_102490	*Pk*	-	+	PFE0395c	*Pf*	+	*merozoite surface during late schizont and merozoite, apical during early schizont and late schizont and merozoite, not parasite plasma membrane during early schizont*
OG4_48203	Positivo	PKH_041100	*Pk*	-	+	PFB0405w	*Pf*	+	*parasite plasma membrane during female gametocyte and gamete and gametocyte and gametocyte stage iv and male gametocyte*
OG4_10180	Positivo	PVX_082960	*Pv*	-	+	MAL13P1.206	*Pf*	+	*parasite plasma membrane during intraerythrocytic*
OG4_10672	Positivo	PVX_091840	*Pv*	-	+	PF11_0256; PY00819	*Pf; Py*	+; +	*nowhere except apicoplast during intraerythrocytic; nowhere except apicoplast during hepatocyte schizont and salivary gland sporozoite, during not schizont*
OG4_10812	Positivo	PVX_092685	*Pv*	-	+	PF11_0427	*Pf*	+	*endoplasmic reticulum*
OG4_10837	Positivo	PVX_084510	*Pv*	-	+	PFL0480w	*Pf*	+	*mitochondrion, apicoplast and not mitochondrion during schizont and trophozoite*
OG4_11632	Positivo	PVX_098840	*Pv*	-	+	PFI0380c	*Pf*	+	*during not schizont and not trophozoite, apicoplast during ring*
OG4_12159	Positivo	PVX_100925	*Pv*	-	+	PFL1915w	*Pf*	+	*apicoplast during intraerythrocytic*
OG4_12828	Positivo	PVX_123795	*Pv*	-	+	PFL1120c	*Pf*	+	*apicoplast during intraerythrocytic*
OG4_19172	Positivo	PVX_118475	*Pv*	-	+	PF14_0382	*Pf*	+	*apicoplast and not mitochondrion during schizont and trophozoite*
OG4_21586	Positivo	PVX_091435	*Pv*	-	+	PF11_0168; PY02159	*Pf; Py*	+; +	*moving junction during invasion and merozoite invasion, rhoptry neck during extracellular merozoite and late schizont and schizont and segmenter, not rhoptry bulb during extracellular merozoite and segmenter, microneme and rhoptry during merozoite; rhoptry neck during schizont, apical during late schizont*
OG4_22398	Positivo	PVX_117270	*Pv*	+	+	PF14_0607	*Pf*	+	*erythrocyte cytosol during 18-36 hours after merozoite invasion, parasitophorous vacuole*
OG4_23359	Positivo	PVX_001780	*Pv*	-	+	PF10_0177b	*Pf*	+	*perinuclear during intraerythrocytic*
OG4_42735	Misto	PVX_000745	*Pv*	-	+	PFD0495c	*Pf*	-	*cytoplasm and erythrocyte cytoplasmic structure during intraerythrocytic, erythrocyte cytosol during 18-36 hours after merozoite invasion and intraerythrocytic*
OG4_43617	Misto	PVX_084315	*Pv*	-	+	PFL0285w	*Pf*	-	*apicoplast during merozoite and schizont and trophozoite*
OG4_48162	Misto	PVX_117660	*Pv*	+	+	PF14_0534	*Pf*	-	*residual body during very late post-mitotic schizont, mitochondrion during ring and schizont and trophozoite*
OG4_54213	Positivo	PVX_090075	*Pv*	-	+	PFD0955w	*Pf*	+	*rhoptry neck and not rhoptry bulb during late trophozoite or schizont, rhoptry during late schizont*
OG4_84674	Positivo	PVX_115165	*Pv*	-	+	PF13_0338	*Pf*	+	*during ring, merozoite surface during late schizont, parasite plasma membrane during merozoite*
OG4_10672	Positivo	PY00819	*Py*	-	+	PF11_0256	*Pf*	+	*nowhere except apicoplast during intraerythrocytic*
OG4_10812	Positivo	PY00944	*Py*	-	+	PF11_0427	*Pf*	+	*endoplasmic reticulum*
OG4_10856	Misto	PY02528	*Py*	-	+	PFF0360w	*Pf*	+	*mitochondrion during intraerythrocytic, apicoplast and not mitochondrion*
OG4_10917	Positivo	PY04302	*Py*	-	+	PF14_0381	*Pf*	+	*apicoplast and not mitochondrion during schizont and trophozoite*
OG4_15900	Positivo	PY06430	*Py*	-	+	PF14_0063	*Pf*	+	*apicoplast*
OG4_21677	Misto	PY00454	*Py*	+	+	PBANKA_100630	*Pb*	+	*microneme and not rhoptry during salivary gland sporozoite, during salivary gland sporozoite and not merozoite and not midgut sporozoite and not ookinete*
OG4_22398	Positivo	PY06776	*Py*	-	+	PF14_0607	*Pf*	+	*erythrocyte cytosol during 18-36 hours after merozoite invasion, parasitophorous vacuole*
OG4_23349	Positivo	PY05180	*Py*	-	+	PBANKA_082420	*Pb*	+	*microneme during ookinete*
OG4_23361	Misto	PY00441	*Py*	-	+	PFB0475c	*Pf*	+	*apical during extracellular merozoite and late schizont*
OG4_24016	Positivo	PY01850	*Py*	-	+	PF13_0180	*Pf*	+	*apicoplast and not mitochondrion during ring and schizont*
OG4_25190	Positivo	PY05738	*Py*	-	+	PFB0680w	*Pf*	+	*cytoplasm during trophozoite, during not late ring, parasitophorous vacuole during early ring, apical during schizont, rhoptry neck during merozoite*
OG4_25260	Positivo	PY03879	*Py*	-	+	PF10_0363	*Pf*	+	*apicoplast and not mitochondrion during intraerythrocytic*
OG4_27696	Positivo	PY00697	*Py*	-	+	PFE1460w	*Pf*	+	*apicoplast during trophozoite*
OG4_34347	Misto	PY02282	*Py*	-	+	MAL8P1.73	*Pf*	-	*rhoptry neck and not rhoptry bulb during extracellular merozoite*
OG4_42709	Positivo	PY01808	*Py*	-	+	PFL2505c	*Pf*	+	*parasitophorous vacuole during ring, rhoptry bulb and not rhoptry neck during schizont*
OG4_42735	Misto	PY04086	*Py*	-	+	PFD0495c	*Pf*	-	*cytoplasm and erythrocyte cytoplasmic structure during intraerythrocytic, erythrocyte cytosol during 18-36 hours after merozoite invasion and intraerythrocytic*
OG4_46055	Positivo	PY03873	*Py*	-	+	PBANKA_141830	*Pb*	+	*rhoptry during merozoite*
OG4_46501	Positivo	PY00539	*Py*	-	+	PF08_0067	*Pf*	+	*apicoplast*
OG4_10575	Negativo	PKH_120830	*Pk*	-	-	MAL13P1.231	*Pf*	-	*endoplasmic reticulum during ring and schizont and trophozoite*
OG4_10611	Misto	PKH_093070	*Pk*	-	-	PF11_0339	*Pf*	-	*mitochondrion during intraerythrocytic*
OG4_34375	Negativo	PKH_124540	*Pk*	+	-	PF14_0578	*Pf*	-	*inner membrane complex during extracellular merozoite and late schizont*
OG4_43617	Misto	PKH_130420	*Pk*	+	-	PFL0285w	*Pf*	-	*apicoplast during merozoite and schizont and trophozoite*
OG4_10425	Negativo	PVX_081665	*Pv*	+	-	MAL7P1.150	*Pf*	-	*mitochondrion and not apicoplast during schizont and trophozoite*
OG4_10575	Negativo	PVX_083205	*Pv*	-	-	MAL13P1.231	*Pf*	-	*endoplasmic reticulum during ring and schizont and trophozoite*
OG4_10790	Negativo	PVX_089930	*Pv*	+	-	PFD0810w	*Pf*	-	*cytoplasmic vesicle and erythrocyte cytoplasmic vesicle during ring and schizont, endoplasmic reticulum and not erythrocyte cytoplasm during merozoite and ring and schizont and trophozoite*
OG4_43900	Misto	PVX_097920	*Pv*	-	-	PFE0355c	*Pf*	-	*parasitophorous vacuole*
OG4_10575	Negativo	PY02510	*Py*	+	-	MAL13P1.231	*Pf*	-	*endoplasmic reticulum during ring and schizont and trophozoite*
OG4_10789	Negativo	PY05100	*Py*	+	-	PFD0311w	*Pf*	-	*cytosol and not food vacuole during ring and schizont and trophozoite*
OG4_10790	Negativo	PY04367	*Py*	-	-	PFD0810w	*Pf*	-	*cytoplasmic vesicle and erythrocyte cytoplasmic vesicle during ring and schizont, endoplasmic reticulum and not erythrocyte cytoplasm during merozoite and ring and schizont and trophozoite*
OG4_11252	Negativo	PY01870	*Py*	+	-	PBANKA_121770	*Pb*	-	*p-body during macrogametocyte*
OG4_11854	Misto	PY05787	*Py*	-	-	PF07_0047	*Pf*	+	*apicoplast*
OG4_12953	Negativo	PY05452	*Py*	+	-	PBANKA_144400; PFL1420w	*Pb; Pf*	-; -	*parasitophorous vacuole during schizont and trophozoite; maurers cleft during ring and trophozoite*
OG4_13750	Misto	PY03772	*Py*	-	-	PFL1465c	*Pf*	-	*mitochondrial matrix and mitochondrion and not mitochondrial membrane, cytosol during ring and schizont and trophozoite, during not ring*
OG4_23358	Negativo	PY03207	*Py*	+	-	PFI1565w	*Pf*	-	*non-invasion-related localisation and not apical and not inner membrane complex and not merozoite surface during extracellular merozoite and late schizont*
OG4_36196	Negativo	PY04971	*Py*	+	-	MAL13P1.185	*Pf*	-	*cytoplasm and nucleus during late ring and schizont and segmenter and trophozoite*

Para exemplificar o potencial da estratégia proposta, o significado biológico da reanotação de 10 proteínas de *P. vivax* foi analisado em um contexto comparativo com *P.*

falciparum. Foram analisadas 4 tRNAs sintetases, 1 tRNA amidotransferase, 2 Fatores de Iniciação da Tradução, 2 subunidades de DNA girases e 1 Ferrodoxina, sendo que todas estas proteínas apresentaram uma predição positiva para a presença de peptídeo sinal após serem reanotadas. As tRNAs sintetases, a amidotransferase e os Fatores de Iniciação estão envolvidos no processo de transcrição/tradução de proteínas, as girases participam da replicação de DNA e a exata função da Ferredoxina permanece desconhecida. As ortólogas destas 10 proteínas em *P. falciparum* apresentam peptídeos sinal e se localizam no lúmen do Apicoplasto. As 4 sintetases e as 2 subunidades das girases apresentam contrapartes, citossólicas e nucleares, respectivamente, que exercem a mesma função bioquímica porém em compartimentos celulares distintos. Estas proteínas não apresentam peptídeos sinal e pertencem a outros grupos ortólogos (**Tabela 6**). Ortólogas destas proteínas citossólicas e nucleares são também encontradas em *P. vivax* e não apresentam peptídeo sinal. Portanto, as reanotações das 10 proteínas de *P. vivax* e, consequentemente, as novas predições de petídeo sinal reproduziram nesta espécie um paralelismo perfeito do quadro encontrado em *P. falciparum* (**Tabela 6**), sugerindo que a correta localização destas proteínas em *P. vivax* seja também o Apicoplasto.

Tabela 6: Concordância das predições de peptídeos sinal entre proteínas de *Plasmodium vivax* reanotadas e suas ortólogas em *Plasmodium falciparum* sugerindo uma possível localização subcelular

Descrição da proteína	Grupo	Proteína em *P. falciparum*			Proteína em *P. vivax*	
		ID do gene	Predição	Localização	ID do gene	Predição*
Prolina tRNA sintetase	OG4_11662	PFI1240c	+	Apicoplasto	PVX_099680 ®	+
	OG4_10599	PFL0670c	-	Citosol	PVX_123380	-
Asparagina tRNA sintetase	OG4_44547	PFE0475w	+	Apicoplasto	PVX_098040 ®	+
	OG4_10251	PFB0525w	-	Citosol	PVX_002940	-
Metionina tRNA sintetase	OG4_47490	PF10_0053	+	Apicoplasto	PVX_094445 ®	+
	OG4_10161	PF10_0340	-	Citosol	PVX_110980	-
Leucina tRNA sintetase	OG4_11105	PF08_0011	+	Apicoplasto	PVX_088945 ®	+
	OG4_10828	PFF1095w	-	Citosol	PVX_114255	-
Glutamanto tRNA amidotransferase subunidade A	OG4_10803	PFD0780w	+	Apicoplasto	PVX_089895 ®	+
Fator de Iniciação-1	OG4_12019	MAL8P1.27	+	Apicoplasto	PVX_089095 ®	+
Fator de Iniciação-3	OG4_14117	PF14_0658	+	Apicoplasto	PVX_117015 ®	+
DNA girase subunidade A	OG4_12828	PFL1120c	+	Apicoplasto	PVX_123795 ®	+
	OG4_10491	PF14_0316	-	Núcleo	PVX_084855	-
DNA girase subunidade B	OG4_12159	PFL1915w	+	Apicoplasto	PVX_100925 ®	+
	OG4_10746	PF10_0412	-	Núcleo	PVX_111340	-
Ferredoxina	OG4_14702	MAL13P1.95	+	Apicoplasto	PVX_122725 ®	+

® Proteínas reanotadas

* Predição de peptídeo sinal baseada na nova sequência (caso tenha ocorrido reanotação)

5.11. Otimização dos valores de corte de predição de peptídeo sinal

Entre as possíveis explicações para a existência de grupos Mistos estão as limitações intrínsecas ao método de predição de peptídeos sinal. Mesmo em casos onde as

sequências estejam corretas, erros de predição poderiam criar grupos Mistos irreais ou, contrariamente, esconder grupos Mistos verdadeiros. Para tentar minimizar estes efeitos realizou-se a otimização das configurações de predição através da busca da melhor combinação de valores de corte para os três escores, NN-Sum, D-escore e Probabilidade de peptídeo sinal por HMM. A conservação esperada entre proteínas ortólogas foi, mais uma vez, o marco teórico que guiou a definição de condição ótima, postulada como a combinação de valores de corte onde a discordância entre ortólogos fosse mínima, ou seja o número de grupos Mistos observados fosse o menor possível. A otimização foi realizada após os resultados da reanotação de proteínas terem sido incorporados ao banco de dados.

Seriam necessárias aproximadamente 40.000 iterações para realizar a combinação de todos os possíveis valores dos três escores (considerando uma precisão de duas casas decimais para os valores de D-escore e Probabilidade sinal por HMM), portanto, para reduzir a demanda computacional e o tempo de análise, na primeira rodada de combinações de valores, o D-escore e a Probabilidade sinal por HMM foram combinados em intervalos de 0,05 unidades (de 0,05 a 1,0) enquanto o NN-Sum variou de 1 a 4, reduzindo para 1.600 o número de iterações. A partir desta primeira rodada, definiu-se que o valor ótimo para o NN-Sum igual a 4 (**Figura 18A**), pois o número mínimo de grupos Mistos encontrados neste primeiro momento foi 476 e foi obtido quando os valores de corte para NN-Sum, D-escore e Probabilidade sinal por HMM foram, respectivamente, 4, 0,45 e 0,90 (**Tabela 7**). É importante ressaltar que a contagem de números mínimos de grupos Mistos, durante as análises de otimização, levou em consideração os grupos contendo múltiplas proteínas por espécie, pois a mudança de valores de corte também os afetaria.

Uma vez definido o valor ótimo para o NN-Sum, foi realizada uma nova rodada de combinações, onde tanto o D-escore quanto a Probabilidade sinal por HMM foram testados com variações de 0,01 unidades em faixas mais restritas (D-escore: 0,40 a 0,55 e Probabilidade sinal HMM: 0,40 a 1,0), onde se concentravam as contagens mais baixas (**Figura 18A**).

Dentro desta faixa mais específica e com a varredura mais detalhada, o número mínimo de grupos Mistos encontrados foi 465 (**Figura 18B**), e as melhores combinações de valores de corte foram atingidas com o NN-Sum=4, o D-escore=0,48 e a Probabilidade sinal por HMM=0,87, 0,89 ou 0,90. A Probabilidade sinal por HMM de 0,90 foi selecionada para as demais análises por ser a mais representativa entre as três apresentadas.

Tabela 7: Número mínimo de grupos Mistos obtidos pelas combinações entre os parâmetros NN-sum, D-escore e Probabilidade sinal por HMM

NN-Sum	D-escore	Probabilidade sinal por HMM	Número mínimo de grupos Mistos
1	0,4 – 1,0	0,7 – 0,9	1104
2	0,45	0,65	558
3	0,45	0,85; 0,9	489
4	0,45	0,9	**476**

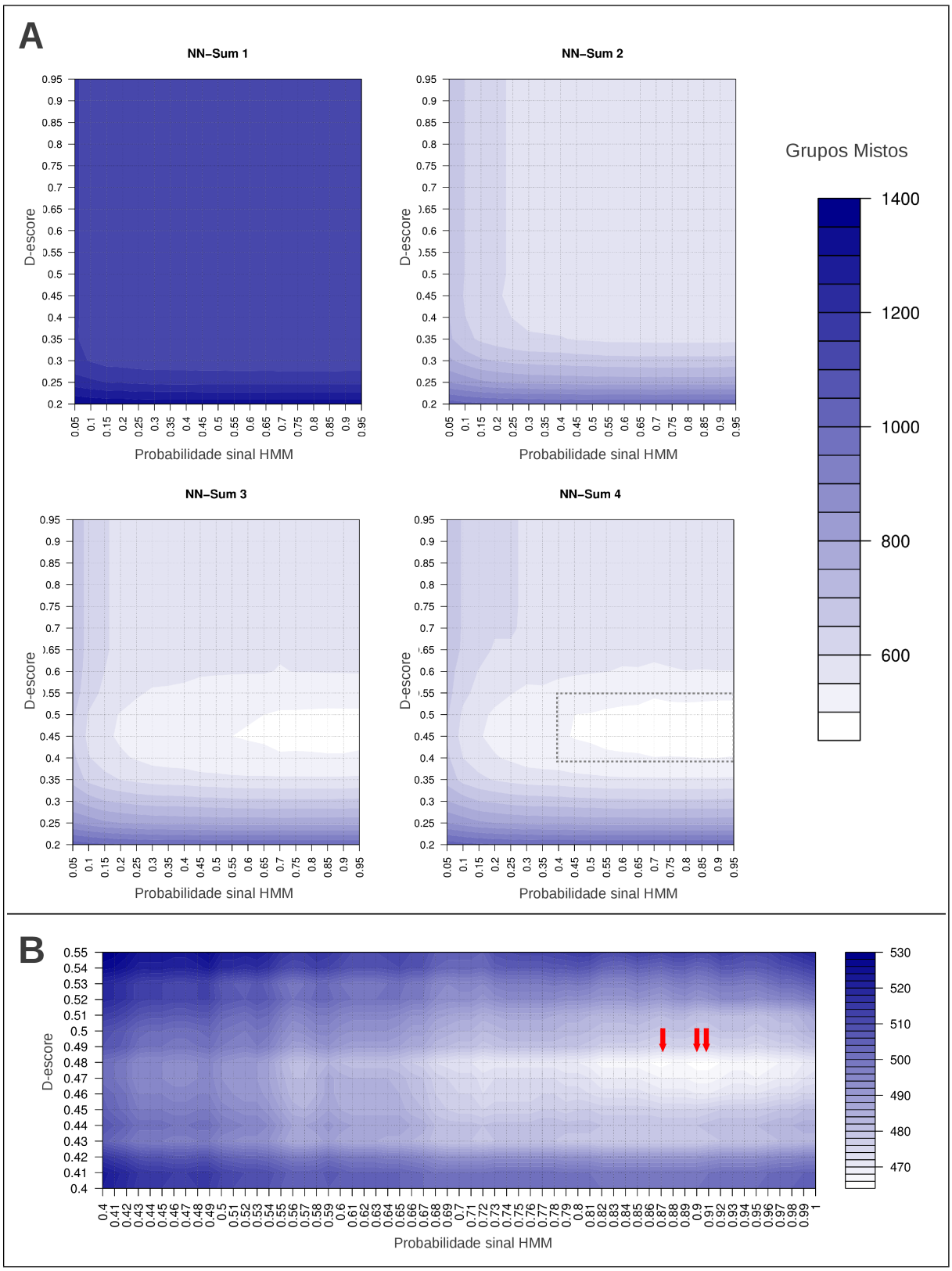


Figura 18: Otimização dos valores de corte para parâmetros de predição de peptídeo sinal. Diversas combinações para valores de corte dos três parâmetros usados na predição de peptídeos sinal foram testados. (A) NN-Sum variando de 1 a 4 e D-escore e Probabilidade sinal por HMM variando, ambos, de 0,05 a 1,0 (em intervalos de 0,05). (B) A área registrando as menores contagens de grupos Mistos (retângulo cinza tracejado em A) foi reanalizada com variações mais finas de D-escore e Probabilidade sinal por HMM (intervalos de 0,01) e NN-Sum fixado em 4. As menores contagens de grupos Mistos (setas vermelhas) foram obtidas com os valores de corte configurados

em: NN-Sum = 4; D-escore = 0,48; Probabilidade sinal por HMM = 0,87; 0,9; 0,91.

A aplicação dos novos valores de corte aos 4319 grupos ortólogos (sem múltiplas proteínas por espécie) provocou a redução na contagem de grupos Mistos de 289 (após reanotações) para 262 (**Figura 19**), passando a representar somente 6% dos grupos ortólogos.

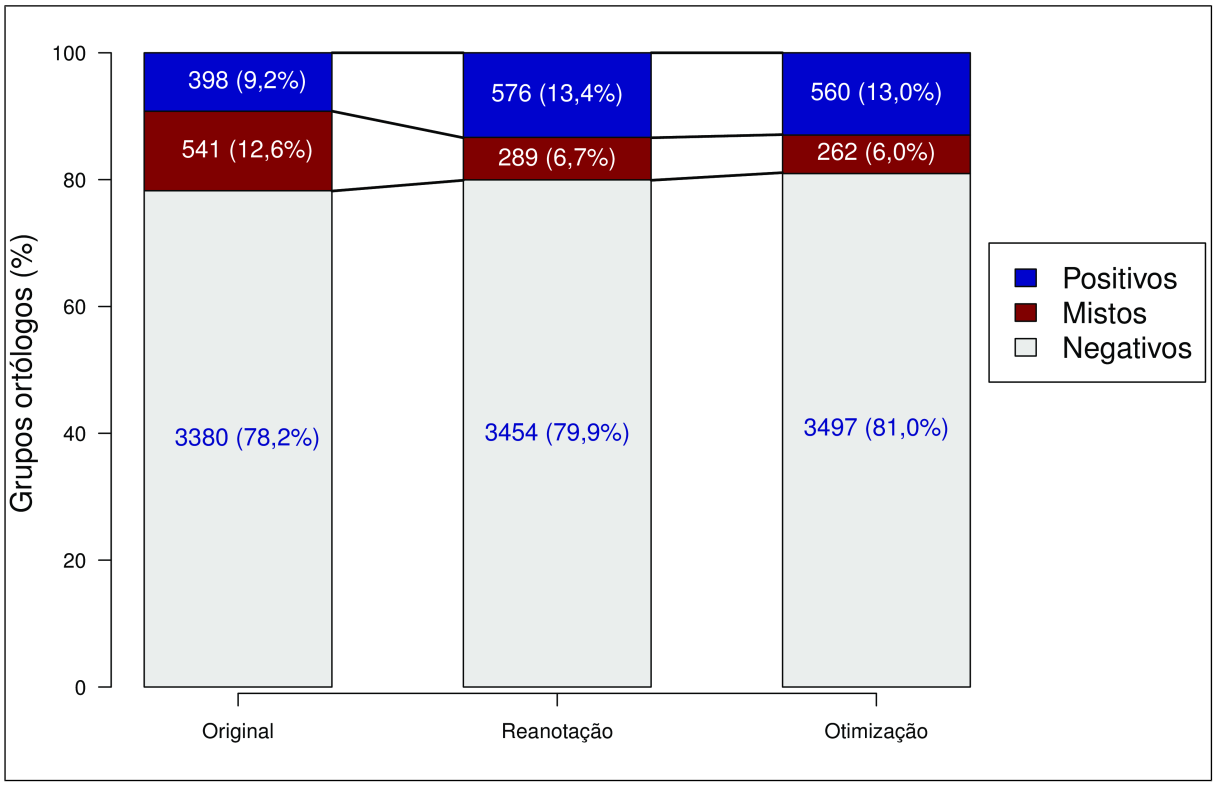


Figura 19: Reclassificação de grupos ortólogos após a otimização das configurações de predição de peptídeo sinal. Após a reanotação de proteínas, novos valores de corta para os parâmetros de predição foram selecionados e os resultados das predições foram usados na reclassificação de grupos. Os novos valores de corte para os parâmetros de predição foram: NN-Sum = 4; D-escore = 0,48; Probabilidade sinal por HMM = 0,9.

A otimização resultou na reclassificação de 61 grupos Mistos em Positivos (11) e Negativos (50), porém, as mudanças nos valores de corte também foram responsáveis pela reclassificação de 34 grupos Positivos (26) e Negativos (6) para Mistos (**Tabela 8**), justificando a pequena diferença observada entre a contagem de grupos Mistos após reanotações e após a otimização. Entre estes 34 grupos reclassificados como Mistos, 8, na verdade, retornaram à sua classificação original (Mistos), uma vez que havia sido reclassificados como Positivos após reanotações. Os demais 26 grupos reclassificados como Mistos foram submetidos aos procedimentos de inspeção visual e, em 13 (50%) destes grupos, constatou-se a presença de proteínas mal anotadas, indicando que a otimização foi vantajosa, pois revelou a necessidade de reanotação de grupos que haviam

sido previamente ignorados por não serem originalmente classificados como Mistos. Em uma observação interessante, a otimização das configurações de predição resultou na reclassificação de grupos Mistos (61/289, 21,1%) em proporções superiores às de grupos Positivos (26/576, 4,5%) e Negativos (8/3454, 0,002%), sugerindo que a otimização atue preferencialmente em grupos nos quais divergências de predição entre ortólogos são mais comumente observadas.

Tabela 8: Trocas de classes provocadas pela otimização dos configurações de predição de peptídeo sinal

Classes	Antes da otimização*	Alteração de classes após a otimização**			
		Negativos	Positivos	Mistos	TOTAL
Negativos	3454	-	0	8	3496
Positivos	576	0	-	26	561
Mistos	289	50	11	-	262
TOTAL	4319	50	11	34	4319

* Configurações da predição de peptídeo sinal: NN-Sum = 3; D-escore = 0,5; Probabilidade sinal por HMM = 0,5

** Configurações da predição de peptídeo sinal: NN-Sum = 4; D-escore = 0,48; Probabilidade sinal por HMM = 0,9

5.12. Padrões de predição refletindo a filogenia de *Plasmodium* são mais frequentes em grupos consistentemente classificados como Mistos

A reclassificação de grupos Mistos foi alcançada, principalmente, pela reanotação de proteínas e, em menor grau, através de modificações dos parâmetros de predição de peptídeos sinal. Entretanto, 228 grupos mantiveram a sua classificação como Mistos durante todo o processo. Entre estes 228 grupos, 87 (70 grupos Parcialmente reanotados e 17 grupos Inconclusivos) ainda podem ser reclassificados futuramente e foram excluídos desta análise. Sobre os 141 grupos restantes, pode-se afirmar que mantiveram a sua classificação (Mistos) mesmo após terem sido submetidos à inspeção, reanotação ou otimização dos parâmetros de predição. Portanto, de acordo com a hipótese original, é plausível que a fonte das divergências de predição observadas dentro destes grupos seja uma diversidade biológica real, fruto do processo de evolução diferencial de proteínas ortólogas que as levam a, genuinamente, divergirem quanto à presença de um peptídeo sinal. Um indício que a divergência entre predições de peptídeos sinal para ortólogos possa ser verdadeira seria a distribuição das predições em um perfil coerente com as relações evolutivas entre as espécies do gênero. Existem três padrões de combinações de predições positivas e negativas que refletem a filogenia do gênero *Plasmodium*, com predições divergentes sendo restritas somente: **1-** ao *P. falciparum*; **2-** às espécies que parasitam roedores (*P. berghei* e *P. yoelii*); ou **3-** ao *P. vivax* e *P. knowlesi* (**Figura 20A**).

Interessantemente, os 141 grupos que foram classificados consistentemente como Mistos mostram uma maior proporção destes padrões filogenéticos quando comparados com 301 grupos que foram originalmente classificados como Mistos mas, por reanotação ou otimização, foram reclassificados (**Figura 20B**). Esta correspondência entre a predição de peptídeos sinal e a filogenia dá suporte à ideia de que estes grupos podem apresentar novidades evolutivas.

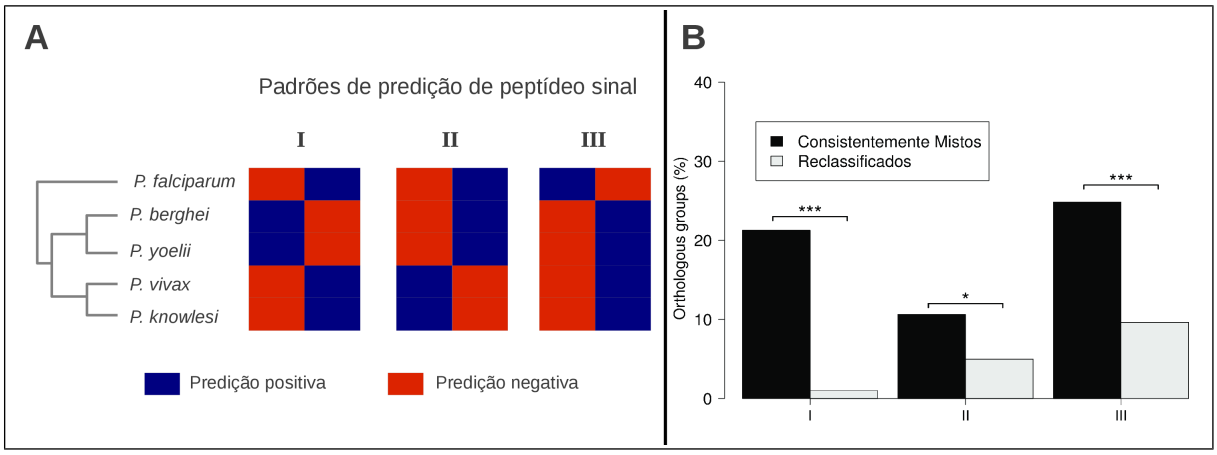


Figura 20: Padrões de predição de peptídeo sinal em grupos Mistos entre as espécies de *Plasmodium*. (A) Três padrões de predição de peptídeo sinal que refletem a filogenia do gênero *Plasmodium*. Padrões com predições diferentes restritas a: **I** - *P. berghei* e *P. yoelii*; **II** - *P. vivax* e *P. knowlesi*; **III** – *P. falciparum*. (B) As proporções destes padrões (I, II e III) foram comparadas entre grupos Mistos que foram reclassificados por reanotação ou otimização da predição de peptídeo sinal (N = 301) e grupos que foram consistentemente classificados como Mistos (N = 141). A significância estatística foi analisada através do Qui-quadrado das diferenças entre proporções (*p < 0,05; ***p < 0,0001).

PARTE II – Automatização da identificação de grupos com erros de anotação

5.13. Descrição dos bancos de dados

Assim como foi feito previamente para as espécies de *Plasmodium*, as informações de ortologia e as predições de peptídeo sinal foram combinadas para os conjuntos de espécies (ou cepas) de cada um dos demais cinco bancos de dados: **Toxoplasma**, **Cryptosporidium**, **Trypanosoma**, **Leishmania** e **Apicomplexa**. Após a classificação dos grupos ortólogos e a exclusão daqueles grupos contendo múltiplas proteínas por espécie, foi revelado o perfil de cada conjunto de dados em relação às distribuições das classes de grupos ortólogos (**Tabela 9**). Os valores referentes ao banco de dados **Plasmodium**, antes de reanotações e otimização, são fornecidos para efeito comparativo. A menor taxa de grupos Mistos foi observada no banco de dados **Toxoplasma**, justamente o banco em que, ao invés de proteínas provenientes de diferentes espécies, estão agrupadas proteínas de diferentes cepas de uma única espécie. Vale ressaltar que no banco **Toxoplasma** as proteínas das três cepas *não são ortólogas entre si*, uma vez que esta nomenclatura

descreve relações entre proteínas de espécies distintas, e sim alelos de uma mesma proteína. Porém a estratégia de combinação de informações pode ser empregada nesta situação sem qualquer prejuízo para o objetivo final que é a reanotação de sequências proteicas, demonstrando mais uma possível aplicação da metodologia proposta. À medida que o número de espécies aumentou, a proporção de grupos Mistos também cresceu, com exceção do banco de dados **Plasmodium**, que mesmo sendo formado por 5 espécies apresentou uma taxa de grupos Mistos bem inferior aos bancos com 5 ou 4 espécies (**Tabela 9**), o que talvez seja um reflexo do interesse diferenciado que as espécies do gênero *Plasmodium* recebem por parte da comunidade científica.

Tabela 9: Classificação dos grupos órtologos

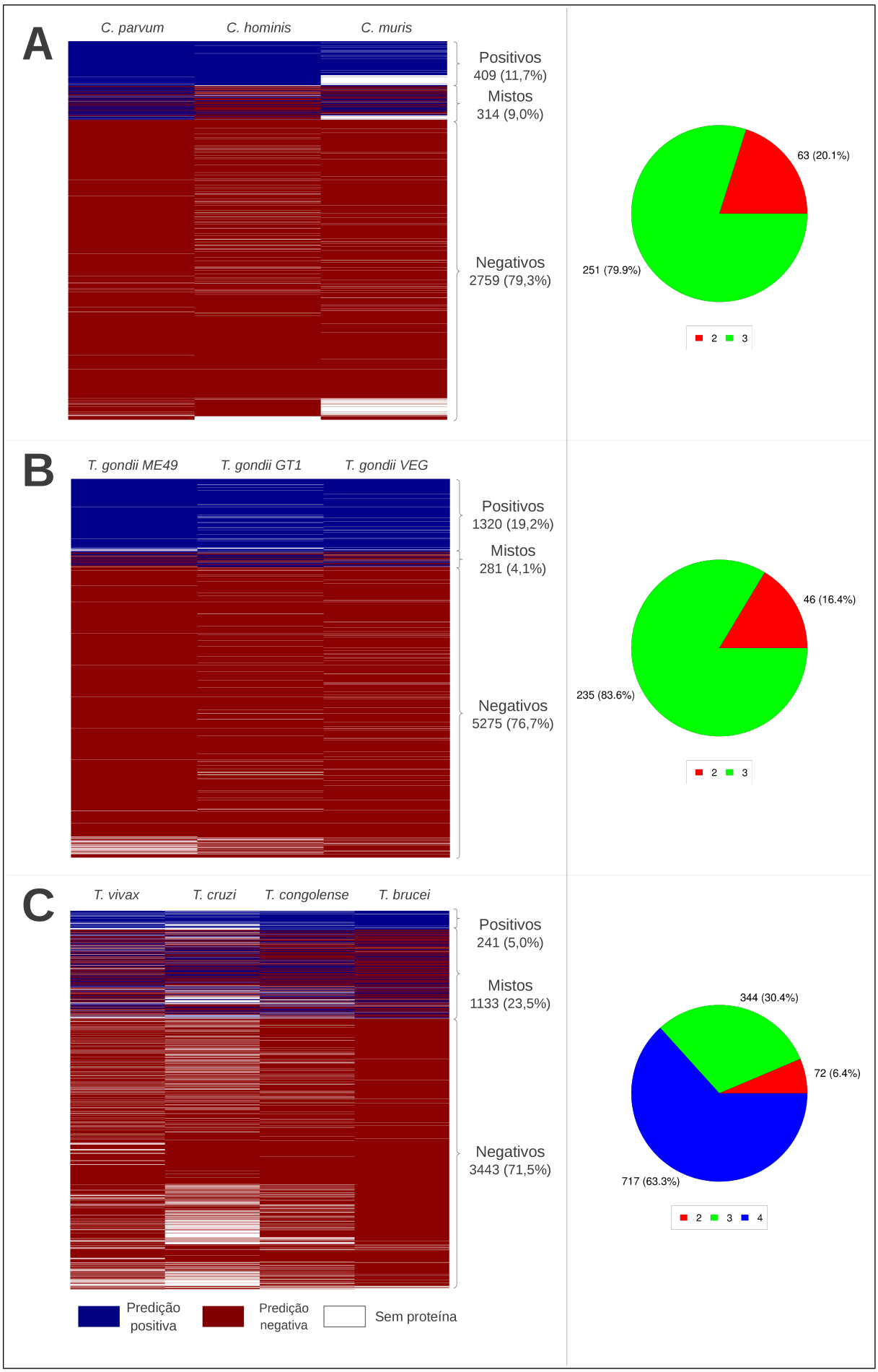
Banco de dados	Espécies	Classes de grupos ortólogos						TOTAL
		Positivos		Mistos		Negativos		
Toxoplasma	*Toxoplasma gondii***	1320	*19,2%*	281	*4,1%*	5275	*76,7%*	6876
Cryptosporidium	*Cryptosporidium parvum, C. hominis, C. muris*	409	*11,7%*	314	*9,0%*	2759	*79,2%*	3482
Trypanosoma	*Trypanosoma brucei, T. cruzi, T. vivax, T. congolense*	241	*5,0%*	1133	*23,5%*	3443	*71,5%*	4817
Leishmania	*Leishmania major, L. infatum, L. braziliensis, L. mexicana, L. tarentolae*	486	*7,1%*	1797	*26,1%*	4589	*66,8%*	6872
Apicomplexa	*Eimeria tenella, Neorospora caninum, Theileria annulata, Babesia bovis, Toxoplasma gondii****	586	*9,5%*	1514	*24,6%*	4047	*65,8%*	6147
Plasmodium*	*Plasmodium falciparum, P. vivax, P. knowlesi, P. berghei, P. yoelii*	398	*9,2%*	541	*12,5%*	3380	*78,3%*	4319

* Números originais, anteriores às reanotações e otimização de parâmetros de predição

** Cepas ME49, GT1 and VEG de Toxoplasma gondii

*** Cepa ME49 de Toxoplasma gondii

É possível observar que em quase todos os bancos de dados os grupos ortólogos são, em sua maioria, bastante representativos, apresentando proteínas de todas as espécies (ou cepas) recrutadas para a sua construção (**Figura 21A-D**). Esta representatividade é mantida, inclusive entre os grupos Mistos (**Figura 21A-D**). A única exceção é o banco de dados **Apicomplexa**, que apresenta grupos de composição extremamente variada (**Figura 21E**), e grupos Mistos formados, em sua maioria por 2 ou 3 proteínas (**Figura 21E**).


A
C. parvum
C. hominis
C. muris
Positivos
409 (11,7%)
Mistos
314 (9,0%)
Negativos
2759 (79,3%)
63 (20.1%)
251 (79.9%)
2
3
B
T. gondii ME49
T. gondii GT1
T. gondii VEG
Positivos
1320 (19,2%)
Mistos
281 (4,1%)
Negativos
5275 (76,7%)
46 (16.4%)
235 (83.6%)
2
3
C
T. vivax
T. cruzi
T. congolense
T. brucei
Positivos
241 (5,0%)
Mistos
1133 (23,5%)
Negativos
3443 (71,5%)
344 (30.4%)
72 (6.4%)
717 (63.3%)
2
3
4
Predição positiva
Predição negativa
Sem proteína

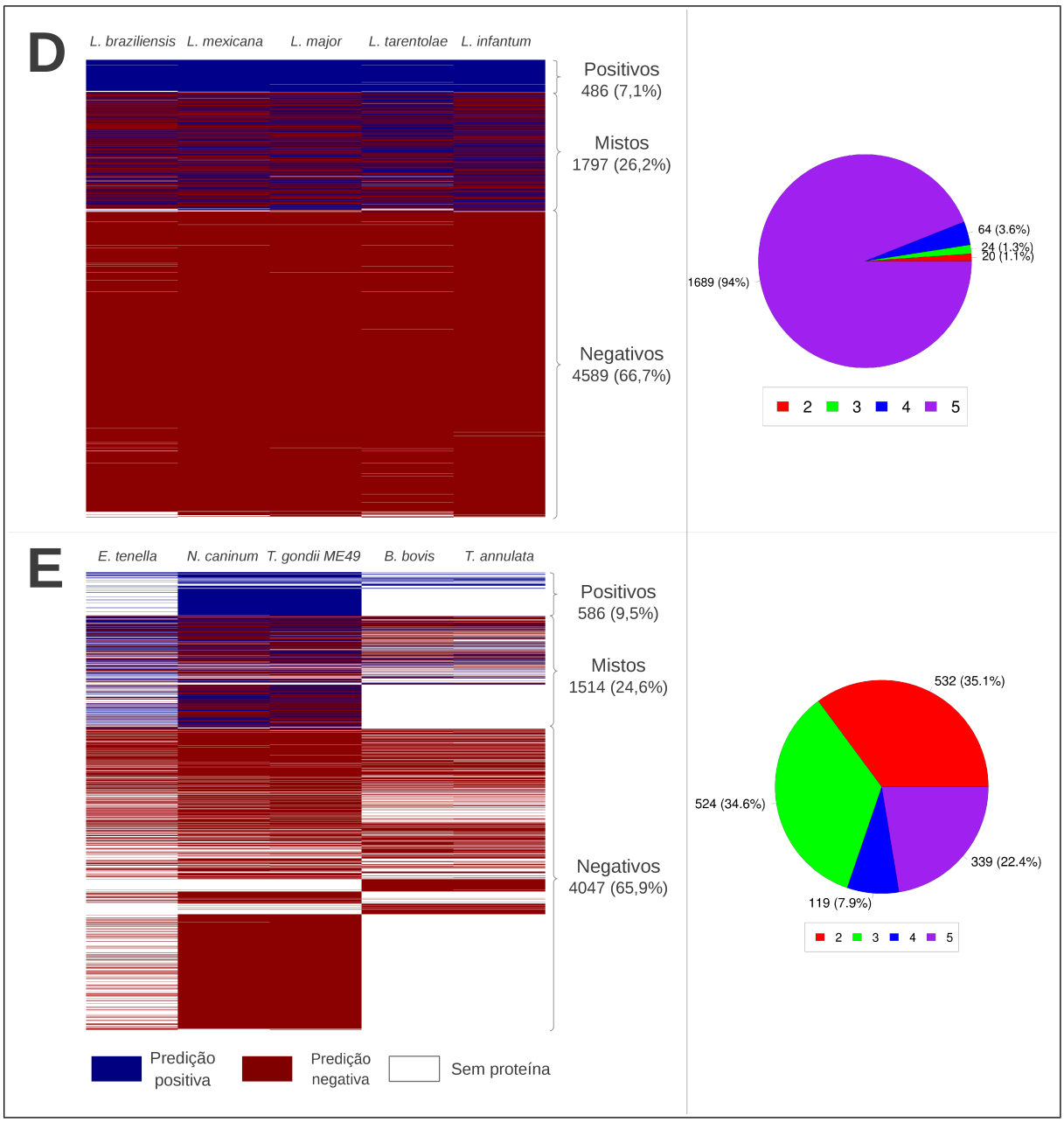


Figura 21: **Distribuição dos grupos ortólogos de diferentes bancos de dados classificados de acordo com a predição de peptídeo sinal de suas proteínas.** (A) Banco de dados **Toxoplasma**: Composição dos grupos por proteínas das diferentes cepas de *Toxoplasma gondii*: ME49, GT1 e VEG. (B) Banco de dados **Cryptosporidium**: Composição dos grupos por proteínas das diferentes espécies do gênero *Cryptosporidium*: *C. parvum*, *C. hominis* e *C. muris*. (C) Banco de dados **Trypanosoma**: Composição dos grupos por proteínas das diferentes espécies do gênero *Trypanosoma*: *T. vivax*, *T. Cruzi*, *T. congolenses* e *T. brucei.* (D) Banco de dados **Leishmania**: Composição dos grupos por proteínas das diferentes espécies do gênero *Leishmania*: *L. braziliensis*, *L. mexicana*, *L. major*, L. *tarentolae* e *L. infantum*. (E) Banco de dados **Apicomplexa**: Composição dos grupos por proteínas das diferentes espécies: *Eimeria tenella*, *Neospora caninum*, *Toxoplasma gondii* (cepa ME49), *Babesia bovis* e *Theleria annulata*. **Painéis da esquerda:** Linhas horizontais representam cada grupo ortólogo e as cores representam as predições de peptídeo sinal positivas (azul), negativas (vermelho) ou a ausência (branco) da proteína da espécie/cepa no grupo. Foram utilizadas as configurações padrão do SignalP implementadas no banco de dados PlasmoDB ou EuPathDB para definir predições positivas: NN-Sum ≥ 3 ou D-escore ≥ 0,5 ou Probabilidade de peptídeo sinal por HMM ≥ 0,5. **Painéis da direita:** Distribuição do número de espécies nos grupos de ortólogos.

Outra característica para a qual os bancos apresentaram distribuições diferenciadas foi a contribuição de cada um dos três escores usados na predição de peptídeos sinal. Dentro dos grupos Mistos, a predição positiva de proteínas por cada escore individual (NN-Sum, D-escore e Probabilidade sinal por HMM) e por todas as combinações possíveis apresentou dois padrões mais comuns: **(1)** Um equilíbrio entre os três escores, com todos sendo positivos na maioria dos casos e a combinação '*NN-Sum + D-escore*' sendo a segunda mais observada ou **(2)** uma concentração de predições positivas somente por Probabilidade sinal por HMM (**Figura 22**). Os dois padrões refletem as duas metodologias empregadas pelo SignalP 3.0, as Redes Neurais e o Modelo Oculto de Markov, respectivamente. O primeiro padrão foi observado para **Plasmodium** e **Cryptosporidium** enquanto os demais bancos apresentaram distribuições condizentes com o segundo padrão.

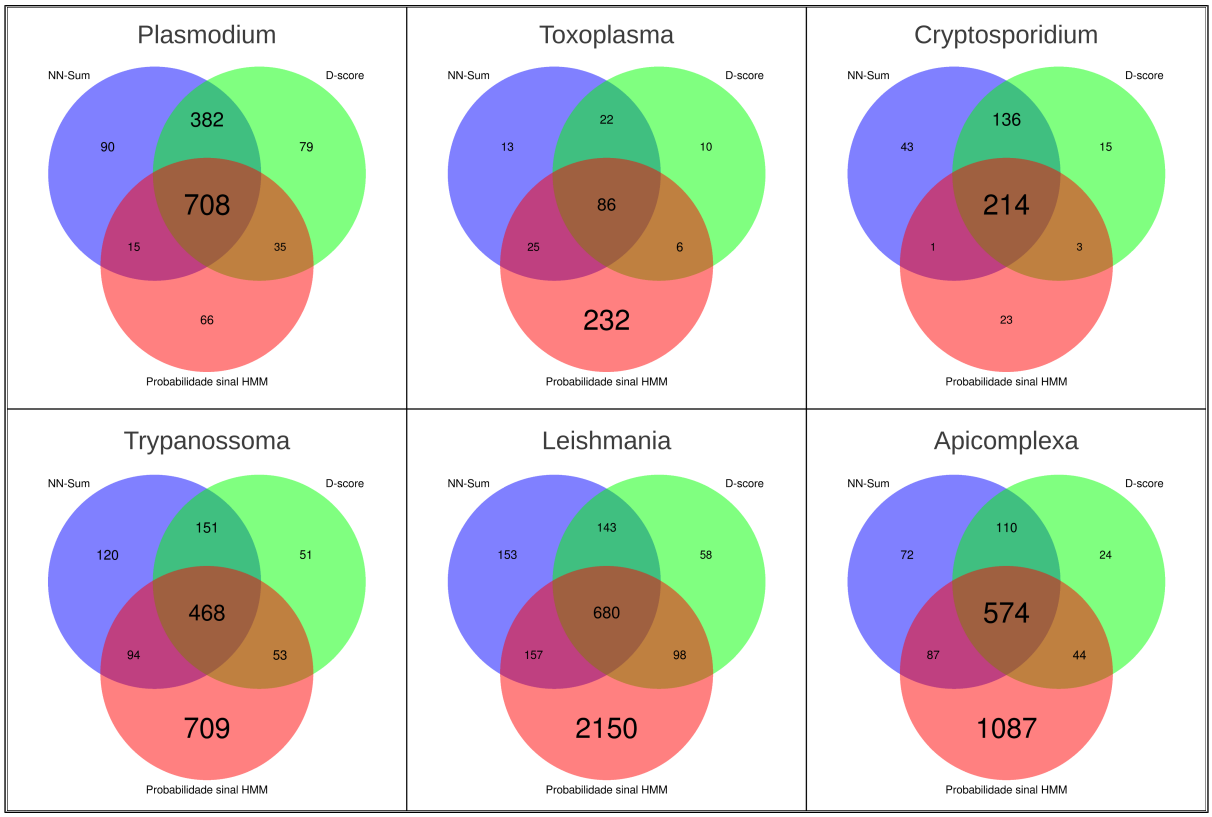


Figura 22: Diagramas de Venn representando a contribuição de cada escore individual para a predição positiva de peptídeo sinal em cada banco de dados. Foi feito o levantamento de quais os escores contribuíam para a predição positiva de peptídeos sinal. Os escores estão representados nas cores azul (NN-Sum), verde (D-escore) e vermelha (Probabilidade sinal por HMM) e as áreas de interseção representam as proteínas positivas por dois ou três escores simultaneamente. Dentro de cada seção do diagrama números totais de proteínas positivas.

5.14. Grupos Mistos contendo proteínas mal anotadas são mais variáveis

A inspeção manual dos grupos Mistos do banco de dados **Plasmodium** resultou na identificação de 111 grupos sem erros de anotação e 352 grupos com ao menos uma proteína putativamente mal anotada (foram considerados também os grupos Parcialmente

reanotados nos quais ao menos uma proteína foi revisada). Para cada um destes 463 grupos foram calculadas todas as 11 métricas do painel de variabilidade (descritas detalhadamente nos Materiais e Métodos), e para cada métrica foi feita a comparação entre os dois conjuntos de dados: grupos Sem erros (111) x grupos Com ao menos uma proteína mal anotada (352). Esta comparação revelou que grupos contendo erros de anotação são intrinsecamente mais variáveis (**Figura 23**), sendo que a diferença observada entre as medianas foi estatisticamente significativa para todas as métricas.

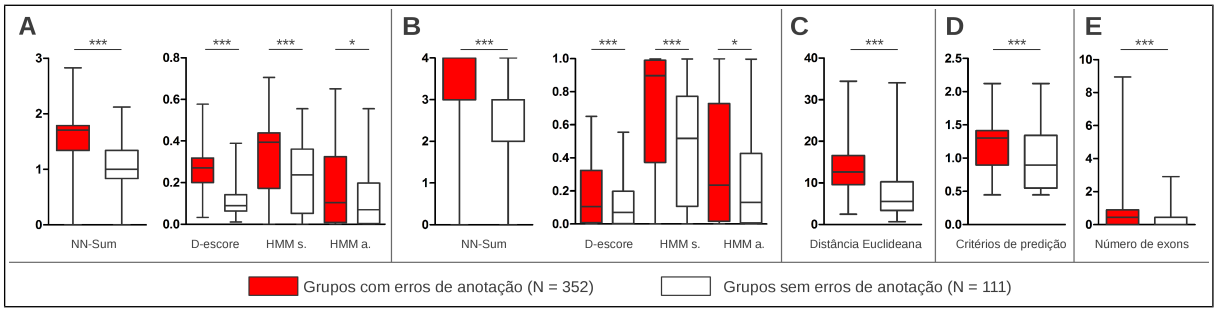


Figura 23: Comparação entre as distribuições das 11 métricas entre os grupos com erros de anotação e grupos sem erros de anotação. Dot-plots indicando a distribuição de métricas calculadas a partir **(A)** dos desvios padrão de NN-Sum, D-escore, Probabilidade sinal por HMM e Probabilidade de âncora por HMM, **(B)** das amplitudes de NN-Sum, D-escore, Probabilidade sinal por HMM e Probabilidade de âncora por HMM, **(C)** do desvio padrão da Distância Euclideana, **(D)** do desvio padrão do número de critérios de predição positivos e **(E)** do desvio padrão do número de exons. As medianas foram comparadas com o teste de Mann-Whitney (* $p < 0,05$; *** $p < 0,0001$).

Quando as proteínas mal anotadas foram desconsideradas e as 11 métricas foram recalculadas utilizando-se somente as proteínas corretamente anotadas em cada grupo, a variabilidade dos grupos foi significativamente reduzida (**Figura 24**).

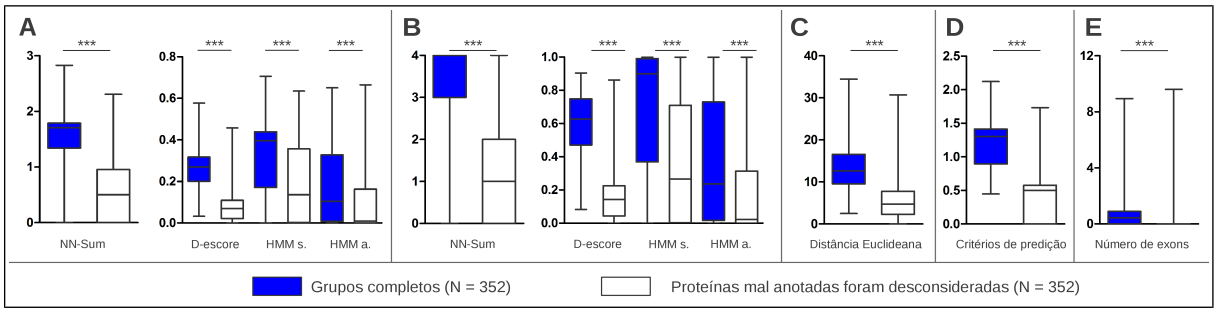


Figura 24: Comparação entre as distribuições das 11 métricas para os grupos com erros de anotação completos e desconsiderando as proteínas mal anotadas. Dot-plots indicando a distribuição de métricas calculadas a partir **(A)** dos desvios padrão de NN-Sum, D-escore, Probabilidade sinal por HMM e Probabilidade de âncora por HMM, **(B)** das amplitudes de NN-Sum, D-escore, Probabilidade sinal por HMM e Probabilidade de âncora por HMM, **(C)** do desvio padrão da Distância Euclideana, **(D)** do desvio padrão do número de critérios de predição positivos e **(E)** do desvio padrão do número de exons. As medianas foram comparadas com o teste pareado de Wilcoxon (*** $p < 0,0001$).

O mesmo fenômeno foi observado quando as proteínas mal anotadas foram revisadas e as métricas foram recalculadas considerando-se os novos resultados da predição de peptídeos sinal (**Figura 25**). A redução da variabilidade foi estatisticamente significativa para todas as métricas.

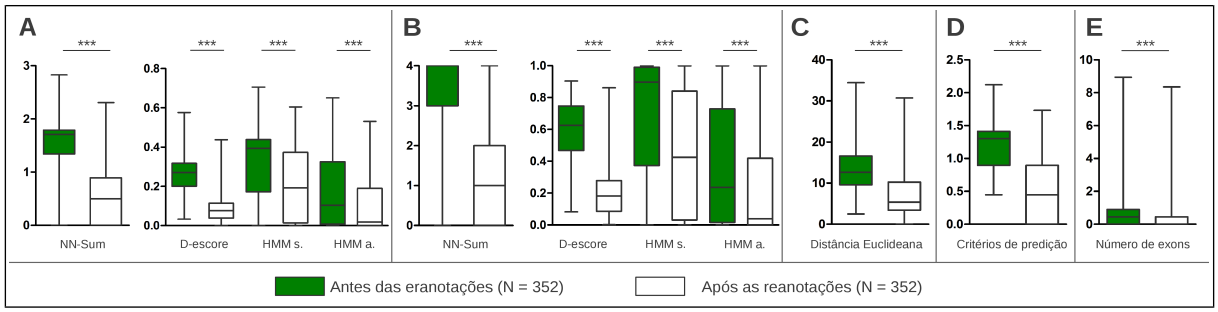


Figura 25: Comparação entre as distribuições das 11 métricas para os grupos com erros de anotação antes e depois das reanotações. Dot-plots indicando a distribuição de métricas calculadas a partir **(A)** dos desvios padrão de NN-Sum, D-escore, Probabilidade sinal por HMM e Probabilidade de âncora por HMM, **(B)** das amplitudes de NN-Sum, D-escore, Probabilidade sinal por HMM e Probabilidade de âncora por HMM, **(C)** do desvio padrão da Distância Euclideana, **(D)** do desvio padrão do número de critérios de predição positivos e **(E)** do desvio padrão do número de exons. As medianas foram comparadas com o teste pareado de Wilcoxon (*** $p < 0{,}0001$).

Estes resultados demonstram claramente que a presença de proteínas com erros de anotação N-terminal é o fator responsável pelo aumento da variabilidade em grupos Mistos ortólogos. Ainda, os resultados embasam o uso das métricas para a tarefa de distinguir entre os dois conjuntos de grupos: Com ou Sem erros de anotação. A partir destes dados, cogitou-se a automatização da etapa de identificação de grupos contendo proteínas mal anotadas, que demanda um longo tempo dedicado à inspeção de alinhamentos múltiplos e apresenta uma carga de subjetividade muito elevada.

5.15. Seleção de métricas para compor o classificador

A estratégia escolhida para a automatização da separação de grupos Mistos foi a criação de um classificador baseado em uma Máquina de Suporte de Vetores (SVM). Primeiramente, foi realizada a seleção de atributos para a composição do classificador. A ferramenta '*Feature selection tool*' foi usada para determinar o poder discriminativo de cada métrica e avaliar qual a combinação ideal de métricas para o problema de classificação dos grupos. Entre as 11 métricas, aquelas derivadas do D-escore (desvio padrão e amplitude) apresentaram os maiores valores de F-escore, seguidas pelas métricas do NN-Sum e pelo desvio padrão das Distâncias Euclidianas (calculadas a partir do resultado completo das predições de petídeo sinal) (**Tabela 10**). As demais 6 métricas apresentaram valores muito baixos de F-escore, inferiores a 0,1. (**Tabela 10**). Entretanto, a análise combinatória, após validação cruzada, indicou que a classificação mais precisa foi alcançada com a combinação de 9 métricas, sendo excluídas as 2 métricas derivadas do parâmetro de predição Probabilidade de âncora por HMM (desvio padrão e amplitude), que haviam apresentado os menores valores de F-escore.

Tabela 10: Seleção de atributos (métricas) para composição do classificador

	Métrica	F-escore	Selecionado?
1	Desvio padrão de Distâncias Euclidianas	0,216169	**Sim**
2	Desvio padrão do Número de exons	0,033971	**Sim**
3	Desvio padrão do Número de critérios	0,061116	**Sim**
4	Desvio padrão de NN-Sum	0,262426	**Sim**
5	Amplitude de NN-Sum	0,315303	**Sim**
6	Desvio padrão de D-escore	0,605367	**Sim**
7	Amplitude de D-escore	0,699938	**Sim**
8	Desvio padrão da probabilidade sinal por HMM	0,055839	**Sim**
9	Amplitude da probabilidade sinal por HMM	0,063895	**Sim**
10	Desvio padrão da probabilidade de âncora por HMM	0,015598	**Não**
11	Amplitude da probabilidade de âncora por HMM	0,017064	**Não**

5.16. Treinamento do SVM

O SVM foi treinado com os dados das inspeções de grupos Mistos do banco de dados **Plasmodium**. O conjunto de treinamento foi formado pelos 463 grupos divididos em 111 exemplos Negativos (Sem erros de anotação) e 352 exemplos Positivos (com pelo menos uma proteína mal anotada). As 9 métricas selecionadas previamente foram usadas no treinamento e os valores ótimos dos parâmetros C e γ foram definidos em $2^{16,8}$ e $2^{-17,4}$, respectivamente (**Figura 26A**). A exatidão do SVM treinado foi de 89,6%, calculada por validação cruzada em 5 vezes. Os 463 grupos Mistos usados na construção do SVM representam todos os grupos para os quais o resultado da inspeção manual foi categórico, portanto, não há entre os grupos Mistos um conjunto de dados para teste, justificando a escolha da validação cruzada. Os dados da validação cruzada foram ainda empregados para a construção de uma curva ROC apresentando uma Área Sob a Curva (AUC) de 0,9284 (**Figura 26B**). A Área Sob a Curva é uma medida de performance do classificador e demonstra que o SVM criado foi capaz de classificar satisfatoriamente os dados do conjunto de treinamento.

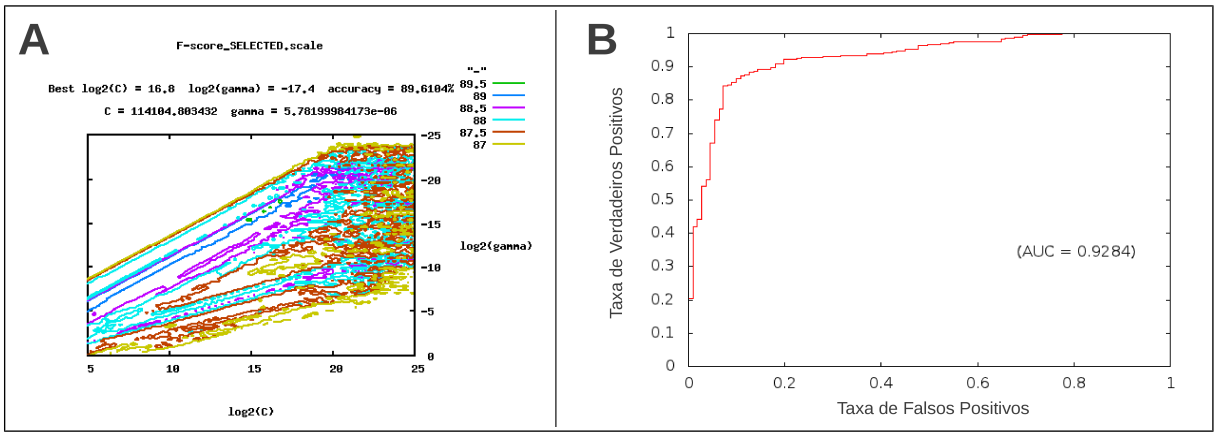


Figura 26: Treinamento do SVM. (A) Definição dos parâmetros C e γ. (B) Curva ROC calculada com os resultados de validação cruzada em 5 vezes e respectiva Área Sob a Curva (AUC).

O SVM foi usado para classificar todos os grupos Mistos de **Plasmodium**, inclusive os que não foram utilizados em sua construção (grupos Inconclusivos e parte dos grupos Parcialmente reanotados). Quando somente os 463 grupos usados para o treinamento foram considerados, a exatidão calculada para o classificador foi de 88,5%. Entretanto, a grande maioria dos 58 grupos Parcialmente reanotados que não fazem parte do conjunto de treino foi classificada entre os grupos **Com erros de anotação**. Esta classificação está amplamente de acordo com os resultados da inspeção, pois todos estes grupos (Parcialmente reanotados) possuem proteínas que foram identificadas como sendo mal anotadas mas não puderam ser revisadas. Portanto, quando todos os grupos Parcialmente reanotados foram considerados para o cálculo, a exatidão do SVM passou a ser de 89,1%, com uma taxa de recuperação (*recall*) de grupos **Com erros de anotação** de 93,2% e uma precisão de 92,9%. Curiosamente, todos os grupos Inconclusivos foram classificados entre os grupos com erros de anotação (**Tabela 11**).

Tabela 11: Comparação entre a classificação com o SVM e a inspeção visual dos grupos Mistos do banco de dados Plasmodium

Classificação com o SVM	**Inspeção visual**				**TOTAL**
	Reanotados	Parciamente reanotados	Sem erros de anotação	Inconclusivos	
Grupos **COM** erros de anotação	307	78	29	17	431
Grupos **SEM** erros de anotação	24	4	82	0	110
TOTAL	331[a]	82[b]	111	17	**541**

[a] Os grupos reanotados contribuem com 328 grupos para o treinamento do classificador

[b] Os grupos Parcialmente reanotados contribuem com 24 grupos para o treinamento do classificador

Os resultados demonstram a possibilidade de realizar satisfatoriamente a separação de grupos Mistos de forma automática, sem que haja a necessidade de uma etapa demasiadamente laboriosa de inspeção visual de cada grupo.

5.17. Aplicação do classificador aos demais bancos de dados

Após a constatação de que a separação automática de grupos é viável, o passo seguinte foi demonstrar que esta estratégia de automatização era flexível e poderia ser aplicada a outras situações, envolvendo novas espécies em diferentes contextos evolutivos. Os demais bancos de dados foram criados com este objetivo. A totalidade de grupos Mistos de cada banco de dados foi submetida à classificação automática dos grupos utilizando o classificador criado com os dados das inspeções de grupos Mistos de **Plasmodium** (SVM-Plasmodium). Entre os bancos de dados, as proporções de grupos classificados em cada categoria (Com ou Sem erros de anotação) variaram muito, desde 4:1 (**Com : Sem erros de**

anotação) em **Cryptosporidium** até 1:1 em **Leishmania**, com os demais bancos apresentando valores intermediários mas sempre com um maior número de grupos **Com erros de anotação** (**Tabela 12**). Em termos absolutos a diferença entre os bancos de dados é também considerável, enquanto **Toxoplasma** apresentou somente 159 grupos Mistos classificados **Com erros de anotação**, **Apicomplexa** apresentou 1077 (**Tabela 12**), refletindo as diferenças estruturais entre os bancos de dados.

Tabela 12: Classificação dos grupos Mistos de diferentes bancos de dados utilizando o SVM

Banco de dados	Classificação de grupos Mistos com o SVM-plasmodium				
	COM erros de anotação	%	SEM erros de anotação	%	TOTAL
Toxoplasma	159	*56,6%*	122	*43,4%*	281
Cryptosporidium	250	*79,6%*	64	*20,4%*	314
Trypanosoma	670	*59,1%*	463	*40,9%*	1133
Leishmania	914	*50,9%*	883	*49,1%*	1797
Apicomplexa	1077	*71,1%*	437	*28,9%*	1514

A performance do classificador foi avaliada através de conjuntos de testes selecionados aleatoriamente a partir dos grupos Mistos de cada banco de dados. A representatividade dos conjuntos de teste foi bastante variável entre os bancos, com a seleção de somente 6,6% dos grupos Mistos de **Apicomplexa** até a seleção de 25,5% em **Cryptosporidium** (**Tabela 13**). Os grupos selecionados aleatoriamente foram submetidos ao mesmo processo de inspeção e identificação de proteínas mal anotadas realizado para o banco de dados **Plasmodium**. Grupos Mistos inspecionados foram classificados em: **Com erros de anotação**, **Sem erros de anotação** ou **Inconclusivos**. As proporções de grupos Inconclusivos foram muito elevadas (**Tabela 13**), um fato intrigante, principalmente para **Toxoplasma**, que é um banco formado por cepas de uma mesma espécie. Mesmo a menor proporção de Inconclusivos registrada, de 8% em **Trypanosoma**, é muito superior aos ~3% de **Plasmodium** (**Figura 7C**).

Tabela 13: Inspeção manual de subconjuntos de grupos Mistos de diferentes bancos de dados

Banco de dados	Grupos Mistos			Inspeção manual de grupos Mistos					
	Total	Selecionados		COM erros de anotação		SEM erros de anotação		Inconclusivos	
Toxoplasma	281	50	*17,8%*	26	*42%*	10	*20%*	14	*28%*
Cryptosporidium	314	80	*25,5%*	26	*32,5%*	17	*21,3%*	37	*46,2%*
Trypanosoma	1133	100	*8,8%*	50	*50%*	42	*42%*	8	*8%*
Leishmania	1797	150	*8,3%*	88	*58,6%*	34	*22,7%*	28	*18,6%*
Apicomplexa	1514	100	*6,6%*	34	*34%*	24	*24%*	42	*42%*

Os resultados da inspeção manual dos conjuntos de teste foram comparados à classificação dos mesmos grupos pelo SVM-Plasmodium (**Tabela 14**). As Curvas ROC que

foram construídas a partir dos resultados da comparação para cada conjunto de testes e as respectivas Áreas Sob as Curvas (AUCs) demonstram que o classificador apresentou performances satisfatórias para todos os bancos de dados (**Figura 27**).

Tabela 14: Comparação entre a classificação com o SVM e a inspeção visual dos grupos Mistos de diferentes bancos de dados

Banco de dados	Classificação com o SVM	Inspeção visual		TOTAL
		COM erros de anotação	**SEM** erros de anotação	
Toxoplasma	**COM** erros de anotação	19	1	20
	SEM erros de anotação	7	9	16
	TOTAL	26	10	36
Cryptosporidium	**COM** erros de anotação	23	6	29
	SEM erros de anotação	3	11	14
	TOTAL	26	17	43
Trypanosoma	**COM** erros de anotação	40	11	51
	SEM erros de anotação	10	31	41
	TOTAL	50	42	92
Leishmania	**COM** erros de anotação	61	4	65
	SEM erros de anotação	27	30	57
	TOTAL	88	34	122
Apicomplexa	**COM** erros de anotação	30	6	36
	SEM erros de anotação	4	18	22
	TOTAL	34	24	58

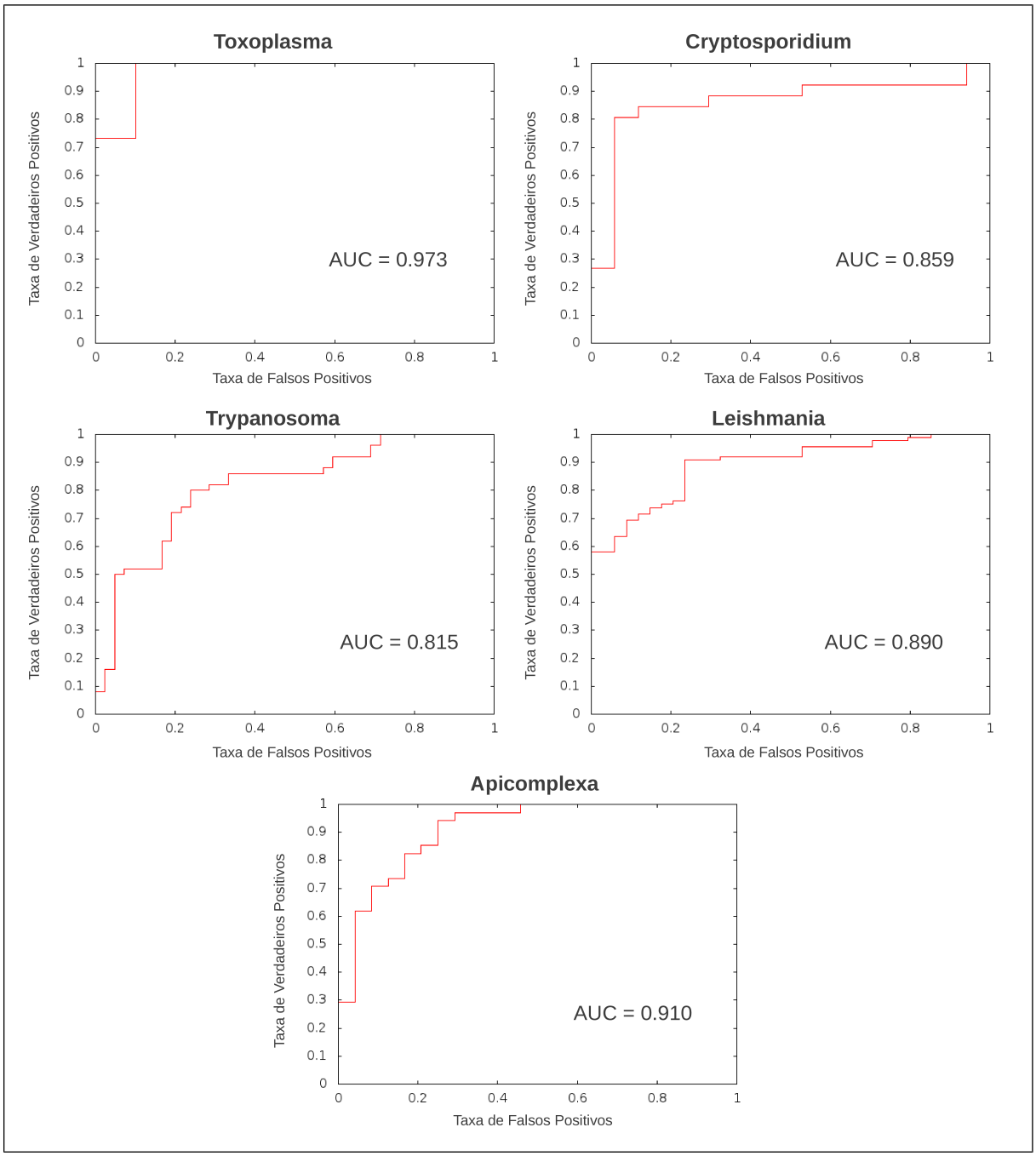


Figura 27: Performance do classificador SVM-Plasmodium nos demais bancos de dados. Os dados da classificação automática pelo SVM-Plasmodium foram comparados à inspeção manual de subconjutos de grupos Mistos de cada banco de dados e os resultados foram usados na construção das curvas ROC e no cálculo das Áreas Sob as Curvas (AUC).

Além das Áreas Sob as Curvas, outras medidas de performance foram calculadas a partir dos resultados da comparação entre a inspeção visual e a classificação automática. Para cada cenário de predição, calculou-se a exatidão, a taxa de recuperação (*recall*) de grupos Com erros de anotação, a precisão e o coeficiente de correlação de Matthews (**Tabela 15**). Em todas as situações o classificador apresentou um comportamento constante, com pouca variação observada entre as medidas avaliadas. O banco de dados **Apicomplexa** pode ser destacado por apresentar avaliaçãoes de performance tão positivas, incluindo o maior valor do coeficiente de Matthews observado, principalmente por se tratar

do único grupo composto por espécies de gêneros distintos.

Tabela 15: Avaliação da performance do classificador SVM-Plasmodium para os diferentes bancos de dados

Banco de dados	AUC*	Exatidão	Recuperação**	Precisão	CCM***
Toxoplasma	0,973	0,778	0,731	0,95	0,569
Cryptosporidium	0,859	0,791	0,885	0,793	0,555
Trypanosoma	0,815	0,772	0,8	0,784	0,539
Leishmania	0,89	0,754	0,693	0,938	0,517
Apicomplexa	0,91	0,828	0,882	0,833	0,642

* Área Sob a Curva (*Area Under the Curve*)

** Taxa de recuperação relativa a grupos **Com erros de anotação**

*** Coeficiente de Correlação de Matthews ou coeficiente phi(φ)

6. DISCUSSÃO

PARTE I – Reanotações de proteínas de *Plasmodium*

Este trabalho propõe uma abordagem inovadora para a interpretação dos resultados de predição de peptídeos sinal ao mudar a percepção destes dados como sendo propriedades exclusivas de proteínas individuais para uma perspectiva mais abrangente e baseada em um enfoque evolutivo e comparativo, na qual a presença ou a ausência de peptídeos sinal passa a ser também uma característica descritiva de grupos de proteínas ortólogas. Ao sobrepor a informação de predição de peptídeos sinal ao agrupamento de proteínas guiado por suas relações de homologia, criou-se um sistema inédito de classificação de grupos ortólogos em Positivos, Negativos e Mistos. Dentro desta classificação, destacam-se os grupos Mistos, aqueles nos quais predições conflitantes de peptídeo sinal são observadas para proteínas ortólogas. Geralmente, espera-se que haja uma conservação funcional entre proteínas ortólogas, razão pela qual os grupos Mistos despertam certa estranheza, pois o peptídeo sinal exerce influência direta sobre o papel biológico de uma proteína. Portanto, a relevância dos grupos Mistos está nas inferências sobre a sua origem e nas consequências biológicas de sua existência.

O agrupamento de duas ou mais proteínas de uma mesma espécie em um único grupo ortólogo indica que existe uma relação de paralogia entre estas proteínas. Mesmo tratando-se de parálogos recentes, como propõe a metodologia do OrthoMCL (LI; STOECKERT; ROOS, 2003), estes grupos foram excluídos das análises de reanotação de proteínas. Esta exclusão justifica-se pela diferença de conservação de funções biológicas observada entre ortólogos e parálogos. Segundo a “conjectura sobre ortologia”, as funções entre proteínas parálogas seriam, geralmente, menos conservadas que entre proteínas ortólogas, pois a evolução de funções divergentes se daria mais rapidamente entre as parálogas (KOONIN, 2005; THEISSEN, 2002). Acredita-se que, após a duplicação que dá origem às proteínas parálogas, uma das cópias, por apresentar a princípio uma redundância funcional, estaria livre para acumular mutações e desenvolver uma nova função biológica (neofuncionalização), ou então ambas as cópias passam por uma evolução acelerada e assumem funcões distintas presentes no gene ancestral multifuncional (subfuncionalização). Portanto, a conservação da presença ou ausência de peptídeos sinal entre ortólogos não é necessariamente esperada entre parálogos, como foi demonstrado para a família de proteínas VIR de *P. vivax* (BERNABEU et al., 2012).

A “conjectura sobre ortologia” é o paradigma por trás do uso difundido da ortologia na biologia comparativa, entretanto, esta conjectura sempre foi uma proposição de caráter fortemente teórico que apenas recentemente foi testada experimentalmente (DESSIMOZ et al., 2012). Alguns estudos contestam a sua validade, especialmente por causa da ligação direta que é feita entre função biológica e similaridade de sequências. Um estudo em

particular demonstrou que proteínas parálogas foram melhores para a predição de função que proteínas ortólogas (NEHRT et al., 2011). Por outro lado, os princípios da conjectura foram reafirmados por outros estudos que mostraram uma conservação significativamente maior de estrutura proteica e de perfis de expressão e, especialmente, da localização subcelular entre ortólogos do que entre parálogos (ALTENHOFF et al., 2012; DESSIMOZ et al., 2012). Estes estudos recentes sinalizam claramente que o debate ainda está aberto. A conservação da presença de peptídeos sinal não havia sido tratada até o momento.

Três hipóteses foram sugeridas para explicar a presença de grupos Mistos: **(1)** A presença de uma ou mais proteínas com erros de anotação na extremidade N-terminal, que interferem na predição, pois são onde geralmente se localizam os peptídeos sinal; **(2)** Falhas das metodologias computacionais de predição ou agrupamento de proteínas, que resultam, respectivamente, em predições imprecisas ou em grupos formados por proteínas que não apresentam relações de homologia; e **(3)** Diversidade biológica resultante da evolução divergente de proteínas ortólogas, criando grupos Mistos reais.

As duas primeiras hipóteses implicam que uma parcela dos grupos Mistos seja artificial, fruto de erros e falhas metodológicas. Neste contexto a estratégia de classificação ganha uma aplicabilidade imediata, que é contribuir para a detecção e correção destas deficiências. A terceira hipótese trata dos casos nos quais grupos ortólogos são verdadeiramente Mistos e a divergência poderia ser justificada por fenômenos biológicos. Nestes casos, a classificação de grupos, aliada a outras análises, atua como um filtro, permitindo a identificação de casos potencialmente interessantes para estudos evolutivos.

A grande maioria dos grupos Mistos apresentou pelo menos uma proteína aparentemente mal anotada, denunciando um grau de imprecisão preocupante do atual estado de anotação das proteínas do gênero *Plasmodium*, principalmente para as espécies *P. vivax* e *P. yoelii*, por razões que serão detalhadas posteriormente. Erros de anotação também foram detectados em grupos Negativos e Positivos, porém em taxas muito inferiores, o que sinaliza que a estratégia de combinação de peptídeos sinal e ortologia para a classificação de grupos viabiliza a concentração de erros de anotação, uma característica desejável para um mecanismo de controle da qualidade da anotação em escala genômica.

Ao serem reanotadas, as predições de peptídeo sinal da maioria das proteínas foram alteradas. Como consequência, vários grupos Mistos acabaram sendo reclassificados como Positivos ou Negativos, demonstrando que, como hipotetizado, a presença de proteínas mal anotadas cria divergências de peptídeo sinal artificiais entre proteínas ortólogas. Após as reanotações a proporção de grupos Mistos, que era próxima a 13%, foi praticamente reduzida à metade e esta redução é, ainda, uma estimativa conservadora, pois tanto os 17 grupos Inconclusivos, quanto os 82 grupos Parcialmente reanotados podem potencialmente ser reclassificados caso algumas de suas proteínas sejam reanotadas futuramente, o que reduziria ainda mais o número de grupos Mistos.

A reanotação de quase 500 proteínas pode causar um impacto significativo em estudos futuros, principalmente tratando-se de pesquisas nas quais a informação sobre peptídeos sinal seja preponderante. Portanto, a validação de modelos gênicos e as análises baseadas em evidências de localização celular ou anotações funcionais tiveram como objetivo oferecer um nível de confiabilidade aos resultados das reanotações.

Apesar da validação por RT-PCR não permitir a confirmação das metioninas iniciais, ficou claramente demonstrado, para todas as proteínas validadas, que os RNAs mensageiros que são expressos pelos parasitos dão suporte aos novos modelos gênicos, mas não aos modelos originais. Embora o número de genes validados tenha sido restrito, os resultados foram contundentes como prova de conceito. Durante a escolha de alvos procurou-se selecionar tanto genes que haviam sofrido alterações de predição (negativa para positiva e positiva para negativa) quanto genes que mesmo sendo reanotados mantiveram a predição original.

Nas análises usando dados de localização recuperados da literatura (ApiLoc), somente 8 proteínas reanotadas já haviam sido alvos de abordagens experimentais, sugerindo que as reanotações se concentraram em um conjunto de proteínas pouco ou nada estudadas, o que pode ser interessante do ponto de vista de estudos de prospecção de novos alvos de intervenção. Quando foram analisados os dados de localização de proteínas ortólogas às reanotadas, verificou-se uma alta concordância entre proteínas reanotadas apresentando predições positivas e a localização de suas ortólogas. Entre as proteínas com predições negativas a concordância foi menor, entretanto, as predições das próprias proteínas ortólogas não estavam de acordo com as localizações descritas para as mesmas.

Mesmo após a revisão e enriquecimento das descrições dos produtos gênicos através do *Blast Description Annotator*, a quantidade de genes sem qualquer evidência funcional permaneceu significativa, reforçando a noção de que as reanotações se concentram em um conjunto de proteínas ainda pouco explorado. Este resultado é, de certa forma esperado, pois a anotação de genes em maior evidência tende a ser revista e ajustada mais frequentemente. A análise da concordância entre descrições proteicas e predições de peptídeo para as proteínas reanotadas de *P. vivax* demonstrou que, em geral, as descrições estavam de acordo com as novas predições. Portanto, todas as evidências analisadas sugerem a pertinência das reanotações.

A presença de erros de anotação na extremidade N-terminal de proteínas foi a fonte mais comum de divergências nas predições de peptídeo sinal e de formação de grupos Mistos, o que reflete a dificuldade encontrada pelos métodos computacionais empregados, durante a etapa de anotação do genoma, na tarefa de predição gênica. A identificação do primeiro exon e, portanto, da metionina inicial é, reconhecidamente, um dos maiores desafios para algoritmos e programas de predição gênica automatizada (BERNAL et al.,

2007), particularmente para organismos eucariotos cuja estrutura gênica é mais complexa e a densidade gênica é muito inferior (DO; CHOI, 2006). Como o peptídeo sinal geralmente se localiza na extremidade N-terminal, e é comumente encontrado no primeiro exon, o fato da predição gênica ser intrinsicamente menos precisa para esta região tem um impacto significativo na gênese de erros de predição de peptídeo sinal em sequências mal anotadas.

Melhorias na predição gênica podem ter um efeito positivo na detecção de importantes atributos biológicos como, por exemplo, peptídeos sinal, promotores e alvos para micro RNAs na região 3'UTR (BERNAL et al., 2007). Os resultados do presente estudo sugerem que uma das maneiras de se aprimorar a predição e a anotação gênica seja a incorporação de informações adicionais como a predição de peptídeo sinal ao processo de descoberta de genes em um contexto comparativo.

Erros de anotação na região N-terminal de proteínas tendem a gerar predições negativas de peptídeo sinal. Este viés resulta da própria natureza do peptídeo sinal, que deve obedecer a certas restrições estruturais e composicionais (EMANUELSSON et al., 2007). A probabilidade que uma sequência aleatória de aminoácidos (≥ a 40 aminoácidos), codificada em um genoma e iniciada por uma metionina, apresente uma predição negativa de peptídeo sinal é mais alta, uma vez que sequências que respeitam as restrições estruturais são naturalmente mais raras, como demonstrado para os genomas de *P. falciparum, P.* berghei e *P. knowlesi*. Portanto, proteínas com metioninas iniciais erradas irão, geralmente, apresentar predições de peptídeo sinal negativas. Então, uma proteína que verdadeiramente não apresente peptídeo sinal, tende a manter a mesma predição mesmo havendo um erro de anotação, enquanto proteínas que deveriam apresentar peptídeo sinal têm, em geral, sua predição invertida caso haja erro de anotação. Este desequilíbrio explica porque grupos Negativos apresentam uma maior taxa de erros de anotação que os grupos Positivos, pois estes tendem a se transformar em grupos Mistos quando apresentam proteínas mal anotadas. Pela mesma razão, a maioria das reanotações sugeridas resultou na mudança de predições negativas para positivas. A mensagem fundamental é que, como regra, esta estratégia de reanotação em particular tende a aumentar o conjunto de proteínas com predições de peptídeo sinal positivas, assim como demonstrado para quatro das cinco espécies estudadas. Este enriquecimento de proteínas positivas pode ser uma vantagem na busca por alvos vacinais.

Entretanto, nem todas as proteínas identificadas como mal anotadas puderam ser revisadas. Nos 82 grupos Parcialmente reanotados, encontram-se 83 proteínas para as quais não foi possível propor novos modelos gênicos. A principal razão impedindo a reanotação destas proteínas foi a ausência de parte da região genômica que flanqueia o CDS do gene (segundo o modelo gênico original) em sua extremidade 5'. Para a reanotação, eram recuperadas as 2000 bases anteriores (5') à posição genômica que delimitava o início da proteína (códon da metionina inicial), porém, em alguns casos onde

somente fragmentos menores foram recuperados, a informação foi insuficiente para realizar a reanotação. Estes truncamentos estão diretamente relacionados ao estado da montagem dos genomas, o que explica o fato de a grande maioria das proteínas que não puderam ser reanotadas pertencer ao *P. yoelii*. Segundo dados do PlasmoDB (versão 7.1) e dos relatos originais dos sequenciamentos (CARLTON et al., 2002, 2008), *P. yoelii* apresenta o genoma mais fragmentado entre as espécies estudadas, com 5687 contigs, seguido por *P. vivax*, com 2764. Outro reflexo do atual estado de montagem do genoma de *P. yoelii* pode ser claramente observado na **Figura 9**, onde proteínas desta espécie estão ausentes de vários grupos ortólogos, entretanto, os grupos apresentam proteínas de *P. berghei* que é a espécie filogeneticamente mais próxima. O aprimoramento da montagem em *P. yoelii* aliado ao sequenciamento de regiões sem cobertura provavelmente resultaria na posterior identificação destes ortólogos "ausentes".

A estratégia para a classificação dos grupos é baseada em duas metodologias computacionais, a predição de peptídeos sinal e o agrupamento de proteínas ortólogas, e suas configurações e limitações intrínsecas influenciam os resultados obtidos. A própria atividade de predição do peptídeo sinal foi investigada como uma possível fonte para divergências observadas entre proteínas ortólogas. A predição inicial seguiu as mesmas configurações encontradas nos bancos de dados do EuPathDB (NN-Sum = 3, D-escore = 0,5 e Probabilidade sinal por HMM = 0,5), porém existia a possibilidade de que estas não fossem as melhores condições de predição e que os resultados de predição para algumas proteínas se encontrassem invertidos, o que poderia criar grupos Mistos falsos ou mascarar grupos Mistos verdadeiros. Portanto, a otimização das configurações de predição objetivou evitar ou, ao menos, reduzir a classificação errônea de grupos causada por predições imprecisas. A condição ótima foi determinada como sendo a combinação de valores que resultasse no menor número possível de grupos Mistos, ou seja, o ponto de maior concordância entre as predições de proteínas ortólogas.

A otimização foi realizada após as reanotações e somente uma vez, porém o ideal seria que reanotações e otimização fossem realizadas consecutivamente em um processo recursivo. Desta forma, a otimização permitiria expandir as reanotações, como foi demonstrado com a identificação de novas proteínas que deveriam ser reanotadas, e por sua vez, as reanotações permitiriam, a cada rodada, o ajuste dos valores de predição. Portanto, os novos valores de corte para os parâmetros de predição sugeridos aqui devem ser considerados com cautela, não devendo ser tratados como valores definitivos, pois ainda há muitos fatores que podem causar alterações futuras (novas reanotações, incorporações de novos genes, mudanças na ortologia).

De qualquer maneira, esta abordagem oferece uma oportunidade inovadora de se aprimorar a predição de peptídeos sinal sem a necessidade de implementar mudanças muito profundas nas metodologias existentes, simplesmente incorporando a informação

sobre ortologia. O desenvolvimento ou o aprimoramento de metodologias de predição também podem se beneficiar de informações provenientes de grupos Mistos, através da identificação das sequências que estão além dos atuais limites de detecção e sua posterior incorporação em conjuntos de treinamento.

A otimização das configurações de predição não corrige limitações intrínsecas da metodologia de predição como, por exemplo, um viés nos dados usados para o treinamento. Apesar do SignalP ser um programa de aplicação robusta, amplamente reconhecido e empregado, o seu conjunto de treinamento para sequências de eucariotos é dominado por sequências de mamíferos (NIELSEN; BRUNAK; VON HEIJNE, 1999), assim é plausível que peptídeos sinal de proteínas de *Plasmodium* não sejam exatamente iguais aos peptídeos sinal de mamíferos no que tange à sua estrutura e composição. Esta diferença poderia causar um desequilíbrio para predições de petídeo sinal dentro de um mesmo grupo ortólogo e responder por uma superestimação da divergência observada.

Falhas da metodologia de agrupamento de proteínas ortólogas também podem potencialmente comprometer a classificação de grupos, principalmente ao agrupar proteínas que não sejam realmente ortólogas ou a desconsiderar proteínas durante o agrupamento. Porém, durante o processo de inspeção visual dos alinhamentos não foram constatados casos onde houvesse aparente erro de montagem de grupos ortólogos (com exceção do erro sistemático responsável pela exclusão de *P. chabaudi* como discutido anteriormente), minimizando, ao menos, a possibilidade de existirem grupos formados por proteínas não ortólogas. A possibilidade de uma avaliação independente do agrupamento de proteínas (com a execução local do OrthoMCL) foi preterida em relação ao agrupamento pré-calculado executado pelo OrthoMCL-DB pois estas são as informações disponíveis para serem recuperadas a partir do PlasmoDB, refletindo assim os dados que estão à disposição da comunidade de pesquisa. Pela mesma razão, as configurações de predição de peptídeo sinal que foram usadas são as encontradas na versão do SignalP 3.0 que é executado a partir do PlasmoDB e não as configurações padrão da sua versão de distribuição. Além disso, ao usar o agrupamento pré-calculado a estratégia demandou uma menor capacidade computacional de processamento, principalmente ao evitar a etapa de BLAST do OrthoMCL. A avaliação independente do agrupamento poderia ter um impacto sobre as reanotações, uma vez que grupos ortólogos poderiam ter sido adicionados ou retirados, porém, os resultados principais e as conclusões finais não teriam sido afetados. Além disso, considerando-se a distância evolutiva entre as espécies e o alto grau de conservação observado entre as proteínas ortólogas, provavelmente o agrupamento resultante de uma avaliação independente seria muito semelhante ou idêntico ao agrupamento pré-calculado encontrado no OrthoMCL-DB.

Seria interessante adaptar esta estratégia para o uso de outras metodologias tanto de predição de peptídeos sinal quanto de agrupamento de proteínas ortólogas. Os

resultados da classificação de grupos por metodologias distintas poderiam ser comparados e combinados em busca de definições ainda mais robustas de grupos Mistos.

A terceira explicação proposta para justificar a existência de grupos ortólogos Mistos é a diversidade biológica entre as espécies do gênero *Plasmodium*. A possibilidade de que existam proteínas ortólogas que divergem em relação à presença do peptídeo sinal suscita implicações interessantes, pois é provável que estas proteínas estejam envolvidas em processos biológicos que são exclusivos de poucas ou mesmo de uma única espécie. Particularmente em *Plasmodium*, estas proteínas poderiam mediar ou interferir com uma série de fenômenos singulares que diferenciam uma espécie das outras, como, por exemplo, o sequestramento capilar de formas sanguíneas de *P. falciparum*, as preferências e tropismos dos merozoítos para invadirem certos tipos celulares (ex.: reticulócitos na malária vivax), a variabilidade no tempo de maturação e na morfologia de gametócitos entre *P. falciparum* e *P. vivax* e a formação de estágios hepáticos latentes, os hipnozoítos, em *P. vivax* (FOTH et al., 2008). A identificação de casos nos quais a divergência de peptídeos sinal entre ortólogas poderia estar relacionada à diversidade interespecífica é de alta relevância para o estudo da malária, entretanto, demonstrações inequívocas de divergência biológica, envolvendo a localização subcelular de proteínas, demandam procedimentos experimentais como, por exemplo, técnicas de imunohistoquímica com anticorpos poli ou monoclonais específicos ou a marcação de proteínas com epítopos fluorescentes através de recombinação, que estão além do escopo deste trabalho. Mesmo assim, os resultados deste trabalho podem ajudar a acelerar a identificação de grupos com maior probabilidade de estarem relacionados às divergências biológicas reais ao direcionar a atenção a um subconjunto de 141 grupos ortólogos que mantiveram a classificação de Mistos, apesar de quaisquer esforços de reanotação ou otimização da predição.

Grupos Mistos reais seriam resultado da evolução divergente entre proteínas ortólogas e, portanto, tenderiam a refletir a história evolutiva das próprias espécies. Segundo a literatura sobre a filogenia do gênero (MARTINSEN; PERKINS; SCHALL, 2008; PERKINS; SCHALL, 2002), dentre as 5 espécies estudadas, o *P. falciparum* seria a espécie mais distante, com as demais espécies ainda apresentando um ancestral comum mesmo após o evento de especiação que originou o ramo em que *P. falciparum* se encontra. As outras 4 espécies podem ser ainda divididas em dois clados, um contendo as espécies que parasitam roedores (*P. berghei* e *P. yoelii*) e outro onde se encontram as espécies que parasitam primatas (*P. vivax* e *P. knowlesi*). Interessantemente, padrões de presença/ausência de peptídeos sinal que concordam com a filogenia de *Plasmodium* são encontrados com maior frequência entre os 141 grupos consistentemente classificados como Mistos, sugerindo uma explicação evolutiva para a divergência entre ortólogas.

A existência de grupos Mistos reais é, na verdade, muito fascinante, pois estes grupos representam focos de diversidade biológica, e suas as proteínas em particular merecem estudos mais aprofundados que confirmem ou rejeitem a sua ligação aos fenômenos biológicos característicos das espécies de *Plasmodium*.

O translocamento de proteínas para o lúmen do Retículo Endoplasmático, seguido pelo transporte vesicular até o Golgi, é conhecido como 'via secretória clássica' por ser a principal rota de direcionamento intracelular de proteínas em organismos eucariotos, porém não é a única. Algumas proteínas podem ser transportadas por vias não convencionais, que operam mesmo na ausência de um sistema ER/Golgi funcional (NICKEL, 2003) e levam proteínas a serem expostas na superfície celular como proteínas integrais de membrana (SCHOTMAN; KARHINEN; RABOUILLE, 2008, 2009) ou a serem secretadas para o meio extracelular (NICKEL; SEEDORF, 2008). Existem várias proteínas de *Plasmodium* que estão incluídas nesta categoria como, por exemplo, RESA, GBP-130, Pf41-2, PfHPRT, FIRA, entre outras (LINGELBACH, 1993). Em um primeiro momento pode-se ter a impressão que as proteínas transportadas por vias não clássicas não estariam sendo consideradas nas análises baseadas na predição de peptídeo sinal, porém, vale ressaltar que proteínas transportadas por vias independentes do RE e do golgi, geralmente, apresentam predição negativa de peptídeo sinal (NICKEL; RABOUILLE, 2009; NICKEL, 2003). Portanto, se uma proteína segue uma via alternativa de transporte, o mesmo processamento biológico é antecipado para suas ortólogas, não havendo razão para divergências nas predições do grupo, que provavelmente encontra-se classificado como Negativo.

Durante as análises para estabelecer correlações entre evidências experimentais de localização subcelular e a predição de peptídeos sinal de proteínas reanotadas ou suas ortólogas, a concordância observada para as proteínas com predições negativas foi baixa. Uma possível explicação seria justamente o transporte destas proteínas ser realizado por vias alternativas, justificando os dados de localização.

As reanotações que estão sendo propostas redefinem o conjunto de proteínas que são direcionadas ao Retículo Endoplasmático de organismos do gênero *Plasmodium*, com um claro viés para o aumento do número de proteínas com peptídeo sinal, e são, portanto, altamente relevantes, pois o transporte de proteínas é essencial para o desenvolvimento destes parasitos e esta classe de proteínas, em particular, está entre os alvos moleculares prioritários para estratégias de combate à malária.

A presença de peptídeos sinal, por exemplo, é amplamente utilizada como filtro em estratégias de vacinologia reversa (GOODSWEN; KENNEDY; ELLIS, 2012; JOHN; JOHN; KHOLIA, 2012; RAPPUOLI, 2000; SEIB; ZHAO; RAPPUOLI, 2012), uma vez que alvos clássicos da resposta imune humoral são geralmente proteínas secretadas ou proteínas da superfície celular. Erros de anotação N-terminal certamente impediriam a identificação de possíveis alvos. A maioria dos principais candidatos vacinais em *Plasmodium* (AMA-1,

Pfs230, CS, PvDBP) (HILL, 2011) são proteínas que apresentam predições positivas para peptídeos sinal, demonstrando o quão importante esta caraterística pode ser na descoberta de novos alvos.

A seleção de alvos para drogas também pode ser beneficiada pela incorporação da informação sobre a presença de peptídeo sinal, especialmente quando é determinado ou esperado que a atividade metabólica alvo da intervenção ocorre dentro de certas organelas membranares ou compartimentos celulares. Neste sentido, *Plasmodium* destaca-se novamente, pois já foi demonstrado que tanto o vacúolo digestivo (EGAN, 2002; FIDOCK et al., 2004) quanto o apicoplasto (FICHERA; ROOS, 1997) são suscetíveis a compostos antimaláricos e o transporte de proteínas para ambas as organelas é dependente de peptideo sinal (TONKIN et al., 2006). Por exemplo, o direcionamento de proteínas para o apicoplasto foi um dos critérios de seleção para identificar alvos de drogas atrativos em *P. falciparum* em um estudo que empregou uma abordagem *in silico* (CROWTHER et al., 2010).

Fica evidente que a busca por novos alvos de intervenção teurapêutica é cada vez mais dependente de abordagens computacionais que manipulam grandes volumes de dados biológicos, o que torna a precisão das anotações uma prioridade, pois dados de entrada imprecisos resultam em resultados de baixa qualidade.

No banco de dados público PlasmoDB, existem informações genômicas e anotações gênicas para seis espécies do gênero *Plasmodium*, no entanto, somente cinco espécies foram analisadas na versão final deste trabalho, tendo sid após a exclusão da espécie *Plasmodium chabaudi*. Originalmente, esta espécie estava incluída nas análises, porém, durante a etapa de inspeção visual de alinhamentos, constatou-se que havia um erro sistemático nos dados depositados no PlasmoDB (versão 7.1) para esta espécie, que resultava na recuperação de sequências de proteínas não ortólogas para a montagem de grupos. Este problema foi detectado em mais de 150 grupos ortólogos. Por mais lamentável que seja a exclusão de *P. chabaudi*, não há razões para se acreditar que a sua ausência tenha qualquer impacto sobre as conclusões deste trabalho. A adição de sequências de uma nova espécie não afetaria as reanotações já propostas. A única consequência de uma eventual incorporação de proteínas de *P. chabaudi* seria o aumento do número de grupos Mistos devido à presença de novas proteínas com erros de anotação.

PARTE II – Automatização da identificação de grupos Mistos Com erros de anotação

Duas questões importantes surgiram após a conclusão do processo que levou à reanotação das proteínas de *Plasmodium*. Primeiro, a substituição da etapa de inspeção manual de alinhamentos e da análise comparativa entre modelos gênicos (e regiões

flanqueadoras) por uma metodologia computacional, capaz de realizar a identificação de grupos Mistos apresentando proteínas com erros de anotação. Segundo, a possibilidade de aplicar tanto a estratégia inicial quanto a ferramenta de classificação de grupos Mistos para auxiliar a reanotação de proteínas em outros conjuntos de espécies, demonstrando a flexibilidade, adaptabilidade e universalidade da metodologia.

A identificação manual de grupos com erros de anotação é um esforço exaustivo, meticuloso e demorado, sendo a etapa limitante de todo o processo. Além disso, é uma tarefa intrinsecamente subjetiva, pautada na experiência do observador. Portanto, a sua substituição por uma alternativa computacional representa um ganho significativo de tempo e reprodutibilidade. Para tanto, foi necessário identificar propriedades intrínsecas aos grupos que permitissem a sua separação em grupos **Com erros de anotação** e grupos **Sem erros de anotação**. Como a classificação dos grupos foi baseada nos resultados de predição de peptídeos sinal, características relativas à predição foram exploradas no intuito de se obter métricas capazes de realizar a distinção. Neste ponto, vale ressaltar, a metodologia da inspeção manual é baseada em princípios diferentes, pois a informação sobre predições individuais não foi consultada na identificação dos grupos por inspeção manual.

O raciocínio por trás do desenvolvimento do painel de métricas é baseado no conceito de que os resultados da predição de peptídeos sinal (os valores finais para cada escore do SignalP) para proteínas ortólogas devem ser próximos entre si, ou seja, devem apresentar uma baixa variância. A presença de proteínas, mesmo que somente uma, com erros de anotação que influenciem nos resultados da predição de peptídeo sinal irão causar o aumento da variância do grupo. Por este motivo as métricas desenvolvidas são todas relacionadas a medidas de variabilidade, a amplitude e o desvio padrão calculados para cada um dos três parâmetros principais (NN-Sum, D-escore e Probabilidade sinal por HMM) e para um quarto parâmetro (Probabilidade de âncora por HMM) não adotado nas predições do EuPathDB e ainda os desvios padrão para o número de critérios de positividade, para a distância Euclidiana e para o número de exons. O número de critérios de positividade varia entre 0 (predição negativa) e 3 (todos os parâmetros positivos), portanto esta métrica é especialmente forte para os grupos cujas proteínas se aproximam do valor máximo, pois o impacto do erro de anotação será maior no desvio padrão. O cálculo da distância Euclidiana foi uma tentativa de extrair o máximo de informação do resultado do SignalP, ao transformar o resultado completo da predição de cada proteína, contendo 21 atributos, em um único vetor. Esta métrica ainda pode ser aprimorada com um tratamento mais criterioso dos atributos, pois não necessariamente todos seguem a premissa da conservação entre ortólogos como, por exemplo, as predições das posições de clivagem do peptídeo sinal e, ainda, alguns dos atributos são atrelados aos valores de corte praticados pela versão de distribuição do SignalP, podendo gerar uma inconsistência nos dados. Entretanto, mesmo apresentando estas ressalvas, a distância Euclideana foi a quinta métrica com maior poder

preditivo, superando as métricas derivadas da Probabilidade sinal por HMM. A única métrica não derivada do resultado do SignalP foi o desvio padrão do número de exons, que foi adicionado após observações feitas na etapa de análise comparativa de modelos gênicos, quando foi constatado que o número de exons entre as proteínas corretas é geralmente bastante conservado. As métricas derivadas do D-escore e do NN-Sum foram as que apresentaram os maiores valores preditivos, entretanto a análise da melhor combinação de métricas resultou na escolha de 9 das 11 métricas iniciais, preservando inclusive algumas métricas com valores de F-escore muito baixos, o que significa que mesmo com valores preditivos baixos estas métricas ainda contribuíram para melhorar o poder final do classificador.

Outro argumento a favor da manutenção de métricas, mesmo com um baixo valor preditivo, é a heterogeneidade dos próprios bancos de dados, exemplificada nos diagramas de Venn da **Figura 22**. Os bancos basicamente se dividem em dois perfis de predição de peptídeo sinal, em um dos perfis existe um equilíbrio entre os três parâmetros de predição, que na verdade representa um equilíbrio entre as duas metodologias de predição do SignalP, as Redes Neurais (NN-Sum e D-escore) e o Modelo Oculto de Markov (Probabilidade sinal por HMM) com uma leve tendência a uma maior influência das Redes Neurais, no outro perfil existe uma predominância das predições pelo Modelo Oculto de Markov. O classificador foi construído com base nos dados de **Plasmodium**, um banco ligado ao primeiro perfil, o que talvez explique o maior poder preditivo do D-escore e NN-Sum. Entretanto, o classificador foi testado com sucesso contra bancos de dados ligados ao segundo perfil (**Toxoplasma**, **Trypanosoma**, **Leishmania** e **Apicomplexa**). Provavelmente, a presença das demais métricas, principalmente daquelas derivadas da Probabilidade sinal por HMM, ajudou a manter a alta performance do classificador. Em observações semelhantes, é provável que a métrica 'desvio padrão do número de critérios' perca valor preditivo nos bancos com o segundo perfil, pois a maioria das proteínas positivas apresentam somente um criério positivo (Probabilidade sinal por HMM), ou que a métrica 'desvio padrão do número de exons' também tenha seu valor preditivo reduzido na classificação de bancos de Trypanossamatídeos, que raramente apresentam proteínas com múltiplos exons.

A razão de existirem estes diferentes perfis de combinação de parâmetros preditivos entre bancos de dados é desconhecida, mas é possível que estes perfis estejam refletindo diferenças estruturais dos próprios peptídeos sinal, uma hipótese certamente interessante.

Finalmente, a exclusão das métricas derivadas da Probabilidade de âncora por HMM é um resultado compreensível, pois este parâmetro serve para definir se há ou não clivagem do peptídeo sinal e não prediz diretamente a presença deste (uma predição negativa pode ocorrer na presença ou ausência de um peptídeo sinal).

Máquinas de Suporte de Vetores (SVMs) são amplamente usadas em aplicações de

biologia computacional em razão da sua alta precisão, da sua habilidade em trabalhar com dados multidimensionais e sua flexibilidade para modelar diversos tipos de dados (BEN-HUR et al., 2008), justificando a sua escolha para a tarefa de classificação dos grupos Mistos. A função kernel de base radial (RBF) é considerada uma boa opção inicial, pois ela é capaz de lidar mesmo com os casos em que há uma relação não-linear entre atributos e classes de instâncias individuais e a sua modelagem é mais simples que a da função polinomial (HSU; CHANG; LIN, 2010). A função de base radial não é recomendada para situações onde o número de atributos seja grande, porém não é o caso do problema de classificação dos grupos Mistos, pois o classificador contou com somente nove atributos.

O objetivo da criação dos cinco bancos de dados foi testar a performance do classificador em diferentes condições, enfatizando, principalmente, **(1)** a flexibilidade na escolha de organismos, **(2)** variações no número de espécies (ou cepas) recrutadas para o agrupamento por ortologia e **(3)** a distância evolutiva entre as espécies recrutadas.

A seleção das espécies para os bancos **Toxoplasma**, **Cryptosporidium** e para os bancos de Tripanosomatídeos (**Trypanosoma** e **Leishmania**) representa um distanciamento filogenético gradual em relação ao banco de dados **Plasmodium**, que foi a fonte dos dados usados no treinamento do classificador, e teve como objetivo avaliar os resultados do classificador à medida que esta distância filogenética aumentava. Em paralelo à diversificação de espécies, foi avaliado também o efeito do número total de espécies selecionadas para o agrupamento de proteínas ortólogas. Os dados usados para treinar o classificador são dominados por grupos contendo cinco proteínas (de *Plasmodium*), o que poderia ter criado um viés com impacto negativo na classificação de grupos com um total de proteínas diferente. Portanto, a performance do classificador foi observada e avaliada nos bancos de dados em que os grupos Mistos apresentavam no máximo 3 (**Toxoplasma**, **Cryptosporidium**) ou 4 proteínas (**Trypanosoma**). O banco de dados **Apicomplexa** foi criado com o objetivo de avaliar o classificador em um cenário onde a distância filogenética entre as espécies dentro de um mesmo conjunto de teste fosse maior que entre as espécies do conjunto de treinamento. Em contraste com os demais bancos, incluindo **Plasmodium**, onde todas as espécies (ou cepas) pertencem a um mesmo gênero (ou espécie, no caso das cepas), em **Apicomplexa** cada uma das cinco espécies pertencia a um gênero distinto dentro do Filo.

Os resultados da classificação pelo SVM para os diversos conjuntos de teste indicaram que nem a escolha de um conjunto de espécies não relacionadas ao conjunto de treinamento, nem o número reduzido de proteínas em grupos Mistos e nem a formação de um conjunto de espécies filogeneticamente distantes entre si influenciaram a performance do SVM. Na prática, os resultados apontam que a utilização futura deste classificador como uma ferramenta de reanotação em escala genômica e de curadoria de bancos de dados pode ser promissora. Provavelmente, o cenário mais comum para o seu uso seria o

agrupamento de duas ou poucas (até 5) espécies evolutivamente próximas, seguindo a tendência dos estudos de genômica comparativa, porém, teoricamente esta ferramenta precisaria ser prontamente aplicável a qualquer combinação de espécies.

A flexibilidade do classificador, apesar deste ser baseado somente nos dados de inspeção de *Plasmodium*, pode ser explicada pelas próprias características da predição pelo SignalP 3.0. Primeiramente, todos os escores do programa variam em uma faixa pré-definida com limites máximo e mínimo fixos (de 0 a 1), o que resulta em distribuições dos perfis de predição de escores individuais muito semelhantes entre espécies. Além disso, o programa conta com duas metodologias de predição distintas, Redes Neurais e Modelo Oculto de Markov, que estão presentes entre os atributos do classificador, como discutido anteriormente, e provavelmente auxiliam na flexibilização ao compensarem eventuais vieses em peptídeos sinal de outras origens.

A incorporação de exemplos positivos e negativos provenientes da inspeção de outros conjuntos de dados é uma perspectiva para o aprimoramento do classificador, pois os conjuntos de treinamento devem sempre ser atualizados e expandidos para contemplar todo o espectro de variabilidade dos dados.

A tendência na interpretação dos resultados do SVM é defini-lo como um selecionador de grupos Com erros de anotação, com um claro viés para a importância e as implicações da presença de erros de anotação. Entretanto, uma interpretação alternativa pode ser considerada, com o classificador funcionando para separar os grupos Mistos entre aqueles que são explicados pela primeira hipótese (erros na extremidade N-terminal de proteínas) e aqueles que são explicados pelas hipóteses 2 (limitações das metodologias computacionais) e 3 (diversidade biológica real). O mérito desta interpretação alternativa é destacar as possíveis aplicações dos grupos Sem erros de anotação. Nos casos onde forem detectadas limitações metodológicas, os grupos serviriam como pontos de partida para eventuais correções e melhorias para as metodologias em questão. Para os casos onde existam indícios de diversidade biológica surgem possibilidades de estudos funcionais e evolutivos que fundamentem a divergência observada.

7. CONCLUSÃO

O ponto principal deste trabalho é a apresentação de uma nova abordagem metodológica simples, mas extremamente útil na retificação ou na análise de uma série de questões de alta relevância prática. A abordagem explora **(1)** as relações de ortologia entre proteínas e suas implicações na determinação de funções biológicas e **(2)** uma característica estrutural marcante e universal de proteínas, o peptídeo sinal, e as consequências de sua presença ou ausência.

O gênero *Plasmodium* foi selecionado como contexto biológico e a metodologia foi testada em um conjunto de cinco espécies. A primeira aplicação da estratégia e, provavelmente, a mais relevante foi na tarefa de reanotação de sequências de proteínas, especialmente a extremidade N-terminal onde se encontra o peptídeo sinal. Foram realizadas mais de 470 correções de proteínas como resultado direto do emprego da metodologia. Houve um número notável de inversões nas predições de peptídeo sinal que acarretaram em um aumento global do número de proteínas com predições positivas. As validações de algumas destas revisões e o corpo de evidências existentes sobre as prováveis funções de várias destas proteínas corroboram as mudanças propostas.

Analisar individualmente os resultados de cada reanotação não era um dos objetivos deste trabalho, entretanto, o impacto destas reanotações para o gênero *Plasmodium*, particularmente para *P. vivax*, será refletido principalmente nos dados que estarão disponíveis para estratégias de buscas, em escala genômica, de potenciais alvos vacinais, alvos para drogas e proteínas relacionadas à interação parasito-hospedeiro. A maioria das reanotações que estão sendo propostas já se encontram disponíveis no banco de dados PlasmoDB na forma de comentários de usuário (*user comments*) e as reanotações restantes serão incorporadas em breve.

Além de auxiliar na correção de erros de anotação, a combinação de peptídeo sinal e ortologia foi explorada para aprimorar o próprio processo de predição de peptídeo sinal. A determinação de novos valores de corte foi baseada nas premissas da conservação de função e da coesão de predições entre proteínas ortólogas. A ortologia foi empregada como uma espécie de calibrador da predição sem que fossem necessárias alterações do algoritmo preditor. Apesar de tratar-se de um resultado específico para os bancos de dados estudados e que não pretende ser definitivo, a revisão das configurações de predição objetivou promover uma mudança tanto da determinação quanto da interpretação dos resultados de predição de peptídeo sinal, ao incorporar princípios evolutivos.

Ainda que de maneira indireta e resultante de um processo de exclusão de

possibilidades, a estratégia sugeriu, por fim, a existência de grupos de proteínas ortólogas apresentando divergências reais quanto à presença de peptídeos sinal. Estes grupos provavelmente guardam ligações com fenômenos biológicos potencialmente muito interessantes, que justificariam as divergências. Portanto, a confirmação da condição destes grupos através de experimentação é uma perspectiva promissora.

Os resultados da inspeção de grupos Mistos do banco de dados **Plasmodium** foram transformados em um classificador, uma Máquina de Suporte de Vetores (SVM) capaz de identificar entre os grupos Mistos aqueles que apresentam proteínas mal anotadas. O classificador foi testado contra um painel de bancos de dados com características variadas e se mostrou robusto e flexível, mantendo uma performance consistente mesmo em cenários consideravelmente diferentes do conjunto de dados usados no treinamento.

O estudo sugere que erros de anotação são encontrados com frequência em bancos de dados. Neste contexto, os resultados da abordagem metodológica, culminando com o classificador são encorajadores, pois a estratégia cumpre um papel essencial ao auxiliar na tarefa de curadoria de dados biológicos, demonstrando um grande potencial para ser implementada como uma ferramenta de revisão e controle de qualidade de dados em escala genômica.

8. ANEXOS

Os seguintes arquivos e pastas são encontrados no CD que acompanha a Tese:

- Pasta (/GRUPOS_PlasmoDB-71) contendo arquivos multifasta com sequências proteicas de todos os grupos ortólogos (dados originais do PlasmoDB 7.1);

- Pasta (/ALINHAMENTOS_PlasmoDB-71) contendo alinhamentos de todos os grupos ortólogos (baseados nos dados originais do PlasmoDB 7.1);

- Pasta (/GRUPOS_REANOTADOS) contendo aquivos multifasta com as sequências proteicas dos grupos que apresentaram proteínas reanotadas (para proteínas reanotadas é fornecida a nova sequência);

- Pasta (/ALINHAMENTOS_REANOTADOS) contendo alinhamentos dos grupos que apresentaram proteínas reanotadas (alinhamentos com as novas sequências);

- Pasta (/NOVOS_MODELOS_GENICOS) com os arquivos embl dos novos modelos gênicos;

- Planilha (/Proteinas_reanotadas.xls) com informações sobre as proteínas reanotadas: ID do gene; espécie; grupo ortólogo; predições antes e depois; descrição do produto original; descrição do produto segundo o BDA; nova sequência proteica;

- Arquivo multifasta (/sequencias_originais.fasta) com as sequências originais das proteínas reanotadas

- Arquivo multifasta (/sequencias_novas.fasta) com as novas sequências das proteínas reanotadas

- Planilha (/Grupos_ortologos.xls) com informações sobre a inspeção, reclassificação e otimização de grupos ortólogos.

9. BIBLIOGRAFIA

ALTENHOFF, A. M. et al. Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. **PLoS computational biology**, v. 8, n. 5, p. e1002514, jan. 2012.

AMINO, R. et al. Quantitative imaging of *Plasmodium* sporozoites in the mammalian host. **Comptes rendus biologies**, v. 329, n. 11, p. 858-62, nov. 2006.

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.

AURRECOECHEA, C. et al. EuPathDB: a portal to eukaryotic pathogen databases. **Nucleic acids research**, v. 38, n. Database issue, p. D415-9, jan. 2010.

BATON, L. A.; RANFORD-CARTWRIGHT, L. C. *Plasmodium falciparum* ookinete invasion of the midgut epithelium of *Anopheles stephensi* is consistent with the Time Bomb model. **Parasitology**, v. 129, n. 6, p. 663-676, dez. 2004.

BEN-HUR, A. et al. Support vector machines and kernels for computational biology. **PLoS computational biology**, v. 4, n. 10, p. e1000173, out. 2008.

BENDTSEN, J. D. et al. Improved prediction of signal peptides: SignalP 3.0. **Journal of molecular biology**, v. 340, n. 4, p. 783-95, 16 jul. 2004.

BERNABEU, M. et al. Functional analysis of *Plasmodium vivax* VIR proteins reveals different subcellular localizations and cytoadherence to the ICAM-1 endothelial receptor. **Cellular microbiology**, v. 14, n. 3, p. 386-400, 21 mar. 2012.

BERNAL, A. et al. Global discriminative learning for higher-accuracy computational gene prediction. **PLoS computational biology**, v. 3, n. 3, p. e54, 16 mar. 2007.

BILLKER, O. et al. Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. **Nature**, v. 392, n. 6673, p. 289-92, 19 mar. 1998.

BILLKER, O. et al. Calcium and a calcium-dependent protein kinase regulate gamete formation and mosquito transmission in a malaria parasite. **Cell**, v. 117, p. 503-514, 2004.

BLOBEL, G.; DOBBERSTEIN, B. Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. **Journal of Cellular Biology**, v. 67, p. 835-851, 1975.

BLUM, T.; BRIESEMEISTER, S.; KOHLBACHER, O. MultiLoc2: integrating phylogeny and Gene Ontology terms improves subcellular protein localization prediction. **BMC bioinformatics**, v. 10, p. 274, jan. 2009.

BOSER, B.; GUYON, I.; VAPNIK, V. **An training algorithm for optimal margin classifier**Fifth Annual Workshop on Computational Learning Theory. **Anais**...ACM, 1992

CARLTON, J. M. et al. Genome sequence and comparative analysis of the model rodent malaria parasite *Plasmodium yoelii yoelii*. **Nature**, v. 419, n. 6906, p. 512-9, 3 out. 2002.

CARLTON, J. M. et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. **Nature**, v. 455, n. October, 2008.

CHANG, C.; LIN, C. LIBSVM : a library for support vector machines. **ACM Transactions on Intelligent Systems and Technology**, v. 2, n. 3, p. 1-39, 2011.

CHEN, Y.; LIN, C. Combining SVMs with Various Feature Selection Strategies. In: **Feature extraction, foundations and applications**. [S.l.] Springer, 2005. p. 315-324.

CHERIAN, B. S.; NAIR, A. S. Protein location prediction using atomic composition and global features of the amino acid sequence. **Biochemical and biophysical research communications**, v. 391, n. 4, p. 1670-4, 22 jan. 2010.

CLEVES, A. Protein transport: the nonclassical ins and outs. **Current Biology**, p. 318-320, 1997.

CONESA, A. et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. **Bioinformatics**, v. 21, n. 18, p. 3674-6, 15 set. 2005.

CORNEJO, O. E.; ESCALANTE, A. The origin and age of *Plasmodium vivax*. **Trends in parasitology**, v. 22, n. 12, p. 558-63, dez. 2006.

COX-SINGH, J. et al. *Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening. **Clinical infectious diseases**, v. 46, n. 2, p. 165-71, 15 jan. 2008.

CRABB, B. S. et al. 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, maio. 2010.

CRAIG, A.; SCHERF, A. Molecules on the surface of the *Plasmodium falciparum* infected erythrocyte and their role in malaria pathogenesis and immune evasion. **Molecular and biochemical parasitology**, v. 115, n. 2, p. 129-43, jul. 2001.

CROWTHER, G. J. et al. Identification of attractive drug targets in neglected-disease pathogens using an in silico approach. **PLoS neglected tropical diseases**, v. 4, n. 8, p. e804, jan. 2010.

DESSIMOZ, C. et al. Toward Community Standards in the Quest for Orthologs. **Bioinformatics**, p. 1-5, 12 fev. 2012.

DO, J. H.; CHOI, D.-K. Computational approaches to gene prediction. **Journal of microbiology**, v. 44, n. 2, p. 137-44, abr. 2006.

DONALDSON, I. et al. PreBIND and Textomy – mining the biomedical literature for protein-protein interactions using a support vector machine. **BMC Bioinformatics**, v. 13, p. 1-13, 2003.

EGAN, T. J. Discovering antimalarials: a new strategy. **Chemistry & biology**, v. 9, n. 8, p. 852-3, ago. 2002.

EKSI, S.; WILLIAMSON, K. C. Male-specific expression of the paralog of malaria transmission-blocking target antigen Pfs230, PfB0400w. **Molecular and biochemical parasitology**, v. 122, n. 2, p. 127-30, jul. 2002.

EMANUELSSON, O. et al. Locating proteins in the cell using TargetP, SignalP and related tools. **Nature protocols**, v. 2, n. 4, p. 953-71, jan. 2007.

ESCALANTE, A. et al. A monkey's tale: the origin of *Plasmodium vivax* as a human malaria parasite. **Proceedings of the National Academy of Sciences of the United States of America**, v. 102, n. 6, p. 1980-5, 8 fev. 2005.

ESCALANTE, A.; AYALA, F. J. Phylogeny of the malarial genus *Plasmodium*, derived from rRNA gene sequences. **Proceedings of the National Academy of Sciences of the United States of America**, v. 91, n. 24, p. 11373-7, 22 nov. 1994.

FICHERA, M. E.; ROOS, D. S. A plastid organelle as a drug target in apicomplexan parasites. **Nature**, v. 390, n. 6658, p. 407-9, 27 nov. 1997.

FIDOCK, D. A et al. Antimalarial drug discovery: efficacy models for compound screening. **Nature reviews. Drug discovery**, v. 3, n. 6, p. 509-20, jul. 2004.

FITCH, W. M. Distinguishing homologous from analogous proteins. **Systematic Biology**, v. 19, n. 2, p. 99, 1970.

FITCH, W. M. Homology. **Trends in Genetics**, v. 16, n. 5, p. 227-231, maio. 2000.

FOTH, B. J. et al. Quantitative protein expression profiling reveals extensive post-transcriptional regulation and post-translational modifications in schizont-stage malaria parasites. **Genome biology**, v. 9, n. 12, p. R177, jan. 2008.

GABALDÓN, T. et al. Joining forces in the quest for orthologs. **Genome biology**, v. 10, n. 9, p. 403, jan. 2009.

GAVEL, Y.; NILSSON, L.; VON HEIJNE, G. Mitochondrial targeting sequences why “non-amphiphilic” peptides may still be amphiphilic. **FEBS letters**, v. 235, n. 1, p. 173-177, 1988.

GOLDBERG, D. E.; COWMAN, A. F. Moving in and renovating: exporting proteins from Plasmodium into host erythrocytes. **Nature reviews. Microbiology**, v. 8, n. 9, p. 617-21, set. 2010.

GOODSWEN, S. J.; KENNEDY, P. J.; ELLIS, J. T. A guide to in silico vaccine discovery for eukaryotic pathogens. **Briefings in bioinformatics**, 24 out. 2012.

GUEIRARD, P. et al. Development of the malaria parasite in the skin of the mammalian host. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 43, p. 18640-5, 26 out. 2010.

HADLEY, T. J. Invasion of erythrocytes by malaria parasites: a cellular and molecular overview. **Annual review of microbiology**, v. 40, p. 451-77, jan. 1986.

HAY, S. I. et al. Estimating the global clinical burden of *Plasmodium falciparum* malaria in 2007. **PLoS medicine**, v. 7, n. 6, p. e1000290, jun. 2010.

HILL, A. V. S. Vaccines against malaria. **Philosophical transactions of the Royal Society of London. Series B, Biological sciences**, v. 366, n. 1579, p. 2806-14, 12 out. 2011.

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

HOWARD, R. J.; MILLER, L. H. Invasion of erythrocytes by malaria merozoites: evidence for specific receptors involved in attachment and entry. **Ciba Foundation symposium**, v. 80, p. 202-19, jan. 1981.

HSU, C.; CHANG, C.; LIN, C. A Practical Guide to Support Vector Classification. v. 1, n. 1, p. 1-16, 2010.

JIN, Y.; KEBAIER, C.; VANDERBERG, J. Direct microscopic quantification of dynamics of *Plasmodium berghei* sporozoite transmission from mosquitoes to mice. **Infection and immunity**, v. 75, n. 11, p. 5532-9, nov. 2007.

JOHN, L.; JOHN, G. J.; KHOLIA, T. A Reverse Vaccinology Approach for the Identification of Potential Vaccine Candidates from *Leishmania* spp. **Applied biochemistry and biotechnology**, 21 mar. 2012.

KAISER, K.; MATUSCHEWSKI, K. et al. Differential transcriptome profiling identifies. **Molecular Microbiology**, v. 51, p. 1221-1232, 2004.

KAISER, K.; CAMARGO, N. et al. A member of a conserved *Plasmodium* protein family with membrane-attack complex/perforin (MACPF)-like domains localizes to the micronemes of sporozoites. **Molecular and biochemical parasitology**, v. 133, n. 1, p. 15-26, jan. 2004.

KATOH, K. et al. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. **Nucleic acids research**, v. 30, n. 14, p. 3059-66, 15 jul. 2002.

KATS, L. M. et al. *Plasmodium* rhoptries: how things went pear-shaped. **Trends in parasitology**, v. 22, n. 6, p. 269-76, jun. 2006.

KATS, L. M. et al. Protein trafficking to apical organelles of malaria parasites - building an invasion machine. **Traffic (Copenhagen, Denmark)**, v. 9, n. 2, p. 176-86, fev. 2008.

KOCH, H.-G.; MOSER, M.; MÜLLER, M. Signal recognition particle-dependent protein targeting, universal to all kingdoms of life. **Reviews of physiology, biochemistry and pharmacology**, v. 146, p. 55-94, jan. 2003.

KOHLER, S. A Plastid of Probable Green Algal Origin in Apicomplexan Parasites. **Science**, v. 275, n. 5305, p. 1485-1489, 7 mar. 1997.

KOONIN, E. V. Orthologs, paralogs, and evolutionary genomics. **Annual review of genetics**, v. 39, p. 309-38, jan. 2005.

KRIEF, S.; ESCALANTE, A.; PACHECO, M. On the diversity of malaria parasites in African apes and the origin of *Plasmodium falciparum* from Bonobos. **PLoS …**, 2010.

KROTOSKI, W. A. et al. Demonstration of hypnozoites in sporozoite-transmitted *Plasmodium vivax* infection. **The American journal of tropical medicine and hygiene**, v. 31, n. 6, p. 1291-3, nov. 1982.

KUTAY, U.; AHNERT-HILGERL, G.; HARTMANN, E. Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane. **EMBO Journal**, v. 14, n. 2, p. 217-223, 1995.

LECLERC, M. C. et al. Evolutionary relationships between 15 *Plasmodium* species from new and old world primates (including humans): an 18S rDNA cladistic analysis. **Parasitology**, v. 129, n. Pt 6, p. 677-84, dez. 2004.

LI, L.; STOECKERT, C. J.; ROOS, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. **Genome research**, v. 13, n. 9, p. 2178-89, set. 2003.

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

LINGELBACH, K.; JOINER, K. A. The parasitophorous vacuole membrane surrounding *Plasmodium* and *Toxoplasma*: an unusual compartment in infected cells. **Journal of cell science**, v. 111, n. 11, p. 1467-75, jun. 1998.

LINIAL, M. How incorrect annotations evolve – the case of short ORFs. **Trends in Biotechnology**, v. 21, n. 7, p. 298-300, jul. 2003.

MAK, M.-W.; WANG, W.; KUNG, S.-Y. Fast subcellular localization by cascaded fusion of signal-based and homology-based methods. **Proteome science**, v. 9 Suppl 1, n. Suppl 1, p. S8, jan. 2011.

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

MARTINSEN, E. S.; PERKINS, S. L.; SCHALL, J. J. A three-genome phylogeny of malaria parasites (*Plasmodium* and closely related genera): evolution of life-history traits and host switches. **Molecular phylogenetics and evolution**, v. 47, n. 1, p. 261-73, abr. 2008.

MIKOLAJCZAK, S. A et al. Distinct malaria parasite sporozoites reveal transcriptional changes that cause differential tissue infection competence in the mosquito vector and mammalian host. **Molecular and cellular biology**, v. 28, n. 20, p. 6196-207, out. 2008.

MINISTÉRIO DA SAÚDE. **Guia prático de tratamento da malária no Brasil.**Brasilia, 2010.

MONGUI, A. et al. Characterization and antigenicity of the promising vaccine candidate *Plasmodium vivax* 34kDa rhoptry antigen (Pv34). **Vaccine**, v. 28, n. 2, p. 415-21, 11 dez. 2009.

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: the first steps of *Plasmodium* sporozoites in the mammalian host. **Cellular microbiology**, v. 6, n. 12, p. 1113-8, dez. 2004.

MUELLER, I. et al. Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite. **The Lancet infectious diseases**, v. 9, n. 9, p. 555-66, set. 2009.

NAIR, R.; ROST, B. Sequence conserved for subcellular localization. **Protein Science**, v. 11, p. 2836-2847, 2002.

NEHRT, N. L. et al. Testing the ortholog conjecture with comparative functional genomic data from mammals. **PLoS computational biology**, v. 7, n. 6, p. e1002073, jun. 2011.

NICKEL, W. The mystery of nonclassical protein secretion. **European Journal of Biochemistry**, v. 270, n. 10, p. 2109-2119, 22 abr. 2003.

NICKEL, W.; RABOUILLE, C. Mechanisms of regulated unconventional protein secretion. **Nature reviews. Molecular cell biology**, v. 10, n. 2, p. 148-55, fev. 2009.

NICKEL, W.; SEEDORF, M. Unconventional mechanisms of protein transport to the cell surface of eukaryotic cells. **Annual review of cell and developmental biology**, v. 24, p. 287-308, jan. 2008.

NIELSEN, H. et al. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. **Protein Engineering**, v. 10, n. 1, p. 1-6, 1997.

NIELSEN, H.; BRUNAK, S.; VON HEIJNE, G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. **Protein engineering**, v. 12, n. 1, p. 3-9, jan. 1999.

OLIVIERI, A. et al. The *Plasmodium falciparum* protein Pfg27 is dispensable for gametocyte and gamete production, but contributes to cell integrity during gametocytogenesis. **Molecular microbiology**, v. 73, n. 2, p. 180-93, jul. 2009.

OLLOMO, B. et al. A new malaria agent in African hominids. **PLoS pathogens**, v. 5, n. 5, p. e1000446, 2009.

ONG, C. W. M. et al. Case Report: Monkey Malaria in Humans: A Diagnostic Dilemma with Conflicting Laboratory Data. v. 80, n. 6, p. 927-928, 2009.

PEI, Y. et al. *Plasmodium* pyruvate dehydrogenase activity is only essential for the parasite's progression from liver infection to blood infection. **Molecular microbiology**, v. 75, n. 4, p. 957-71, fev. 2010.

PERKINS, S. L.; SCHALL, J. J. A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. **The Journal of parasitology**, v. 88, n. 5, p. 972-8, out. 2002.

PETERSON, M. E. et al. Evolutionary constraints on structural similarity in orthologs and paralogs. **Protein science : a publication of the Protein Society**, v. 18, n. 6, p. 1306-15, jun. 2009.

PRADEL, G.; FREVERT, U. Malaria sporozoites actively enter and pass through rat Kupffer cells prior to hepatocyte invasion. **Hepatology**, v. 33, n. 5, p. 1154-65, maio. 2001.

PRZYBORSKI, J. M.; LANZER, M. Protein transport and trafficking in *Plasmodium falciparum*-infected erythrocytes. **Parasitology**, v. 130, n. 4, p. 373-388, abr. 2005.

RAPPUOLI, R. Reverse vaccinology. **Current opinion in microbiology**, v. 3, n. 5, p. 445-50, out. 2000.

RICH, S. M. et al. The origin of malignant malaria. **Proceedings of the National Academy of Sciences of the United States of America**, v. 106, n. 35, p. 14902-7, set. 2009.

RUTHERFORD, K. et al. Artemis: sequence visualization and annotation. **Bioinformatics**, v. 16, n. 10, p. 944-5, out. 2000.

SCHNEIDER, G.; FECHNER, U. Advances in the prediction of protein targeting signals. **Proteomics**, v. 4, n. 6, p. 1571-80, jun. 2004.

SCHOTMAN, H.; KARHINEN, L.; RABOUILLE, C. dGRASP-mediated noncanonical integrin secretion is required for *Drosophila* epithelial remodeling. **Developmental cell**, v. 14, n. 2, p. 171-82, fev. 2008.

SCHOTMAN, H.; KARHINEN, L.; RABOUILLE, C. Integrins mediate their unconventional, mechanical-stress-induced secretion via RhoA and PINCH in Drosophila. **Journal of cell science**, v. 122, n. Pt 15, p. 2662-72, 1 ago. 2009.

SEIB, K. L.; ZHAO, X.; RAPPUOLI, R. Developing vaccines in the era of genomics: a decade of reverse vaccinology. **Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases**, v. 18 Suppl 5, p. 109-16, out. 2012.

SHAN, S.; WALTER, P. Co-translational protein targeting by the signal recognition particle. **FEBS letters**, v. 579, n. 4, p. 921-6, 7 fev. 2005.

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

SILVA, J.; EGAN, A. Genome sequences reveal divergence times of malaria parasite lineages. …, v. 138, n. 13, p. 1737-1749, 2010.

SINDEN, R. E. *Plasmodium* differentiation in the mosquito. **Parassitologia**, v. 41, n. 1-3, p. 139-48, set. 1999.

SIRI, J. G. et al. Quantitative urban classification for malaria epidemiology in sub-Saharan Africa. **Malaria journal**, v. 7, p. 34, jan. 2008.

SNOW, R. W. et al. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. **Nature**, v. 434, n. 7030, p. 214-7, 10 mar. 2005.

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 out. 2006.

TASHIRO, Y. Subcellular compartments and protein topogenesis. **Cell structure and function**, v. 8, n. 2, p. 91-107, jun. 1983.

TATEM, A. J.; SMITH, D. L. International population movements and regional *Plasmodium falciparum* malaria elimination strategies. **Proceedings of the National Academy of Sciences of the United States of America**, v. 107, n. 27, p. 12222-7, 6 jul. 2010.

THEISSEN, G. Secret life of genes. **Nature**, v. 415, n. 6873, p. 741, 14 fev. 2002.

TONKIN, C. J. et al. Protein targeting to destinations of the secretory pathway in the malaria parasite *Plasmodium falciparum*. **Current opinion in microbiology**, v. 9, n. 4, p. 381-7, 2006.

TOPOLSKA, A. E. et al. Characterization of a membrane-associated rhoptry protein of *Plasmodium falciparum*. **The Journal of biological chemistry**, v. 279, n. 6, p. 4648-56, 6 fev. 2004.

VLACHOU, D. et al. Real-time, in vivo analysis of malaria ookinete locomotion and mosquito midgut invasion. **Cellular microbiology**, v. 6, n. 7, p. 671-85, 2004.

VON HEIJNE, G. Signal Peptide. **The Journal of Membrane Biology**, v. 201, p. 195-201, 1990.

WALLER, R. et al. Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. **The EMBO journal**, v. 19, n. 8, 2000.

WATERS, A. P.; HIGGINS, D. G.; MCCUTCHAN, T. F. *Plasmodium falciparum* appears to have arisen as a result of lateral transfer between avian and human hosts. **Proceedings of the National Academy of Sciences of the United States of America**, v. 88, n. 8, p. 3140-4, 15 abr. 1991.

WHITE, N. J. *Plasmodium knowlesi*: the fifth human malaria parasite. **Clinical infectious diseases : an official publication of the Infectious Diseases Society of America**, v. 46, n. 2, p. 172-3, 15 jan. 2008.

WILSON, R. J. et al. Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. **Journal of molecular biology**, v. 261, n. 2, p. 155-72, 16 ago. 1996.

WINDSOR, A. J.; MITCHELL-OLDS, T. Comparative genomics as a tool for gene discovery. **Current opinion in biotechnology**, v. 17, n. 2, p. 161-7, abr. 2006.

WORLD HEALTH ORGANIZATION. **World Malaria Report**. [S.l.] WHO Press, 2011.

YAMAUCHI, L. M. et al. *Plasmodium* sporozoites trickle out of the injection site. **Cellular microbiology**, v. 9, n. 5, p. 1215-22, maio. 2007.

YEH, E.; DERISI, J. L. Chemical Rescue of Malaria Parasites Lacking an Apicoplast Defines Organelle Function in Blood-Stage *Plasmodium falciparum*. **PLoS Biology**, v. 9, n. 8, p. e1001138, 30 ago. 2011.

ZIELER, H.; DVORAK, J. A. Invasion in vitro of mosquito midgut cells by the malaria parasite proceeds by a conserved mechanism and results in death of the invaded midgut cells. **PNAS**, n. Track II, 2000.