Ministério da Saúde
Fundação Oswaldo Cruz
Centro de Pesquisas René Rachou
Programa de Pós-graduação em Ciências da Saúde

Identificação e caracterização computacional de proteínas do tipo IUP no proteoma predito de *Schistosoma mansoni*

por

Raul Torrieri

Belo Horizonte
Fevereiro/2010

DISSERTAÇÃO MBCM – CPqRR R. TORRIERI 2010

Ministério da Saúde
Fundação Oswaldo Cruz
Centro de Pesquisas René Rachou
Programa de Pós-graduação em Ciências da Saúde

Identificação e caracterização computacional de proteínas do tipo IUP no proteoma predito de *Schistosoma mansoni*

por

Raul Torrieri

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

Orientação: Jeronimo Conceição Ruiz

Belo Horizonte
Fevereiro/2010

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

T695i 2010

Torrieri, Raul.

Identificação e caracterização computacional de proteínas do tipo IUP no proteoma predito de S*chistosoma mansoni* / Raul Torrieri. – Belo Horizonte, 2010.

xx, 116 f.: il.; 210 x 297mm.

Bibliografia: f.: 132 - 136

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

1. Esquistossomose mansoni/genética 2. *Schistosoma mansoni*/ultraestrutura 3. Proteoma/ultraestrutura 4. Biologia computacional/métodos i. Título. ii. Ruiz, Jeronimo Conceição (Orientação).

CDD – 22. ed. – 616.963

Ministério da Saúde
Fundação Oswaldo Cruz
Centro de Pesquisas René Rachou
Programa de Pós-graduação em Ciências da Saúde

Identificação e caracterização computacional de proteínas do tipo IUP no proteoma predito de *Schistosoma mansoni*

por

Raul Torrieri

Foi avaliada pela banca examinadora composta pelos seguintes membros:

Prof. Dr. Jeronimo Conceição Ruiz (Presidente)

Prof. Dra. Cristiana Ferreira Alves de Brito

Prof. Dr. Richard John Ward

Suplente: Prof. Dr. Paulo Marcos Zech Coelho

Dissertação defendida e aprovada em: 26/02/2010.

COLABORADORES

Centro de Pesquisas René Rachou – Belo Horizonte
Dra. Rosiane Aparecida da Silva Pereira

SUPORTE FINANCEIRO

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – nº 131682/2008-3).

“A coisa mais bela que podemos experimentar é o mistério. Essa é a fonte de toda a arte e ciências verdadeiras.”

Albert Einstein

Dedico esse trabalho a minha família, que desde muito cedo me ensinou que o conhecimento, o respeito, e acima de tudo, a perseverança e a humildade são que tornam um homem digno de seus méritos.

AGRADECIMENTOS

Ao meu orientador Jeronimo, pela hospitalidade e apoio durante o processo de seleção para essa instituição, pela amizade e orientação durante os dois anos de trabalho e pela ajuda inestimável para a finalização desse projeto.

A Dra. Rosiane Aparecida da Silva Pereira, pelo apoio, paciência e dedicação na realização dos experimentos de proteômica.

A toda a equipe envolvida nos trabalhos de proteômica, pela inestimável ajuda e dedicação durante a realização dos experimentos.

Ao Laboratório de Parasitologia Celular e Molecular, pela acolhida e pela oportunidade de desenvolver o meu projeto.

A minha família. Meu pai, minha mãe e meu irmão. Meu mais sincero e profundo obrigado. Pelo apoio incondicional, acima de qualquer esforço ou razão. Mesmo a distância, me fizeram crescer, me fizeram amar, me fizeram enxergar, me fizeram mais do que eu era há dois anos. Devo e dedico esse trabalho a vocês.

A minha avó Maria, pelo apoio e torcida essenciais em todos os aspectos para que eu conseguisse concluir esse projeto.

A uma pessoa com importância especial, Patrícia. Pela amizade sincera e incondicional, pela atenção e pelo carinho. Também pelo companheirismo em momentos de alegria, de tristeza, de muito trabalho, de chuvas, de frio e de vitórias. Pelo conforto de saber que nunca estive sozinho e que juntos nós conseguimos.

A duas pessoas especiais, Antonio e Tatiana, por terem me acolhido como um amigo. Por ótimos momentos e apoio que me proporcionaram durante esses dois anos. Agradeço também ao Antonio pela oportunidade de ter me tornado um amigo da sua família. Agradeço também a eles pelos ótimos momentos e pela acolhida.

Aos amigos da bioinformática, Marco Aurélio, Nesley, Antonio e Luciana por todo o apoio e amizade.

AGRADECIMENTOS

A toda a turma da peteca (Armando, Sabrina, Bruno, Antonio, Tatiana, Patrícia, Flavio, Silvia, Marcele, Maira, Batata e Carol), que durante esses dois anos também me acolheram como um amigo, e me proporcionaram ótimos momentos de diversão e amizade.

A minha amiga Claudia, pela sincera amizade, pelo apoio e carinho de irmã, pelo exemplo de força e de caráter e pela presença constante mesmo com tamanha distância.

A todos os amigos do Laboratório de Parasitologia Celular e Molecular, por todo o apoio e amizade que me dedicaram.

À Biblioteca do CPqRR em prover acesso gratuito local e remoto à informação técnico-científica em saúde custeada com recursos públicos federais, integrante do rol de referências desta dissertação, também pela catalogação e normalização da mesma.

Ao Centro de Pesquisas René Rachou pala infra-estrutura técnica.

Ao programa de pós-graduação em Ciências da Saúde do CPqRR pela oportunidade e pela hospitalidade.

Ao CNPQ pela bolsa de estudos.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE GRÁFICOS

LISTA DE ABREVIATURAS E SÍMBOLOS

aa: aminoácidos

API: *Application Programming Interface* (Interface de Programação para Aplicações)

Da: abreviação da unidade de medida de peso molecular Dalton

DER: Diagrama de Entidades e Relacionamentos

FN: F*alse Negative* (Falso Negativo)

FP: *False Positive* (Falso Positivo)

FPR: F*alse Positive Rate* (Taxa de Falso Positivo)

GeneDB: Repositório de dados genômicos (www.genedb.org)

GO: *Gene Ontology*

HCA: *Hydrophobic Cluster Analysis* (Análise de Cluster Hidrofóbico)

ISO: *International Organization for Standardization*

IUP: *Intrinsically Unstructured Protein* (Proteínas Intrinsecamente Desestruturadas)

IUPAC: *International Union of Pure and Applied Chemistry*

Kb: kilobase

LDR: Long Disordered Region (Regiões Longas de Desordem)

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

MER: Modelo Entidade Relacionamento

MR: Modelo Relacional

MR: Modelo Relacional

NMR: *Nuclear Magnetic Resonance* (Ressonância Nuclear Magnética)

MAS: *Multiple Sequence Alignment* (Alinhamento múltiplo de Seqüências)

PDB: *Protein Data Bank*

RLD: Região Longa de Desordem

ROC: *Receiver Operating Characteristics*

S. mansoni: *Schistosoma mansoni*

SCOP: *Structural Classification Of Proteins*

SGBD: Sistema Gerenciador de Banco de Dados

SQL: *Structured Query Language*

TN: *False Negative* (Falso Negativo)

TP: *True Positive* (Verdadeiro Positivo)

TPR: T*rue Positive Rate* (Taxa de Verdadeiro Positivo)

WGS: *Whole Genome Shotgun* ()

XML: *eXtensible Markup Language*

LISTA DE DEFINIÇÕES

Arquivo multi-fasta: arquivo que apresenta duas ou mais seqüências no formato fasta. Fasta é um formato de apresentação de seqüências biológicas, no qual, para cada seqüência existe uma linha de identificação começando com o símbolo “>” e que descreve a seqüência com informações variadas, sendo seguida por outras linhas contendo a seqüência propriamente dita em um total de 60 a 80 caracteres por linha.

Bash: é um interpretador de comandos, uma espécie de tradutor entre o sistema operacional e o usuário, normalmente conhecido como Shell. Permite execução de seqüências de comandos diretamente no terminal do sistema ou escritas em arquivos de texto, conhecidos como Shell *scripts*.

CSV: Comma-Separated Values (Valores Separados por Vírgula). Formato de arquivo texto no qual os valores são separados por vírgulas. Quando importados em uma planilha eletrônica, todo valor entre vírgulas será inserido em uma coluna diferente.

Hash: é uma estrutura de dados especial, que associa chaves de pesquisa a valores. Seu objetivo é, a partir de uma chave simples, fazer uma busca rápida e obter o valor desejado.

Perl (*Practical Extraction and Reporting Language*): linguagem de programação interpretativa bastante popular que vem sendo extensivamente utilizada em diferentes áreas como programação de web e bioinformática.

Pipeline: é um programa que integra e coordena diferentes instruções a serem executadas de maneira automática a fim de atingir um objetivo final.

MySQL: é um sistema gerenciador de banco de dados ou SGBD (ver lista de abreviaturas), responsável por implementar no sistema operacional, o modelo de dados relacional projetado e permitir consultas complexas.

Script: linguagem de computador interpretada, uma série de instruções formais escritas para um interpretador.

Vetor: é um modo de armazenamento e organização de dados em um programa de computador, cuja organização e método de acesso se assemelham a uma pasta de documentos.

RESUMO

A relação entre estrutura e função protéica é um dos conceitos mais bem estabelecidos da biologia molecular. O acúmulo de evidências experimentais, cujos primeiros trabalhos datam de 1890, suportam essa hipótese com grande embasamento científico. Apesar da existência de evidências de mais de um século de estudos, somente no inicio da década de 90 começaram a surgir trabalhos mostrando de forma conclusiva a existência de proteínas funcionalmente ativas, mas incapazes de manter uma conformação estável em condições fisiológicas. Tais proteínas, hoje conhecidas como IUPs (do inglês *Intrinsically Unstructured Proteins*) estão envolvidas em importantes processos de saúde e doenças, tais como o câncer e diversos processos de interação parasito/hospedeiro. A presente dissertação tem como proposta o estabelecimento de um *pipeline* computacional visando à avaliação dos diferentes algoritmos de predição de desordem estrutural, seu desempenho e a posterior aplicação dessa ferramenta no estudo *in silico* do conteúdo de IUPs presentes no proteoma predito de *Schistosoma mansoni*. Complementarmente, foi desenhado um banco de dados MySQL capaz de albergar toda a informação de desordem estrutural juntamente com diferentes dados de caracterização das IUPs para *S. mansoni*. Foram analisados um total de 10.417 proteínas, 7.373 predições de desordem estrutural, mais de 24.600 predições de características estruturais e funcionais, desenvolvidos 21 *scripts*, e todas essas predições e scripts desenvolvidos foram integrados em um *pipeline* totalmente automático e inédito para análise de desordem estrutural. Nossas análises de sensibilidade e especificidade implementadas pela análise de gráficos ROC e pela integração de resultados utilizando bancos de dados relacionais indicam que a predição integrativa (consenso de quatro diferentes metodologias de predição) de desordem estrutural apresenta um ganho de 40% na correta identificação de regiões desordenadas se comparada às predições de cada metodologia individualmente. Aproximadamente 5,5% das regiões desordenadas identificadas tiveram suas coordenadas limítrofes ajustadas após comparação com as coordenadas de domínios conservados. Nossos resultados indicam que aproximadamente 33,6% do proteoma predito de *S. mansoni* apresenta desordem estrutural. Destas, 2% apresentam domínios transmembrana e 7% apresentam peptídeo sinal. A comparação do perfil funcional das IUPs com as proteínas globulares de *S. mansoni* demonstra uma maior proporção de IUPs envolvidas em processos de regulação celular e componentes extracelulares.

ABSTRACT

The relationship between protein structure and function is one of the more well-established concepts of molecular biology. The accumulation of experimental evidence, dating from 1890, put this hypothesis on a strong scientific base. Despite the evidence of more than half a century of studies, only in the early 90's began to surface studies showing conclusively the existence of functionally active proteins, but unable to maintain a stable conformation under physiological conditions. These proteins, today known as IUPs (Intrinsically Unstructured Proteins) are involved in important processes in health and diseases such as cancer and various processes of host and parasite interaction. This work is a proposal to establish a computational pipeline in order to evaluate different algorithms for prediction of structural disorder, their performance and the posterior application of this tool in the in silico study of the content of IUPs present in the S. mansoni predicted proteome. In addition, a MySQL database was developed to store all the information of structural disorder together with different data of IUPs characterization for *S. mansoni*. We analyzed a total of 10,417 proteins, 7,373 predictions of structural disorder, more than 24,600 predictions of structural and functional characteristics, developed 21 scripts, and all these predictions and scripts were integrated in an original and totally automatic pipeline for analysis of structural disorder. Our analysis of sensitivity and specificity implemented by the analysis of ROC graphics and by the integration of results using relational databases, indicate that the integrative prediction (consensus of four different methods of prediction) of structural disorder shows an increase of 50% in correctly identifying disordered regions compared to the predictions of each single method. Approximately 5.5% of identified disordered regions had their boundaries coordinates adjusted after comparison with conserved domain coordinates. Our results indicate that approximately 33.6% of the predicted proteome of *S. mansoni* presents structural disorder. 2% of these, have at least one transmembrane domain and 7% had signal peptide. The comparison of functional profile of IUPs with globular proteins of *S. mansoni* shows the biggest proportion of IUPs are involved in process of cellular regulation and extracellular components.

1 INTRODUÇÃO

1.1 Contexto biológico das IUPs

1.1.1 Relação estrutura função e IUPs

Até a década de 90, a explicação mais amplamente aceita sobre a determinação da função de uma proteína estava centrada no chamado paradigma estrutura/função. Esse paradigma postula que a seqüência de uma proteína determina sua estrutura tridimensional, e que essa estrutura determina sua função.

De acordo com esse conceito, uma proteína só pode desempenhar sua função biológica após assumir uma conformação tridimensional estável e única. Quando uma proteína assume essa conformação específica, diz-se que a proteína assumiu seu estado nativo.

Uma enorme quantidade de evidências experimentais tem sido acumulada desde 1890, dando suporte a esse paradigma estrutura/função (Radivojac, Iakoucheva *et al.*, 2007). Alguns trabalhos que se transformaram em referência na área merecem destaque: a) os modelos teóricos postulados por Pauling, Corey e Branson (Pauling, Corey *et al.*, 1951); b) o modelo chave-fechadura introduzido por Fischer (Fischer, 1894); c) o primeiro modelo de uma estrutura cristalizada a partir de uma proteína globular (Kendrew, Bodo *et al.*, 1958; Kendrew, Dickerson *et al.*, 1960) e d) os estudos que mostram a possibilidade de renaturação de uma proteína ao seu estado funcional, partindo de seu estado desnaturado (Anson e Mirsky, 1925; Anfinsen, 1973).

Embora tenha havido experimentos durante todo o século XX demonstrando a relação estrutura função, alguns estudos mostraram vários exemplos de proteína que desempenhavam funções biológicas, porém não eram capazes de manter uma conformação tridimensional estável em condições fisiológicas. A maioria desses casos foi ignorada ou se mantiveram inexpressivos diante do sucesso dos experimentos que demonstravam de maneira elegante, a relação entre estrutura e função (Radivojac, Iakoucheva *et al.*, 2007).

Apesar de terem sido ignoradas por muitos anos sendo inclusive consideradas erros de predição computacional, proteínas com desordem estrutural representam um importante recurso utilizado pelos organismos para a realização de

algumas funções biológicas essenciais.

Um exemplo clássico, e com amplo suporte experimental, é o processo de ativação do tripsinogênio, forma precursora da enzima pancreática tripsina.

Se o tripsinogênio fosse convertido em sua forma ativa, a tripsina, dentro do pâncreas, causaria a destruição intracelular e o extravasamento deste conteúdo no meio intersticial. Esse processo leva a necrose tecidual. Para evitar essa progressão indesejada, a tripsina é produzida em sua forma inativa, o tripsinogênio. Depois de sintetizado (traduzido), o tripsinogênio se enovela em uma forma tridimensional estável, mas se comparado a tripsina, o enovelamento é incompleto, e a proteína é inativa. O tripsinogênio se mantém inativo, pois o sítio de ligação ao substrato (lisina ou arginina) não se forma completamente mantendo-se estruturalmente desordenado. O enovelamento na sua forma ativa é impedido por uma curta região desordenada na extremidade N terminal. Uma vez que o tripsinogênio é exportado para fora da célula, essa região desordenada é clivada pela tripsina. Assim, a característica altamente carregada da extremidade N terminal é substituída por isoleucinas seguidas de valinas, conferindo um caráter hidrofóbico a essa região, permitindo o correto enovelamento, e a conseqüente ativação da enzima (Daughdrill, Pielak *et al.*, 2005).

Tais exemplos evidenciam a importante relação entre estrutura e função protéica, mas também traz a tona uma discussão importante, o papel da desordem estrutural no desempenho de funções biológicas importantes.

1.1.2 IUPs e sua inter-relação com estados de saúde e doença

Em 2002, uma revisão da literatura realizada por Dunker e colaboradores (Dunker, Brown *et al.*, 2002) envolvendo estudos de aproximadamente 90 proteínas (com evidências experimentais relacionando desordem e função), revelou que a grande maioria das IUPs ou domínios desordenados conhecidos estão envolvidos na regulação ou sinalização celular e que esses processos estão diretamente ligados a interações não catalíticas com DNA, RNA ou outras proteínas.

Essas regiões desordenadas se tornam ordenadas após a ligação com seus alvos, confirmando assim a hipótese de que uma estrutura 3D estável não é estritamente necessária para o reconhecimento biomolecular (Spolar e Record, 1994), (Demchenko), (Dyson e Wright, 2002).

O reconhecimento molecular envolvendo regiões desordenadas de proteínas tem duas características que resultam em importantes vantagens funcionais para sinalização e regulação. Primeiro, regiões desordenadas podem se ligar a seus alvos com alta especificidade e baixa afinidade (Dunker, Lawson *et al.*, 2001). Segundo, a desordem estrutural permite a diversidade de ligação, pois possibilita a interação com vários alvos diferentes (Dyson e Wright, 2002).

O estudo realizado por Chervitz e colaboradores (Chervitz, Aravind *et al.*, 1998) comparando os genomas completos de uma levedura unicelular, a *Saccharomyces cerevisiae*, e de um nematódeo multicelular, o *Caenorhabditis elegans*, sugere que organismos multicelulares desenvolveram um controle muito mais elaborado de regulação e transdução de sinal que envolve a utilização de domínios compostos que nos organismos unicelulares não estão ligados a esses processos. Segundo Chervitz, esses "novos domínios" seriam gerados a partir da junção de domínios já existentes, sendo o acoplamento desses domínios realizado por segmentos desordenados.

Uma vez que o número de estruturas associadas aos domínios é restrita, os autores sugerem que a existência de múltiplos domínios conectados por regiões de desordem reflita um mecanismo fundamental para o surgimento da diversidade observada nos organismos multicelulares.

Também tem sido relatado que a evolução do estado normal de uma célula para um estado cancerígeno está intimamente relacionada a problemas no controle do ciclo celular (Hartwell e Kastan, 1994). De fato, um estudo comparativo envolvendo dois grupos distintos de proteínas revelou uma grande diferença na porcentagem de desordem estrutural presente em proteínas humanas envolvidas no câncer se comparada a proteínas depositadas no PDB (*Protein Data Bank*) e do Swiss-Prot (exceto proteínas de controle do ciclo celular).

Outro fato que merece atenção está relacionado ao elevado grau de exposição das cadeias polipeptídicas das IUPs que podem sofrer extensivas modificações pós-traducionais, tais como fosforilação, acetilação e/ou ubiqüitinação, permitindo a modulação de sua função ou atividade biológica.

Como exemplo desse fato, a proteína p27, que pode ser classificada como uma IUP e tem sido relacionada ao câncer de mama pode passar pelo processo de fosforilação de duas maneiras diferentes. A primeira via, considerada normal, envolve a fosforilação da p27 no resíduo Thr187, evento esse que leva a proteína a

ser ubiqüitinada e posteriormente degradada pela subunidade 26S do proteassoma. A segunda via, documentada nos processos cancerígenos, envolve uma fosforilação no peptídeo de localização nuclear, que impede a interação da p27 com a maquinaria de importação nuclear, fazendo com que a proteína se mantenha no citoplasma. Apesar de normalmente localizada no núcleo, a p27 encontra 'novos' alvos no citoplasma e exibe um ganho de função oncogênica (Galea, Pagala *et al.*, 2006).

Outro aspecto interessante relacionado à existência de domínios desordenados em proteínas está relacionado ao desenvolvimento de drogas. Trechos da seqüência protéica apresentando desordem estrutural, pela flexibilidade e grau de exposição, podem representar um facilitador no desenvolvimento de novas drogas, pois uma vez identificados podem representar potencias sítios de interação de fármacos (Galea, Pagala *et al.*, 2006).

Finalmente, mas não menos interessante, existe a constatação do envolvimento de proteínas do tipo IUP nos processos de interação parasito hospedeiro (Feng, Zhang *et al.*, 2006). A necessidade de interação com um ligante para adotar uma conformação definida atrelada a alta promiscuidade das interações proteína-proteína que são capazes de realizar conferem vantagens cinéticas a essas proteínas. Tais características as tornam especialmente versáteis nos processos de adesão, invasão e sobrevivência dentro do hospedeiro, sendo responsáveis, por exemplo, por garantir ao *Plasmodium falciparum* sua natureza polimórfica (Feng, Zhang *et al.*, 2006).

Dentro do contexto exposto acima fica evidente a relação das IUPs com relevantes processos saúde e doença.

1.1.3 Desordem Estrutural

1.1.3.1 Definição e Nomenclatura

Vários termos têm sido utilizados para descrever trechos ou proteínas inteiras que não são capazes de formar estruturas tridimensionais específicas, dentre eles: **'*natively denatured*', '*natively unfolded*', '*intrinsically unstructured*', '*intrinsically disordered*'**.

Contudo, nenhum desses termos é completamente apropriado. A utilização do termo '*natively*', por exemplo, deve ser feita com cautela uma vez que a constatação de que a proteína se encontra no seu estado nativo não é fácil. Mesmo em condições fisiológicas uma proteína pode falhar no seu enovelamento uma vez que a presença de ligantes essenciais atrelada à existência de um ambiente apresentando elevada concentração de moléculas semelhantes na maioria das vezes representa condições indispensáveis ao enovelamento.

Assim sendo, levando-se em conta tais incertezas com relação ao real estado nativo das proteínas, o termo '*intrinsically*' em geral é mais utilizado do que o '*natively*'.

Os termos ***'denatured'***, ***'unfolded'*** e '***unstructured***' são também regularmente empregados para evidenciar a falta de organização estrutural da molécula. Apesar disso, proteínas classificadas como nativamente desenoveladas (***natively unfolded***) freqüentemente conservam sua estrutura secundária, algumas vezes de modo transiente, algumas vezes de modo persistente.

Por fim, existe o termo '***disordered'*** também bastante utilizado e que define a completa falta de organização estrutural. Esse termo tem sido proposto como o mais adequado para a definição de desordem estrutural protéica.

Dentro desse repertório de termos selecionamos dois deles que, a partir de agora, serão utilizadas nesse trabalho, são elas: ***'intrinsically unstructured'*** e **'*intrinsically disordered'***. Proteínas intrinsecamente desordenadas (ou de maneira mais simples, proteínas com desordem estrutural) é a terminologia que utilizaremos para referenciar o fenômeno de falta de estrutura protéica em português.

Vale ainda ressaltar que não existe consenso na literatura especializada sobre todos os fatores globais que influenciam a desordem estrutural, e como vimos, tão pouco existe consenso na nomenclatura empregada. A falta de um vocabulário bem definido além de dificultar as buscas léxicas na literatura especializada provoca equívocos na interpretação das inter-relações estabelecidas. Como alguns termos são "sinônimos" e se referem a estados transientes de ordem e desordem, faz-se necessário o estabelecimento de uma ontologia, que descreva os diferentes estados estruturais da molécula, de sua completa falta de estrutura, até a forma final e estável.

1.1.3.2 Características associadas à predição computacional

Ao longo da década de 90 e se estendendo até os dias de hoje existe uma efervescência nas metodologias associadas à predição de desordem estrutural.

Inúmeros algoritmos utilizando diferentes propriedades físico-químicas e características de composição protéica têm sido empregadas com relativo sucesso (Ferron, Longhi *et al.*, 2006). A seguir enumeramos as principais propriedades das IUPs que servem como alvo para predição computacional.

1.1.3.2.1 Composição de aminoácidos

IUPs usualmente têm uma predileção por determinados aminoácidos. Dois trabalhos independentes (Dunker, Lawson *et al.*, 2001); (Linding, Russell *et al.*, 2003) resultaram em uma mesma regra empírica: a) os aminoácidos G, S e P são promotores de desordem estrutural, ou seja, contribuem para a falta de estabilidade estrutural da molécula. b) os aminoácidos W, F, I, Y, V e L são promotores de ordem estrutural, ou seja, contribuem para a estabilidade estrutural da molécula. c) os aminoácidos H e T são considerados neutros com respeito à desordem estrutural.

1.1.3.2.2 Ausência de predição de estruturas secundárias

A predição de estruturas secundárias é baseada na probabilidade de cada aminoácido fazer parte de um tipo de estrutura secundária conhecida. A probabilidade é calculada através de uma janela deslizante sobre a seqüência da proteína. O consenso de vários preditores de estrutura secundária é avaliado. Se todos resultam na mesma predição de estrutura secundária para um mesmo aminoácido, a probabilidade de que esse aminoácido esteja em uma região estruturada é alta. A probabilidade cai à medida que um número menor de preditores esteja de acordo. Regiões longas (>70 aa) sem nenhum tipo de predição de estrutura secundária são geralmente desordenadas (Ferron, Longhi *et al.*, 2006).

1.1.3.2.3 Baixa complexidade da seqüência

Regiões de baixa complexidade são regiões que apresentam algum viés

composicional. Essas regiões podem apresentar trechos contendo repetições curtas de alguns aminoácidos, uma freqüência sutilmente maior de alguns resíduos ou ainda trechos homopoliméricos. IUPs usualmente apresentam regiões de baixa complexidade em suas seqüências, entretanto isso não é uma regra (Ferron, Longhi *et al.*, 2006).

Nos anos 20, Wu e colaboradores demonstraram que proteínas nativas (ordenadas) são mais resistentes a degradação por proteases do que proteínas desnaturadas (ou desestruturadas, como as IUPs) (Wu, 1995). Considerando uma protease já bem estudada, a tripsina, a maioria dos seus resíduos alvos, lisina e arginina, estão localizados na superfície da proteína, entretanto, poucos desses resíduos são sítios de digestão se estiverem localizados em regiões estruturalmente ordenadas da molécula. Em 1994, Hubbard e colaboradores, utilizando métodos de simulação molecular, realizou ensaios para avaliar o encaixe de resíduos potencialmente alvos de ação da tripsina no sítio ativo dessa protease. Esse estudo sugere que é necessário o desenovelamento (desestruturação) de pelo menos dez resíduos próximos ao resíduo alvo para permitir o ajuste necessário no sítio ativo da tripsina (Hubbard, Eisenmenger *et al.*, 1994). Esse resultado corrobora a afirmação de Wu sobre a alta suscetibilidade das proteínas desnaturadas a ação de proteínas de digestão, uma vez que proteínas desnaturadas apresentam seus aminoácidos alvos expostos ao sítio ativo das proteases. Partindo desse pressuposto, recentemente foi demonstrado por Dyson e colaboradores, que regiões de baixa complexidade são mais sensíveis a degradação por proteases do que outras porções da proteína (porções ordenadas) (Dyson, Shadbolt *et al.*, 2004).

1.1.3.2.4 Alta variabilidade da seqüência

Regiões desordenadas são em média muito mais variáveis do que regiões ordenadas (Brown, Takayama *et al.*, 2002), isto é, apresentam mais mutações em suas seqüências. Em um estudo realizado por Dunker e colaboradores, foi demonstrado que a região desordenada de oito proteínas (de uma mesma família, *calcinerium*) apresenta menor similaridade de seqüência do que a região ordenada das mesmas oito proteínas (Dunker, Garner *et al.*, 1998). Shaiu e colaboradores destacaram que a região desordenada da topoisomerase II apresenta mais substituições de aminoácidos, inserções e deleções do que a região ordenada da

mesma proteína (Shaiu, Hu *et al.*, 1999). A razão pela qual elas apresentam tal variabilidade ainda não é clara, entretanto já se sabe há tempos que existe correlação entre variabilidade de seqüência e flexibilidade da molécula. Quando uma proteína não cristaliza após várias tentativas, os cristalógrafos removem as regiões hipervariáveis da seqüência, pois são supostamente regiões extremamente flexíveis da molécula. Alta variabilidade da seqüência não é por si só uma evidência de desordem estrutural, mas é um indicador. Um alinhamento múltiplo de proteínas pode ser uma ferramenta bastante útil na identificação de tais regiões (Ferron, Longhi *et al.*, 2006).

1.1.3.2.5 Baixa hidrofobicidade e alta carga de rede

O enovelamento de uma proteína se define pelo equilíbrio de forças de atração (de natureza hidrofóbica) e de forças de repulsão eletrostática entre resíduos de cargas similares.

IUPs em geral apresentam uma menor quantidade de resíduos hidrofóbicos e uma grande quantidade de resíduos com cargas similares. Como resultado desse contexto composicional, domínios hidrofóbicos (agrupamento estrutural que protege aminoácidos hidrofóbicos do solvente aquoso) são raramente encontrados nas regiões de desordem e isso atrelado a elevada repulsão entre a maioria dos resíduos da molécula dificulta o enovelamento (Ferron, Longhi *et al.*, 2006).

A identificação dessas características de composição particular das IUPs é utilizada como alvo de estratégias computacionais para predição de desordem estrutural, como veremos no item 1.4.2.

1.1.4 Predição computacional de desordem estrutural

A maioria dos métodos de identificação computacional de desordem estrutural é baseada na busca das características descritas no item 1.3.3.2. Em linhas gerais, a maioria dos programas passa por um processo de “treinamento” onde um conjunto de seqüências de IUPs, cujas evidências de desordem estrutural possuem validação experimental, é fornecido ao programa. Essas seqüências são utilizadas pelos programas para a identificação de padrões em regiões que apresentam desordem e regiões ordenadas. Esses padrões podem ser representados internamente pelos

programas das mais variadas formas, tais como matrizes de peso posição específicas, modelos estatísticos, dentre outros. Os padrões identificados durante a fase de treinamento dos programas são utilizados posteriormente para classificar seqüências desconhecidas.

Para cada conjunto de características descritas no item 1.3.3.2, há um programa (no contexto de identificação de características biológicas, esses programas são conhecidos como preditores) que desempenha melhor as buscas. Essa diferença de desempenho dos preditores está relacionada à existência de um grande número de metodologias distintas. Como vimos anteriormente, a inexistência de um consenso na definição de IUPs, levou ao desenvolvimento de algoritmos muito diversos e assim não podemos eleger um deles como sendo o estado da arte na predição de desordem estrutural protéica. Na verdade, a melhor abordagem de predição seria a integração de diferentes preditores, como foi demonstrado por Ferron e colaboradores em 2006 (Ferron, Longhi *et al.*, 2006).

Veremos adiante alguns dos preditores recomendados pelo *Center for Computational Biology & Bioinformatics* da *Indiana University/USA.* Esse grupo mantém o único banco de dados dedicado a IUPs existente atualmente, o DisProt (www.DisProt.org).Abordaremos também os preditores que foram selecionados para esse trabalho.

1.1.4.1 PONDR

O PONDR, talvez o mais famoso dos preditores de desordem estrutural, classifica uma região como desordenada, se esta apresentar baixa complexidade de seqüência, alta flexibilidade, baixa hidrofobicidade e alto valor de carga. Diferentes versões de redes neurais são oferecidas, cada uma com as suas peculiaridades, com maior especificidade para seqüências curtas, ou para a determinação de sítios de ligação ou domínios funcionais (Ferron, Longhi *et al.*, 2006). Trata-se de um programa de uso restrito (não há livre distribuição, nem para uso acadêmico), mas mesmo nessas condições é um dos mais citados na literatura especializada.

1.1.4.2 SEG

O programa SEG foi originalmente desenvolvido para o cálculo da complexidade de seqüência protéica, não tendo como principal objetivo o uso na identificação de desordem estrutural, entretanto, Koonin e seu grupo obtiveram êxito nessa tarefa utilizando o SEG (Koonin e Galperin, 2003). Para tanto, o que Koonin fez foi utilizar parâmetros diferentes do padrão para possibilitar a identificação de regiões com desordem estrutural, tendo como idéia central, a variação do tamanho da janela na qual se avalia o desvio composicional da seqüência.

1.1.4.3 Disopred2

O Disopred2 por sua vez, incorpora a informação de alinhamentos múltiplos, pois seus dados de entrada são baseados em perfis de seqüência gerados pelo PSI-BLAST. As seqüências pras quais se deseja realizar as predições de desordem são alinhadas com um banco de dados contendo seqüências de referência. Essas seqüências contêm evidências experimentais de ordem e desordem estrutural. A comparação da variabilidade das seqüências com o banco de dados de referência (utilizando-se o algoritmo PSI-BLAST) da origem a uma matriz de *scores,* chamada PSSM (*Position Specific Score Matrix*). Essa matriz informa ao algoritmo quão similar é a variabilidade das seqüências fornecidas (em coordenadas específicas) com relação às seqüências do banco de referencia. Dessa forma, a informação das regiões de alta variabilidade de seqüência é utilizada para auxiliar na identificação de regiões de desordem estrutural. Além da informação dos perfis de seqüência, o Disopred2 realiza sua classificação baseado nas decisões de um SVM (*Support Vector Machine*) treinado com um conjunto de dados contendo proteínas globulares e IUPs.

1.1.4.4 NORSp

O NORSp baseia-se no princípio de que longas regiões de desordem não apresentam predições de qualquer tipo de estrutura secundária, e além disso são acessíveis ao solvente. Entretanto existem exceções a essa regra geral empregada pelo algoritmo. O domínio *Kringle,* por exemplo, não apresenta estrutura secundária

regular, mas ainda assim é ordenado. Composto por um *loop* triplo, um domínio ligado por pontes dissulfeto, encontrado em algumas proteases de serina e algumas proteínas do plasma.

1.1.4.5 PreLink

O PreLink realiza suas predições baseando-se nas diferenças de composição de aminoácidos das regiões desordenadas, e na identificação de grupos de resíduos com baixo conteúdo hidrofóbico. Esse preditor tem a peculiaridade de predizer como ordenada uma região desordenada que apresenta potencial de ser ordenada na presença de um ligante. Se associado a outros preditores, o resultado das predições do PreLink pode ser utilizado para auxiliar a identificação de regiões da proteína que irão adquirir ordem estrutural quando interagirem com um ligante. Uma característica que auxilia a identificação de domínios funcionais nas IUPs.

1.1.4.6 Análise carga/hidropatia

Como o enovelamento de uma proteína depende do equilíbrio entre forças de atração e repulsão, é possível avaliar a estabilidade estrutural de uma proteína calculando-se a razão entre essas duas forças. Apesar disso, esse cálculo serve apenas para fornecer uma indicação global sobre a estabilidade estrutural da molécula, não sendo possível dizer, por exemplo, se a molécula possui trechos estruturados e trechos desestruturados. Entretanto um método chamado FoldIndex resolve essa questão calculando a razão carga/hidropatia em uma janela deslizante ao longo da extensão da proteína. Avaliando a razão entre forças de atração e repulsão em pequenos trechos ao longo da extensão da proteína, e não na sua seqüência como um todo, o algoritmo consegue identificar regiões da seqüência protéica com maior propensão a desordem estrutural do que outras.

1.1.4.7 HCA

A técnica chamada HCA (*Hydrophobic Cluster Analysis*) oferece informações não somente sobre o estado de ordem e desordem de um determinado aminoácido, mas também do seu potencial de enovelamento. É especialmente útil para ajudar a

definir os limites de uma região desordenada, em conjunto com outras técnicas de predição.

O HCA usa a representação da seqüência de proteína em padrões de α hélice. A premissa postulada é que o polipeptídio nascente é uma α hélice flutuante quando produzida pelo ribossomo, antes da estrutura nativa da proteína (Figura 3).

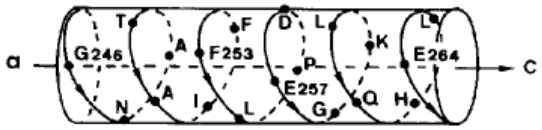


Figura 1: Representação em α hélice da seqüência de uma proteína. Adaptado de Gaboriaud, 1987 (Gaboriaud, Bissery *et al.*, 1987).

Para facilitar a manipulação desse modelo tridimensional, o cilindro hipotético formado pela α hélice é cortado em um plano paralelo ao seu eixo, e então desenrolado, assumindo uma forma plana (Figura 4).

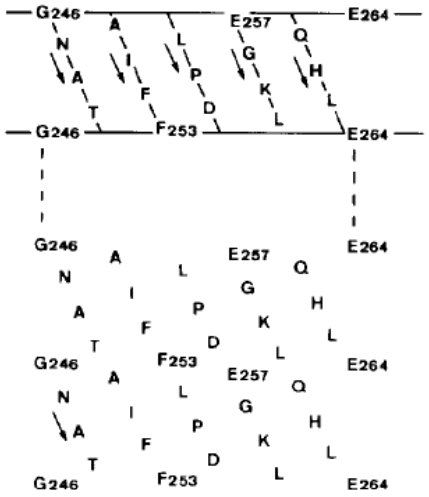


Figura 2: Representação plana da α hélice apresentada na figura anterior. Adaptado de Gaboriaud, 1987 (Gaboriaud, Bissery *et al.*, 1987).

Como alguns aminoácidos adjacentes ficam muito distantes na representação plana do cilindro, a sua representação é duplicada, tornando a seqüência mais fácil de ser seguida ao longo da hélice e dando uma melhor representação da vizinhança de cada aminoácido.

Conjuntos de resíduos hidrofóbicos adjacentes são então circulados, definindo um “cluster hidrofóbico” (Figura 5).

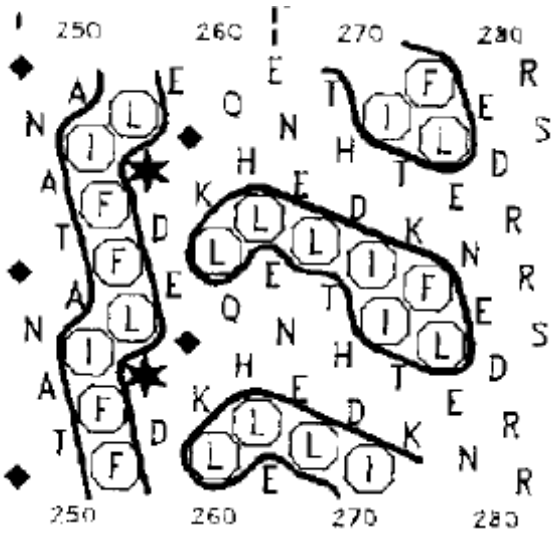


Figura 3: Representação dos clusters hidrofóbicos. Clusters circulados na representação plana da α hélice. Adaptado de Gaboriaud e colaboradores, 1987 (Gaboriaud, Bissery *et al.*, 1987).

Com os clusters hidrofóbicos devidamente marcados no gráfico HCA, é possível identificar as coordenadas limítrofes de possíveis regiões desordenadas entre cada cluster hidrofóbico.

A desvantagem dessa metodologia é não fornecer um dado quantitativo quanto à predição de desordem, exigindo assim uma interpretação humana de seu resultado.

Existem ainda preditores que implementam outras técnicas de identificação de desordem estrutural e que são abordados em detalhes adiante, pois foram os selecionados para esse trabalho.

1.1.4.8 DisEMBL

O DisEMBL incorpora três metodologias diferentes de predição, todas implementadas na forma de redes neurais. Redes neurais são algoritmos computacionais de classificação. Dado um conjunto de indivíduos pertencentes a duas classes, e conhecendo-se a verdadeira classificação de cada indivíduo, após um período de treinamento, o algoritmo é capaz de aprender a identificar as características que definem indivíduos de cada uma das duas classes. Na fase de classificação, o algoritmo utiliza as características aprendidas na fase de treinamento para classificar indivíduos desconhecidos. Para possibilitar o aprendizado durante a fase de treinamento e a posterior classificação, o algoritmo utiliza um sistema de pesos para as diversas características de cada indivíduo.

Esses pesos são calculados por funções sigmóides conectadas de maneira esquemática que se assemelha a sinapses entre neurônios de um sistema nervoso, daí o nome sugestivo de redes neurais (Gibas e Jambeck, 2001).

Duas das três metodologias de predição do DisEMBL se baseiam em conceitos propostos por Kabsch e Sander em 1983, no dicionário de estruturas secundárias de proteínas (DSSP na sigla em inglês) (Kabsch e Sander, 1983).

A primeira metodologia baseada no DSSP é chamada de *COILS*. Nesse preditor, resíduos são classificados como pertencentes a algum tipo de região ordenada, caso estejam localizados em uma α hélice ou folha β. Todos os resíduos que se localizem nas demais estruturas, *random coils, loops, turns*, dentre outras, são considerados como pertencentes a uma região estruturalmente desordenada. De acordo com Linding e colaboradores, coils não são necessariamente desordenados, entretanto, desordem estrutural só é encontrada dentro de tais regiões (Linding, Jensen *et al.*, 2003). Assim, para um resíduo estar em uma região de desordem, deve necessariamente estar dentro de um *loop/coil*, mas só essa premissa não é suficiente.

HOTLOOPS é a segunda metodologia baseada no DSSP. *Hotloops* são *loops* com um alto grau de mobilidade, determinado pelo fator de temperatura do carbono α (*B-factor*). O *B-factor* de um cristal de proteínas reflete a flutuação dos átomos ao redor de suas posições médias, fornecendo informações importantes acerca da dinâmica da proteína (Yuan, Bailey *et al.*, 2005). Considera-se que loops com alto grau de mobilidade podem ser considerados regiões desordenadas estruturalmente. O poder de discriminação entre regiões de ordem/desordem dessa rede neural é muito maior, se comparado com o *COILS*.

Relatórios do PDB possuem um campo específico (REMARK 465) para armazenar informações sobre resíduos que não possuem coordenadas definidas em uma estrutura de raio X. Um resíduo pode receber a etiqueta remark465 por diversas razões, dentre elas, por estar dentro de uma região desordenada. Valendo-se dessa característica, um conjunto de treinamento foi construído com relatórios do PDB, contendo exemplos de proteínas estruturadas e desestruturadas, e uma rede neural foi treinada para discriminar entre os dois tipos de proteínas. Esse preditor apresenta uma alta taxa de falsos positivos (16%) (Linding, Jensen *et al.*, 2003), e isso provavelmente se deve ao fato de que o campo remark465 pode ter sido atribuído a um resíduo por motivos que não tem relação com desordem estrutural.

1.1.4.9 GlobPipe

Nesse programa, a predição de regiões de desordem é baseada na propensão de cada resíduo de uma proteína estar em uma região desordenada (*randon-coil* segundo DSSP), ou em uma estrutura secundária regular (segundo DSSP). Assim, a propensão resultante para cada resíduo se define por:

$$P = RC - SS$$

Onde P é a propensão resultante para cada resíduo, RC é a propensão do resíduo estar em uma região de *randon-coil* e SS é a propensão do resíduo estar em uma região de estrutura secundária. Os valores de propensão de cada aminoácido, para cada uma das situações (RC e SS), foram calculados segundo metodologia desenvolvida por Deleage and Roux (Deléage e Roux) baseando-se em um conjunto de seqüências obtidas do banco de dados SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) (Lo Conte, Brenner *et al.*, 2002). Para a construção desse conjunto de seqüências, selecionou-se uma representante de cada superfamília contemplada no SCOP. O algoritmo de predição se define por:

$$\Omega(a_i) = \sum_{j=1}^{i=1} \Omega(a_j) + \ln(i+1) \cdot P(a_i) \quad \text{for } i = 1, \ldots, L$$

Para cada aminoácido 'a', foi definida uma propensão '$P(a_i)$'. Dada uma proteína de extensão L, define-se a soma Ω, onde $P(a_i)$ é a propensão resultante do enésimo aminoácido, e ln é o logaritmo natural. Após o cálculo da soma Ω, um filtro Savitzky-Golay (Linding, Russell *et al.*, 2003) é aplicado sobre a soma, para suavizar a curva e obter um valor numérico da derivada de primeira ordem. É construído então o gráfico da curva suavizada obtida. Nesse gráfico, picos positivos em uma janela definida pelo usuário representam regiões desordenadas, e picos negativos representam regiões globulares.

1.1.4.10 IUPred

O conceito de desordem empregado nesse programa baseia-se na energia resultante das interações inter-resíduos. Proteínas globulares realizam uma grande quantidade de interações inter-resíduos, em proporção suficiente para superar a

entropia durante o enovelamento. Assim, a estrutura se mantém estável. Em contra partida, IUPs não são capazes de realizar interações em quantidade e força suficientes para garantir a estabilidade da estrutura. Para discriminar proteínas globulares de proteínas estruturalmente desordenadas, o IUPred utiliza uma abordagem que estima o potencial dos polipeptídios de formarem contatos estabilizantes, através do uso de um modelo de interação estatístico (Thomas and Dill, 1996; Dosztányi et al., 2005). A contribuição de um aminoácido para a estabilidade estrutural da molécula não depende somente de suas propriedades físico-químicas, mas também do seu potencial de interação com os aminoácidos vizinhos (Dosztányi et al., 2005). Assim, a soma das energias de interação pode ser estimada por uma expressão quadrática para cada aminoácido da seqüência.

$$Z_i = \sum_k (e_i^k - N_i^k \sum_{j=1}^{20} P_{ij} n_j^k)^2$$

O primeiro elemento da subtração é a energia calculada de interação do aminoácido i com o aminoácido j. O segundo elemento da subtração estima como a energia do aminoácido i depende do aminoácido j. A soma Z_i define a contribuição energética do aminoácido i para a estabilidade estrutural da molécula.

1.1.4.11 VSL2

O VSL2 é um preditor de desordem estrutural baseado em métodos de aprendizado de máquina. Ou seja, possui um algoritmo inteligente, que aprende a classificar novos exemplos comparando-os com exemplos vistos durante sua fase de treinamento.

O algoritmo de aprendizado utiliza nesse preditor é o SVM (*Support Vector Machine*). Trata-se de uma modificação no princípio do algoritmo de redes neurais (implementado no preditor DisEMBL (vide item 1.3.4.8)).

Para alimentar o SVM com informações que possibilitem o aprendizado e a posterior classificação de novas seqüências, é necessário que se forneça informações relevantes da seqüência de aminoácido no que diz respeito à desordem estrutural.

O VSL2 pode fornecer ao SVM características da seqüência de diversas naturezas, permitindo assim uma predição mais robusta. As características

compreendem: a) predições realizadas somente a partir da seqüência de aminoácidos; b) informações sobre a variabilidade da seqüência (derivadas de alinhamentos); e c) informações sobre a predição de estruturas secundárias.

O VSL2 possui três variantes: a) VSL2B: utiliza somente informações obtidas a partir da seqüência de aminoácidos; b) VSL2P utiliza informações do VSL2B mais informação da variabilidade da seqüência e; c) VSL2 utiliza os três tipos de características descritas.

Qualquer uma das três variantes do VSL2 consiste de três componentes preditores em dois níveis de arquitetura. No primeiro nível, há dois preditores especializados: um preditor de desordem curta (VSL2-S) para regiões desordenadas ≤ 30 resíduos e um preditor de desordem longa (VSL2-L) para regiões desordenadas > 30 resíduos. No segundo nível, há um *meta preditor* (*M1*) que combina os resultados dos dois preditores especializados em uma predição final.

Para a variante VSL2B (única utilizada nesse trabalho) o vetor de características é extraído da seqüência de aminoácidos utilizando uma janela deslizante. No total são extraídas 26 características da seqüência cujas origens são: a) flexibilidade da molécula; b) hidrofobicidade c) carga elétrica d) entropia e) razão carga/hidropatia f) informação composicional (freqüência individual de resíduos).

Durante o desenvolvimento do preditor, o SVM foi treinado com um conjunto de teste contendo 1.327 seqüências de origens variadas, compreendendo 1606 regiões desordenadas.

O *M1* é treinado independentemente do VSL2-S e do VSL2-L, mas com o mesmo conjunto de características. Se o resultado do *M1*, O_M é aproximadamente 1 (ou 0), o resíduo estará em ou próximo a uma região desordenada longa (ou curta) e se estiver próximo a 0,5 o resíduo está em uma região ordenada. A predição final pode ser calculada por $O_L \times O_M + O_S \times (1-O_M)$, onde O_L e O_S são os resultados do VSL2-L e do VSL2-S, respectivamente (Peng, Radivojac *et al.*, 2006).

1.1.5 Métodos experimentais de identificação de desordem estrutural

Inúmeras metodologias experimentais têm sido desenvolvidas para o estudo de proteínas individuais e do proteoma (conjunto de proteínas preditas) dos diferentes genomas estudados. Dentro desse conjunto existe um grande número de métodos experimentais capazes de detectar ausência de estrutura em proteínas

(para uma revisão sobre essas metodologias veja (Daughdrill, Pielak *et al.*, 2005)). Um apanhado que inclui um resumo das principais abordagens experimentais que podem ser utilizadas na identificação de regiões protéicas sem estrutura pode também ser encontrado no sitio http://www.DisProt.org/view_detection.php.

As técnicas disponíveis para a identificação de IUPs são muito diversas, e vão de métodos baseados em dicroísmo circular, sensibilidade a proteases, estabilidade térmica, viscometria, mobilidade aberrante em gel SDS-PAGE, fluorescência intrínseca, ressonância nuclear magnética, dentre outros.

Dentre todos esses métodos existentes, o mais utilizado na identificação de IUPs é a cristalografia de raio X. Nessa técnica, a ausência de coordenadas para alguns átomos pode indicar que tais átomos estão em uma região desordenada. A alta flexibilidade dos átomos em regiões desordenadas leva a um espalhamento não coerente dos raios X, impedindo que tais átomos sejam observados.

Abordagens proteômicas também têm sido empregadas na identificação de proteínas com desordem estrutural. Em 2007 Csizmók e colaboradores (Csizmók, Dosztányi *et al.*, 2007) descreveram técnicas proteômicas específicas para identificação em larga escala de IUPs.

Em ambos os protocolos, os autores se valem do fato de que IUPs possuem um nível baixo de aminoácidos hidrofóbicos, que contribuem para a estabilidade estrutural da molécula. Assim, com processos como aquecimento do extrato protéico ou indução de condições desnaturantes, é possível fazer com que proteínas globulares se comportem de maneira diferente das IUPs.

Num processo originalmente descrito por Csizmók em 2006 (Csizmók, Szollosi *et al.*, 2006), a separação das proteínas em duas dimensões ocorre de tal forma, que as IUPs migram para a diagonal do gel. As duas dimensões de separação são realizadas em gel de acrilamida, não havendo focalização isoelétrica. O extrato protéico todo é aquecido e posteriormente depositado em uma única canaleta de um gel nativo (não desnaturante), preparado sem SDS. Assim, a separação das proteínas se dará por sua razão de carga/massa. Após o fim dessa primeira corrida, o gel é corado, e a canela contendo as proteínas é então recortada do gel, e servirá como *strip* para a segunda dimensão do gel, assim como ocorre com a *strip* de focalização isoelétrica. A segunda dimensão então ocorre em um gel que utiliza uréia 8M como agente desnaturante. O princípio de separação das IUPs nesse protocolo se deve ao fato das proteínas globulares não conseguirem realizar

um deslocamento longo no primeiro gel, devido a sua estrutura tridimensional complexa, ao passo que as IUPs com longos trechos desordenados, percorrerão longas distâncias, pois se encontram praticamente desnaturadas. No segundo gel, esse então desnaturante, as proteínas globulares percorrerão uma distância maior, pois agora se encontram alongadas e flexíveis. As IUPs por sua vez percorrerão a mesma distância percorrida no primeiro gel, migrando assim para a diagonal do gel.

Outra metodologia proteômica descrita por Csizmók consiste no enriquecimento da quantidade de IUPs presente no extrato protéico a ser aplicado no gel através do aquecimento da amostra. Proteínas globulares, quando aquecidas tendem a se agregar e precipitar, restando somente as IUPs no sobrenadante, além de uma pequena proporção de globulares resistentes ao aquecimento (Galea, Pagala *et al.*, 2006). Essa segunda metodologia foi empregada nesse trabalho como será visto adiante.

1.2 Modelo Entidade de Relacionamento - MER

Um dos desafios do trabalho de produzir dados é conhecer e representar de maneira adequada a relação entre esses dados. Existem hoje diversas técnicas que possibilitam a representação de relações entre várias coleções de dados. Técnicas modernas de representação de dados para sistemas computacionais, como a XML, são muito eficientes na troca de informações estruturadas entre sistemas, porém ainda são precárias quanto à recuperação rápida e integrada de dados. Os bancos de dados XML nativos são utilizados em sistemas em que a origem de dados se da por meio de arquivos XML, e a sua posterior utilização também ocorre pelo mesmo tipo de arquivos. Nesses sistemas, a utilização de bancos XML nativos se justifica e é bastante simplificada pelo uso de linguagens específicas, tais como a Xquery (http://www.w3.org/TR/xquery), implementada em APIs para diversas linguagens de programação.

Modelos de Entidades e Relacionamentos (MER) é uma metodologia que se mostrou eficiente na representação da relação entre dados de naturezas diferentes. Com o uso dessa técnica, torna-se viável a utilização de computadores no armazenamento e recuperação de dados e/ou informação.

Em um MER, todo e qualquer conceito que se deseje representar, que possa ser classificado como substantivo torna-se uma entidade. Uma entidade é algo que

pode possuir características (atributos), e que possa ter algum tipo de relação com outra entidade. As relações entre as entidades são representadas por um conceito conhecido como 'relacionamento'.

As entidades e os relacionamentos de MER podem ser convertidos em um Modelo Relacional (MR). Um MR representa as entidades e relacionamentos de MER em termos de estruturas de dados que podem ser implementadas em um sistema computacional, tais como tabelas por exemplo.

A simplicidade do MER na representação de dados, e a possibilidade de converter um MER em um Modelo Relacional (MR) levaram ao surgimento de uma classe de programas extremamente poderosos; os Sistemas Gerenciadores de Bancos de Dados, ou SGBDs. Um SGBD é um conjunto de programas, que trabalha de maneira integrada e sincronizada, para viabilizar a armazenagem, a manipulação e a recuperação de dados. Tais sistemas deram origem aos bancos de dados relacionais.

Em um banco de dados relacional, as entidades de um MER são representadas por tabelas, e os relacionamentos são representados por identificadores que se repetem nas tabelas. As colunas de cada tabela representam cada uma das características de uma entidade, e em cada uma das linhas são armazenados os elementos daquela entidade. Cada elemento de uma entidade (tabela) recebe um identificador exclusivo, conhecido como chave primária. Essa chave primária se repete como chave estrangeira, nas outras tabelas que tenha relação com a primeira tabela.

A construção de um DER (Diagrama de Entidades e Relacionamentos) é o primeiro passo para a implementação de um sistema de armazenamento de dados eficiente. Nesse diagrama, as entidades, seus atributos e suas relações são representados graficamente, pois a avaliação visual do modelo é muito mais eficiente e segura.

Após o desenvolvimento correto do DER, o emprego de técnicas de mapeamento (Elmasri e Navathe, 2005) possibilitam a transformação desse diagrama, em um conjunto de tabelas com suas respectivas chaves primárias e chaves estrangeiras. A partir daí, pode-se criar as tabelas resultantes em um SGBD e então povoá-lo.

Para ter acesso aos dados armazenados em um banco de dados relacional, é preciso fazer uso de uma linguagem computacional específica para consultas. Essa

linguagem é conhecida com SQL (*Structured Query Language*). Trata-se de uma linguagem declarativa, e não 'procedural', o que a torna bastante simples de aprender e utilizar. Em SQL, os resultados que se deseja obter com a consulta ao banco de dados devem ser explícitos, entretanto, os mecanismo algébricos entre as entidades para obter esses resultados, são implícitos, sendo realizados automaticamente pelo SGBD. Essa linguagem foi adaptada diversas vezes pelos desenvolvedores de programas para computador, porém foi padronizada em 1992 pela ISO (*International Organization for Standardization*), e hoje é um padrão mundialmente aceito como método de consulta e manipulação de um banco de dados relacional (Oliveira, 2002).

1.3 Otimização de métodos de classificação

Durante a segunda guerra mundial, após o ataque japonês a base de Pearl Harbor, a marinha americana iniciou pesquisas para aprimorar a capacidade de seus radares de diferenciar aeronaves japonesas de suas próprias aeronaves. Essas pesquisas deram origem a uma nova metodologia de analise de métodos de classificação, chamada ROC (*Receiver Operating Characteristic*) (Green e Swets, 1966). O Objetivo dessa metodologia é avaliar métodos de classificação com relação a suas taxas de verdadeiros positivos e falsos positivos. Trata-se de uma maneira eficiente de relacionar a sensibilidade de um método no reconhecimento de membros de diferentes classes, e a especificidade na correta classificação de um elemento em sua classe real(Fawcett, 2004).

A comparação do desempenho de classificação de vários algoritmos é realizada, pela comparação das classificações corretas e erradas de cada método com o desempenho ideal de classificação. O desempenho de cada método é calculado com base nos seus valores de verdadeiro positivo (TP ou *true positives*), falso positivo (FP ou *false positive*), verdadeiro negativo (TN ou *true negative*) e falso negativo (FN ou *false negative*) (Fawcett, 2004).

Exemplificando, em uma situação onde um radar detecta vários objetos, e dentre eles há aviões amigos e inimigos, um algoritmo classificador ideal deve reconhecer a presença de dados referentes a duas classes distintas, e ainda classificar corretamente cada elemento como amigo ou inimigo. O número de elementos classificados como amigos, e que de fato são amigos, somado ao número

de elementos classificados como inimigos, e que de fato são inimigos, totaliza o número de verdadeiros positivos desse classificador.

O número de falsos positivos é dado pela contagem de elementos classificados como amigo, mas que na realidade são inimigos, somado ao número de elementos classificados como inimigos, mas que na realidade são amigos.

O número de falsos negativos se da pela contagem de objetos que são aviões, e que deveriam ser classificados como membros de uma classe (amigo ou inimigo), mas não o foram.

Por fim, o número de verdadeiros negativos se da pela contagem de objetos que não são aviões, ou seja, não deveriam ser classificados e que de fato não o foram.

Com esses quatro valores calculados para um algoritmo é possível a construção de uma tabela chamada matriz de confusão. Essa tabela relaciona esses quatro valores e torna possível o cálculo de duas razões que são à base do método ROC.

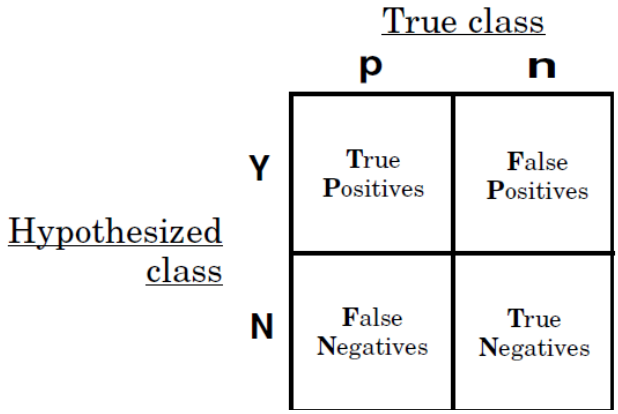


Figura 4: Matriz de confusão. Relaciona TP, FP, TN e FN. Fonte: (Fawcett, 2004).

Partindo-se da matriz de confusão, é possível se calcular duas razões conhecidas como **TPR** (T*rue Positive Rate*) e **FPR** (F*alse Positive Rate*), descritas por TPR = TP / (TP+FN) e FPR = FP / (FP + TN). Esses valores exprimem a quantidade de acertos do classificador em relação ao total de acertos possíveis.

Tais razões resultam em valores entre zero e um, sendo possível a construção de um gráfico relacionando as duas medidas, tendo TPR no eixo y e FPR no eixo x.

Esse gráfico representa o espaço ROC, no qual o par ordenado (0,1) representa o desempenho ideal de classificação. Quanto mais perto desse ponto um classificador se encontrar, melhor o seu desempenho de classificação (Fawcett,

2004).

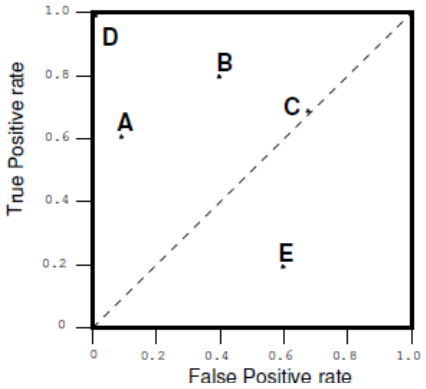


Figura 5: Espaço ROC. Gráfico representando o espaço ROC e pontos representando a distribuição do desempenho de classificação de quatro classificadores (A, B, C e E) em relação ao desempenho ideal (D).

Esse é o princípio básico da construção e de análise de um gráfico ROC, que foi utilizado originalmente para análises de dados de radar, mas que posteriormente teve sua aplicação estendida para diversas áreas tais como a psicologia, a medicina, o aprendizado de máquina e o *data mining (Green e Swets, 1966; Spackman, 1989; Zweig e Campbell, 1993; Obuchowski, 2003; Pepe, 2003)*

Hoje em dia, o emprego de gráficos ROC para análise de classificadores/preditores na bioinformática e biologia computacional se tornou comum, e a apresentação desses gráficos quando a comparação de preditores é realizada, é vista como uma prática elegante.

1.4 Organismo modelo

1.4.1 A esquistossomose

A esquistossomose é uma doença parasitaria que infecta principalmente indivíduos de países tropicais e subtropicais em desenvolvimento. O parasito causador da doença pertence ao gênero *Schistosoma*, sendo *S. mansoni* a única espécie causadora da esquistossomose no Brasil (Bergquist, 2002).

Sendo a segunda maior causadora de morbidade no mundo, dentre as doença parasitárias, perdendo somente para a malária (Chitsulo, Engels *et al.*, 2000), a esquistossomose é endêmica em 74 países. Do total de infectados no mundo, 80% vivem na África. (Figura 1) (TDR, 2005). A incidência maior dessa doença em países tropicais subdesenvolvidos se deve a falta de saneamento básico. Somando-se a esse problema, nesses países, uma boa porcentagem da população depende do contato íntimo com grandes coleções de água, tais como lagos, rios e canais de irrigação para suas atividades diárias, domésticas ou profissionais (Steinmann, Keiser *et al.*, 2006). Tais coleções de água podem estar contaminadas com o parasito (devido à falta de saneamento básico), propiciando a infecção dos indivíduos. A situação é grave, pois anualmente, essa doença leva cerca de 11 mil crianças e adolescentes à morte (com faixa etária entre 10 e 19 anos).

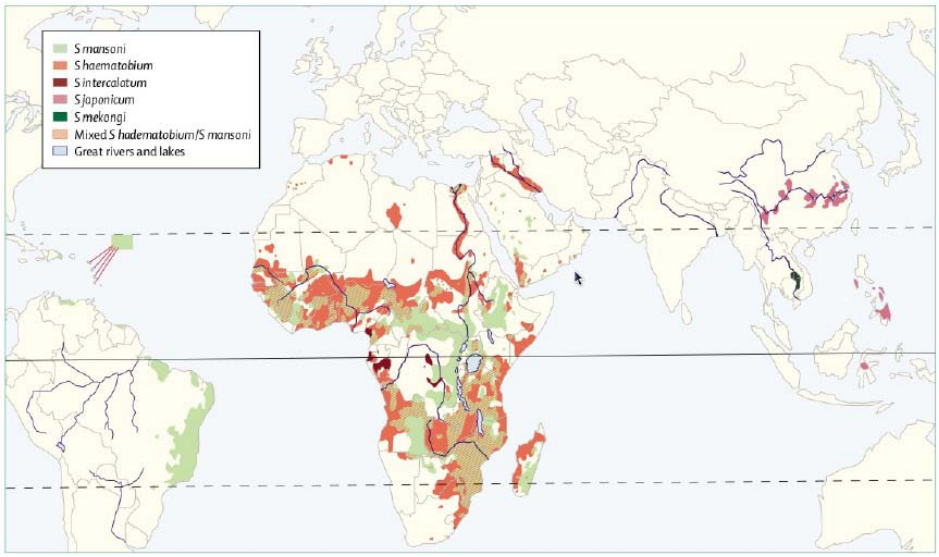


Figura 6: Distribuição da esquistossomose pelo mundo.
S. mansoni: África subsaariana, Brasil, Suriname, Venezuela, Caribe, Egito e Península Árabe; *S. haematobium*: África subsaariana, Egito, Sudão, Magrebe e Península Árabe; *S. japonicum*: China; Mindanau, província de Leyte e algumas ilhas das Filipinas e Indonésia; *S. mekongi*: Laos e Camboja; *S. intercalatum*: África ocidental e central (Gryseels, Polman *et al.*, 2006).

Dentre as espécies do gênero *Schistosoma* que causam esquistossomose no homem, as principais são *S. mansoni*, *S. japonicum* e *S. haematobium*. A infecção causada por *Schistosoma mansoni* é denominada esquistossomose mansoni ou intestinal, devido à localização dos parasitos nas vênulas da parede do intestino grosso, sigmóide e reto, com sintomas predominantemente intestinais (TDR, 2005). Estima-se que existam mais de seis milhões de pessoas infectadas pela esquistossomose mansoni no Brasil, o que o torna o pais mais afetado na America Latina (Katz e Peixoto, 2000).

O quadro sintomático depende do número de ovos e órgão onde estão localizados, sendo a forma Intestinal a mais comum - caracterizada por diarréias repetidas que pode ser muco-sanguinolento, com dor ou desconforto abdominal, podendo também ser assintomática (Disponível em http://www.cve.saude.sp.gov.br/htm/hidrica/IF_ESQUI05.htm. Acessado em 7 de janeiro de 2010).

Como as diferentes manifestações clínicas da esquistossomose podem confundir-se com outras doenças, o diagnóstico considerado padrão ouro é realizado pelo achado de ovos nas fezes, sendo o exame parasitológico de fezes o mais adequado.

1.4.2 O ciclo de vida do *Schistosoma mansoni*

O trematódeo *S. mansoni* apresenta um ciclo de vida complexo, alternado entre hospedeiro intermediário (invertebrado, caramujo do gênero *Biomphalaria* (Davis, 1984) seu hospedeiro definitivo (vertebrado) (Figura 2).

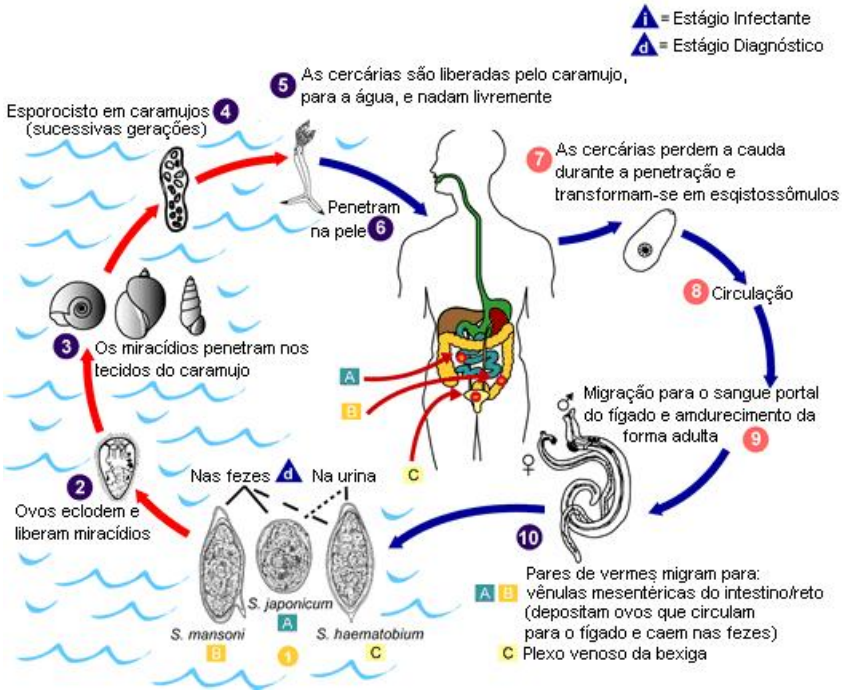


Figura 7: Ciclo de vida do *Schistosoma mansoni.*
Sob condições ótimas, os ovos eclodem e liberam os miracídios (2) que nadam e penetram no caramujo, hospedeiro intermediário especifico (3). Os estágios no caramujo incluem duas gerações de esporocistos (4) e a produção de cercarias (5). Ao abandonar o caramujo, as cercarias infectantes nadam, penetram na pele do hospedeiro humano (6), e perdem sua cauda bifurcada e tornam-se esquistossômulos (7). Os esquistossômulos migram através de diversos tecidos e desenvolvem-se ate atingirem a veia porta onde se acasalam (8, 9). Vermes adultos, nos humanos, residem nas vênulas mesentéricas em varias localizações (10). *S. mansoni* ocorre mais freqüentemente nas veias mesentéricas superiores que drenam o intestino grosso (B). As fêmeas depositam ovos nas pequenas vênulas dos sistemas porta e perivesical. Os ovos são movidos progressivamente para o lúmen do intestino e são eliminados com as (d) fezes (1). (INFORME-NET DTA Schistosoma mansoni/Esquistossomiase. Disponível em: <http://www.cve.saude.sp.gov.br/htm/hidrica/IF_ESQUI05.htm>. Acessado em: 7 de janeiro de 2010).

1.4.3 Tratamento da esquistossomose

Atualmente as estratégias para controle da doença em larga escala dependem quase que exclusivamente da administração de dose única do paraziquantel (PZQ) (Chitsulo, Engels *et al.*, 2000). Droga eficiente contra todas as espécies de *Schistosoma* que infecta o homem, possui alta taxa de cura, baixa toxicidade e baixo custo, sendo, portanto a droga de escolha da Organização Mundial da Saúde (WHO | Schistosomiasis. Disponível em: <http://www.who.int/mediacentre/factsheets/fs115/en/index.html>. Acesso em 7 de janeiro de 2010).

O pico de ação do PZQ ocorre entre 1 e 2 horas após administração e sua eliminação ocorre essencialmente por meio da urina e fezes, sendo que mais de 80% da droga são eliminados após 24 h (Cioli, Botros *et al.*, 2004). O percentual de

cura da doença causada por *S. mansoni* é de 60% a 90%, causando redução do número de ovos de 90% a 95%, dependendo do nível de infecção (De Silva, Guyatt *et al.*, 1997).

A maior desvantagem do PZQ é a sua baixa eficácia contra as formas jovens de *Schistosoma* (Gryseels, Mbaye et al., 2001). Algumas alternativas têm sido propostas para solucionar tal problema, como a associação de PZQ com artemeter, uma droga efetivamente utilizada no tratamento da malária e ativa contra as formas juvenis de *Schistosomas* (Utzinger, Xiao *et al.*, 2001).

Apesar do relativo sucesso nos tratamentos da esquistossomose com PZQ, investigações recentes demonstraram casos de parasitas resistentes ao PZQ no Senegal e Egito, bem como de resistência induzida em condições de laboratório (Doenhoff, Cioli *et al.*, 2008).

Na tentativa de encontrar uma solução mais eficiente no tratamento da esquistossomose, pesquisadores desenvolveram vários candidatos à vacina, mas nenhum produziu um nível alto ou completo de proteção (Wilson e Coulson, 2006). Ainda que muito tenha se aprendido sobre o a resposta imune contra esquistossomose, métodos tradicionais não geraram os resultados esperados e novas abordagens serão necessárias para o desenvolvimento de vacinas contra o *S. mansoni* (Wilson, Curwen *et al.*, 2004; Hokke, Deelder *et al.*, 2007; Loukas, Tran *et al.*, 2007).

1.4.4 O genoma de Schistosoma mansoni

A iniciativa internacional para o seqüenciamento do genoma completo do organismo foi conduzida pelo instituto *The Institute for Genomic Research* - TIGR em associação com o *Welcome Trust Sanger Institute* – WTSI, por meio de financiamento do *National Institutes of Health* - NIH e da *Welcome Trust*, respectivamente (Loverde, Hirai *et al.*, 2004). As últimas versões do seqüenciamento genômico e todas as análises efetuadas estão disponíveis nos bancos de dados online GeneDB (GeneDB. Disponível em: <http://genedb.org>. Acessado em: 08 de janeiro de 2010) e SchistoDB (Zerlotini, Heiges *et al.*, 2009), cujas informações foram recentemente aceitas para publicação pelo periódico Nature (Berriman, Haas *et al.*, 2009). A seqüência do genoma nuclear de *S. mansoni* foi obtida por meio da metodologia WGS (*Whole Genome Shotgun*). Essa técnica consiste na quebra de

DNA genômico em pequenos fragmentos para seqüenciamento, e a posterior organização da seqüência desses fragmentos, para recompor a seqüência original. Este processo é realizado *in silico*.

O parasito *S. mansoni* tem um genoma haplóide de aproximadamente 363 mbp contidos em 7 pares de cromossomos autossômicos e um par de cromossomos sexuais Z e W (Berriman, Haas *et al.*, 2009).

Em sua última versão, os fragmentos foram agrupados em 5.745 *scaffolds* com tamanho superior a 2kb, totalizando 363mbp. Além disso, foram identificados 11.812 genes que codificam 13.162 transcritos. Apesar de 45% do genoma ser composto de elementos repetitivos, 50% das bases estão presentes em contigs de tamanho maior que 16,3 kbp e em *scaffolds* maiores que 824,5kbp. Ainda, a localização cromossomal de 43% da montagem genômica utilizando-se hibridização *in situ* foi identificada (Berriman, Haas *et al.*, 2009).

2 JUSTIFICATIVA

Três fatos se complementam na motivação desse projeto. O primeiro deles está relacionado à recente descoberta de que proteínas com desordem estrutural podem estar envolvidas em processo de interação parasito hospedeiro (Feng, Zhang *et al.*, 2006).

A segunda descoberta que motivou a realização desse projeto, se relaciona a identificação de regiões desordenadas em proteínas envolvidas em processos saúde doença. Estudos sugerem que a presença dessas regiões pode representar um importante facilitador no desenvolvimento de novas drogas.

Por fim, com a recente publicação dos novos dados gerados por pesquisas sobre o genoma do parasito *Shistosoma mansoni* (Berriman et al., 2009), a identificação em larga escala dessas proteínas com desordem estrutural no proteoma predito de *S. mansoni* se faz essencial, uma vez que pode representar uma contribuição de grande importância na busca por terapias mais eficientes.

Somando-se a isso, o estabelecimento de uma metodologia robusta e automatizada de identificação de IUPs em escala proteômica será uma ferramenta útil para o estudo de outros parasitos.

3 OBJETIVOS

3.1 Objetivo geral

A presente dissertação tem como proposta o estabelecimento de um *pipeline* computacional visando à avaliação dos diferentes algoritmos de predição de desordem estrutural, seu desempenho e a posterior aplicação dessa ferramenta no estudo *in silico* do conteúdo de IUPs presentes no proteoma predito de *S. mansoni.*

3.2 Objetivos específicos

- Selecionar preditores de desordem estrutural disponíveis para instalação local.
- Analisar o desempenho dos preditores, individualmente e combinados, para identificar uma possível combinação de preditores que apresente melhores resultados do que cada preditor individualmente.
- Selecionar a combinação de preditores com o melhor desempenho de predição.
- Desenvolver um banco de dados que seja capaz de integrar todas as predições de desordem estrutural e predições de caracterização das proteínas.
- Integrar todos os passos em um *pipeline* automático de predição de desordem estrutural.
- Aplicar o *pipeline* ao proteoma predito de *Schistosoma mansoni*.

4 MATERIAIS E MÉTODOS

Todos os *scripts* desenvolvidos e citados estão disponíveis como material suplementar no endereço eletrônico: iup.cpqrr.fiocruz.br/IUPipeline.

4.1 Proteoma preditos de *S. mansoni*

As seqüências protéicas preditas para o genoma de *S. mansoni* foram obtidas, do repositório de genomas de parasitos GeneDB (ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma/mansoni/genome/gene_predictions/). A versão 4.0e, utilizada nesse trabalho, apresenta um total de 13.175 seqüências.

4.2 Pré-processamento

Todas as atividades de pré-processamento descritas no item 4.1 foram realizadas utilizando o *script* 'pre_processing.perl'.

Arquivo de entrada:

- arquivo texto no formato multi-fasta, contendo as seqüências de todas as proteínas preditas.

Saída gerada:

- arquivo texto no formato multi-fasta, contendo as seqüências de proteínas hipotéticas
- arquivo texto no formato multi-fasta, contendo as seqüências de proteínas com função predita.

Parâmetros utilizados:

- -l = 100 (comprimento mínimo das seqüências)
- -m = T(seleciona seqüências que iniciam com Metionina)
- -d = T (separa as proteínas em hipotéticas e com função predita)
- -a = T (seleciona seqüências sem erros de anotação)
- -i = <nome do arquivo multi-fasta com todas as seqüências preditas>

Os dois arquivos multi-fasta gerados pelo *script* de pré-processamento são utilizados como entrada para os preditores de desordem estrutural executados automaticamente nas etapas seguintes do *pipeline*.

4.3 Predição de desordem estrutural

4.3.1 Seleção de preditores de desordem estrutural

A seleção dos preditores de desordem estrutural utilizados nesse trabalho foi feita através de uma revisão bibliográfica em conjunto com as informações disponíveis no site do banco de dados DisProt (http://www.DisProt.org/predictors.php).

Todos os preditores inicialmente escolhidos foram avaliados segundo os sete critérios descritos abaixo:

1. Disponibilidade do artigo original descrevendo a metodologia de predição de desordem estrutural empregada;
2. Disponibilidade dos programas para download gratuito para Instituições Acadêmicas e de pesquisa;
3. Disponibilidade de programas acessórios e/ou bancos anexos necessários a execução do preditor.
4. Disponibilidade de documentação suficiente para a correta instalação e execução do preditor.
5. Funcionamento correto do preditor quando corretamente instalado.
6. Número de citações em trabalhos relacionados.
7. Implementação de diferentes definições de desordem estrutural.

Considerando os critérios acima enumerados, os preditores de desordem estrutural selecionados foram DisEMBL versão 1.4 (Linding, Jensen *et al.*, 2003), GlobPipe versão 2.3 (Linding, Russell *et al.*, 2003), IUPred versão 1.0 (Dosztányi, Csizmók *et al.*, 2005) e VSL2 versão 2 (Obradovic, Peng *et al.*, 2005); (Peng, Radivojac *et al.*, 2006). O resultado da avaliação segundo os sete critérios estabelecidos para os quinze preditores considerados inicialmente é apresentada em uma tabela (Anexo 4).

4.3.2 Execução dos preditores

4.3.2.1 DisEMBL

Para a execução do DisEMBL, foram utilizados os parâmetros pré-estabelecidos pelo autor, pois foram descritos como sendo os de melhor desempenho de predição. Os parâmetros foram:

- smooth_frame = 8
- peak_frame = 8
- join_frame = 4
- fold_coils = 1.2
- fold_hotloops = 1.4
- fold_rem465 = 1.2

4.3.2.2 GlobPipe

Para a execução do GlobPipe também foram utilizados os parâmetros pré-estabelecidos pelo autor uma vez que também foram descritos como sendo os de melhor desempenho de predição. Os parâmetros utilizados foram:

- smooth_frame = 10
- DOM_join_frame = 15
- DOM_peak_frame = 74
- DIS_join_frame = 4
- DIS_peak_frame = 5

4.3.2.3 IUPred

Para a execução do IUPred foi necessária a separação de cada uma das seqüências dos arquivos multi-fasta em arquivos individuais, pois este algoritmo só aceita como entrada uma seqüência por arquivo.

Para a separação das seqüências em arquivos individuais, foi utilizado o *script* Perl ‘cut_fasta.pl’ Assim, a execução desse preditor para cada uma das seqüências foi realizada através de um laço (comando ‘for’) implementado no módulo *shell* que atua como interface entre o usuário e o sistema operacional Linux.

Não existem parâmetros que alterem a sensibilidade de predição do IUPred. É possível selecionar apenas a extensão da região de desordem que se deseja encontrar. Como estávamos interessados no estudo de regiões longas de desordem estrutural (trechos desordenados com extensão mínima de 40 aa), a opção de predição selecionada foi 'long'.

4.3.2.4 VSL2B

Para a execução do preditor VSL2, foi necessária a separação do arquivo multi-fasta de entrada em arquivos individuais contendo exclusivamente a seqüência da proteína analisada sem o tradicional cabeçalho contido no formato fasta.

Para a separação das seqüências em arquivos individuais, utilizamos o *script* cut_fasta.pl. A remoção do cabeçalho das seqüências fasta foi realizada por um laço (comando *for*), que executou uma expressão regular para remoção da primeira linha de cada arquivo.

A execução consecutiva do programa para todos os arquivos de seqüências individuais sem cabeçalho foi realizada por um laço. O nome de cada arquivo foi fornecido ao programa como parâmetro. O VSL2 implementa internamente mais de um preditor, e a seleção de qual preditor será utilizado para predições depende de quais parâmetros foram fornecidos. Como nenhum outro argumento foi especificado ao programa além do nome do arquivo, o preditor executado pelo VSL2 foi o 'VSL2B'.

4.4 Predição de domínios transmembrana

Para a realização das predições de domínios transmembrana, utilizamos o programa Phobius em sua versão 1.01(Käll, Krogh *et al.*, 2004). O programa utilizou como arquivo de entrada um arquivo multi-fasta contendo 10.417 seqüências resultantes do pré-processamento. Não existem parâmetros que alterem a sensibilidade ou a especificidade das predições. Os parâmetros existentes se referem à forma de apresentação do resultado. Utilizamos a opção padrão de apresentação de resultados chamada 'long'.

4.5 Predição de características físico–químicas

Para a realização das predições de características físico-químicas, utilizamos o programa Pepstats que é parte do pacote EMBOSS (Rice, Longden *et al.*, 2000) em sua versão 5.0.0. Não existem parâmetros que alterem a sensibilidade ou a especificidade das predições.

4.6 Predição de localização celular

Para a realização das predições de localização sub-celular, utilizamos o programa TargetP, em sua versão 5.0.0 (Emanuelsson, Nielsen *et al.*, 2000). Para a execução do TargetP, utilizamos os seguintes parâmetros:

- -N (realiza predições para organismos não-vegetais)
- -c (inclui predição de sítios de clivagem)

4.7 Anotação funcional segundo termos do Gene Ontology

A anotação funcional das IUPs foi obtida por meio de similaridade de seqüência, utilizando o algoritmo BLAST (Altschul, Gish *et al.*, 1990) versus o banco de dados do *Gene Ontology*. Utilizamos como critério de corte um valor de corte de Expect (E) Value de 1.0x10-6. Selecionamos como item anotador da função das IUPs aqueles atribuídos aos 'best hits' de cada seqüência. Foi considerado somente o primeiro nível de anotação, ou seja, a classificação funcional mais geral para os três níveis de classificação do GO: função, componente e processo.

Para realizar essa análise utilizamos um *parser* (nesse trabalho chamado de 'Blaster') desenvolvido pelo Dr. José Marcos Ribeiro. Após essas análises desenvolvemos um *parser* para o 'Blaster' (descrito no item 4.8.2.9.8).

4.8 Integração das predições

4.8.1 Criação do banco de dados relacional

O DER desenvolvido para integrar todas as predições foi desenhado no programa MySQL Workbench (http://dev.mysql.com/workbench/) (Anexo 1). Em sua

versão de livre distribuição, esse programa auxilia na construção do DER (no qual as entidades já são tratadas como tabelas), além de criar automaticamente comandos SQL (*Structured Query Language*) para a implementação do MR (Modelo Relacional). O banco de dados foi implementado no SGBD MySQL versão 5 (http://www.mysql.com/).

4.8.2 Inserção das predições no banco de dados relacional

Para realizar a inserção das predições realizadas no banco de dados relacional, utilizamos *scripts* (em linguagem Perl) exclusivamente desenvolvidos para essa tarefa.

4.8.2.1 Inserção das predições do DisEMBL

Para realizar todas as tarefas descritas acima, utilizamos o *script* 'parser_disembl.perl'.

Os parâmetros utilizados para esse *script* foram:

- -l = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -m = Schistosoma_mansoni (nome do organismo do qual as seqüências derivam)
- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo DisEMBL>

4.8.2.2 Inserção das predições do GlobPipe

Para inserção das predições do programa GlobPipe no banco de dados utilizamos o *script* 'parser_globpipe.perl'.

Os parâmetros utilizados para esse *script* foram:

- -l = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -m = Schistosoma_mansoni (nome do organismo do qual as seqüências derivam)

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo GlobPipe>

4.8.2.3 Inserção das predições do IUPred

Para inserção das predições do programa IUPred no banco de dados utilizamos o *script* 'parser_iupred.perl'.

Os parâmetros utilizados para esse *script* foram:

- -l = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -m = Schistosoma_mansoni (nome do organismo do qual as seqüências derivam)
- -pr = 0.5 (probabilidade de que um aminoácido esteja em uma região desordenada)
- -f = <arquivo multi-fasta com todas as seqüências preditas>
- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo IUPred>

4.8.2.4 Inserção das predições do VSL2

Para inserção das predições do VSL2 no banco de dados utilizamos o *script* 'parser_vsl2b.perl'. Os parâmetros utilizados para esse *script* foram:

- -l = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -m = Schistosoma_mansoni (nome do organismo do qual as seqüências derivam)
- -f = <arquivo multi-fasta com todas as seqüências preditas>
- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>

- -i = <nome do arquivo texto contendo as predições feitas pelo VSL2>

4.8.2.5 Inserção das predições do Phobius

Para inserção das predições do Phobius no banco de dados utilizamos o *script* 'parser_phobius.perl'.

Os parâmetros utilizados para esse *script* foram:

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo Phobius>

4.8.2.6 Inserção das predições do Pepstats

Para inserção das predições do programa Pepstats (que é parte do pacote EMBOSS) no banco de dados utilizamos o *script* 'parser_pepstats.perl'.

Os parâmetros utilizados para esse *script* foram:

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo Pepstats>

4.8.2.7 Inserção das predições do TargetP

Para inserção das predições do programa TargetP no banco de dados utilizamos o *script* 'parser_targetp.perl'.

Os parâmetros utilizados para esse *script* foram:

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo TargetP>

4.8.2.8 Inserção das predições de anotação funcional

Para inserção da anotação funcional realizada por busca de similaridade de seqüências contra o banco de dados Gene Onthology, utilizamos o *script* 'parser_functional_annotation.perl' anexo.

Os parâmetros utilizados para esse *script* foram:

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo gerado pelo *parser* do Dr. José Marcos Ribeiro (Blaster)>

4.8.2.9 Pseudocódigo dos *parsers*

4.8.2.9.1 DisEMBL

O DisEMBL aceita como entrada um arquivo no formato fasta, contendo um cabeçalho identificado pelo caractere '>', e logo abaixo dessa linha de cabeçalho, a seqüência da proteína. Após a identificação das regiões de desordem ao longo da extensão da seqüência, o DisEMBL cria uma cópia do arquivo fasta fornecido como entrada, e acrescenta as respectivas coordenadas no próprio cabeçalho do arquivo, além do método de predição (COILS / HOTLOOPS / REM465). Assim, o processo de extração do resultado do DisEMBL se resume a identificação de todos os cabeçalhos, e nesses cabeçalhos, a identificação do método de predição e das coordenadas, que aparecem no formato 'a-b, a-b, a-b'. Onde 'a' representa a coordenada inicial e 'b' a coordenada final de uma região desordenada. Exemplo de uma linha de cabeçalho gerada pelo DisEMBL:

```
> 2AHR_A_COILS 1-8, 20-29, 45-56, 64-84, 98-111, 141-166, 181-189,
                                    202-226
```

Nessa linha, o id da seqüência é '2AHR_A', 'COILS' é o método de predição, e em seguida as coordenadas de cada região desestruturada.

A seleção dos cabeçalhos foi feita de maneira simples utilizando-se o módulo 'Bio::SeqIO' do pacote Bioperl (http://www.bioperl.org/wiki/Main_Page), uma coleção

de módulos Perl que facilitam o desenvolvimento de *scripts* para aplicativos de bioinformática.

Utilizamos o método 'description' para obter o conteúdo do cabeçalho de cada seqüência, e a partir daí, utilizando expressões regulares, selecionamos o método de predição e as coordenadas de cada uma das regiões de desordem. A extração de dados básicos referentes à seqüência foi feita utilizando-se outros métodos do módulo Bio::SeqIO, tais como 'seq' e 'length'. Esse *parser* recebe obrigatoriamente quatro parâmetros. São eles:

- tamanho mínimo da região desestruturada
- organismo ao qual a seqüência pertence
- nome do banco de dados no qual os resultados serão inseridos
- nome do arquivo de entrada (resultado da execução do DisEMBL)

A conexão com o SGBD MySQL é feita utilizando-se o módulo Mysql da linguagem Perl. A captura de parâmetros passados na linha de comando do shell é feita utilizando-se os módulos ' Getopt::Long' e ' IO::File'.

4.8.2.9.2 GlobPIPE

A exemplo do DisEMBL, as coordenadas das predições realizadas pelo GlobPipe também são incluídas na linha de cabeçalho do arquivo fasta.

Uma característica interessante do GlobPIPE é que ele também faz predição de domínios globulares. Por esse motivo, há uma distinção no cabeçalho entre as coordenadas de domínios globulares e regiões desordenadas. A identificação da palavra 'Disorder' é a diferença fundamental desse *parser* praquele descrito no item anterior, pois todas as coordenadas que aparecem após a palavra 'Disorder' referem-se a regiões desestruturadas.

Exemplo de uma linha de cabeçalho gerada pelo GlobPIPE:

```
>Smp_176370 ||29660.m000203|conserved hypothetical
protein|Schistosoma mansoni|chr unknown01|||Auto|GlobDoms:2-252, 357-475,
496-634, 654-822|Disorder:1-10, 253-267, 339-356, 476-495, 550-554, 605-
610, 635-653, 679-686, 776-784, 816-824
```

Os procedimentos realizados por esse *parser* são exatamente os mesmos

descritos para o *script* desenvolvido para o preditor DisEMBL (descrito no item anterior).

4.8.2.9.3 IUPred

O IUPred aceita como entrada um arquivo fasta (contendo somente uma seqüência), mas gera um arquivo de saída completamente diferente, ao contrario dos outros dois preditores descritos anteriormente.

O fato de aceitar como entrada um arquivo com somente uma seqüência, implica na criação de um arquivo de saída para cada seqüência. Surge ai a necessidade de se utilizar um *script* bash, para que esse *parser* seja executado diversas vezes. O arquivo de saída produzido pelo IUPred apresenta o seguinte formato:

```
# IUPred
# Copyright (c) Zsuzsanna Dosztanyi, 2005
#
# Z. Dosztanyi, V. Csizmok, P. Tompa and I. Simon
# J. Mol. Biol. (2005) 347, 827-839.
#
#
# Prediction output
# Smp_158090
    1 M     0.4220
    2 I     0.5685
    3 H     0.5182
    4 I     0.5704
    5 L     0.6426
    6 D     0.7392
    7 G     0.8529
    8 P     0.7460
    9 D     0.3599
   10 G     0.2988
```

Todas as linhas que se iniciam com o caractere '#' são irrelevantes para nossas análises, com exceção da última. Esta trás o nome da seqüência que foi avaliada. Assim, o processo de extração de informação desse arquivo se inicia no *script* bash, com a remoção das 8 primeiras linhas, deixando somente informação relevante no arquivo. Abaixo da linha que contém o nome da seqüência, estão

listados cada um dos resíduos da seqüência, e as suas respectivas probabilidades de se encontrar em uma região desordenada.

A identificação das regiões de desordem estrutural e suas coordenadas se da pela presença de resíduos consecutivos, com um valor de probabilidade acima do estabelecido como mínimo (valor passado como parâmetro no momento da execução do *parser*, por padrão usa-se 0.5).

Para identificar tais regiões, o *parser* lê o arquivo linha a linha, e utilizando expressões regulares, identifica a coordenada de cada resíduo, e sua respectiva probabilidade. Caso a probabilidade seja maior ou igual ao mínimo estabelecido, uma região de desordem é criada, e então o procedimento se repete. Caso o resíduo seguinte também possua probabilidade aceitável, a região é estendida. A região se estende até que se encontre um resíduo com probabilidade menor do que o mínimo estabelecido ou até que o arquivo termine. Após o fechamento de uma região de desordem por um dos motivos descritos anteriormente, o *parser* avalia se a região possui o tamanho mínimo especificado. Caso o arquivo ainda não tenha terminado após o fechamento da região de desordem, todo o processo se repete, até o fim do arquivo.

Como o IUPred não gera um arquivo fasta como saída, informações básicas sobre a seqüência para inserção na tabela IUP não podem ser obtidas pelo *parser*. Por esse motivo, o *parser* exige um parâmetro adicional em relação aos *parsers* descritos anteriormente. O arquivo fasto com o proteoma predito é fornecido, e é desse arquivo que o *parser* retira as informações básicas, utilizando métodos do módulo Bio::SeqIO.

Esse *parser* recebe obrigatoriamente seis parâmetros. São eles:

- tamanho mínimo da região desestruturada
- organismo ao qual a sequencia pertence
- menor valor aceitável de probabilidade de um resíduo estar em uma região desordenada
- arquivo multi-fasta do proteoma predito do organismo em questão
- nome do banco de dados no qual os resultados serão inseridos
- nome do arquivo de entrada (resultado da execução do IUPred)

A conexão com o SGBD MySQL é feita utilizando-se o módulo Mysql da linguagem Perl. A captura de parâmetros passados na linha de comando do shell é

feita utilizando-se os módulos ' Getopt::Long' e ' IO::File'.

Baseando-nos em dados da literatura (Feng, Zhang *et al.*, 2006), estabelecemos como valor mínimo de probabilidade 0,5 e como tamanho mínimo da região desestruturada quarenta resíduos.

4.8.2.9.4 VSL2B

A exemplo do IUPred, o VSL2B também só aceita uma seqüência por arquivo, e isso exige a utilização de um *script* bash pra execução do *parser* pra todos os arquivos de predição gerados.

O VSL2B tem uma peculiaridade em relação aos outros preditores. Não admite a presença de uma linha de cabeçalho no arquivo de entrada. Por essa razão o arquivo de saída gerado (arquivo com as predições) não conterá referência a nenhuma seqüência. Essa referência tem que ser feita pelo nome do arquivo de saída.

Portanto, a primeira atividade do *parser* do VSL2B é a obtenção do nome da seqüência pra qual foram realizadas predições. Utilizando expressões regulares à extensão do nome do arquivo de saída do VSL2B é removida, e o nome é então obtido.

As coordenadas das regiões de desordem preditas são obtidas do relatório da predição, que apresenta o seguinte formato:

```
VSL2 Predictor of Intrinsically Disordered Regions
Center for Information Science and Technology
Temple University, Philadelphia, PA

Predicted Disordered Regions:
20-136
183-187

Prediction Scores:
=======================================
NO.     RES.    PREDICTION      DISORDER
---------------------------------------
1     M     0,406404    .
2     L     0,358619    .
3     V     0,333387    .
```

Apenas as linhas destacadas em alaranjado contêm informação de interesse. Os valores representam as coordenadas iniciais e finais de cada região de desordem estrutural predita. Lendo o arquivo linha-a-linha e utilizando expressões regulares, as linhas de interesse (iniciadas por números seguidas de um hífen) são identificadas pelo *parser*. Novamente fazendo uso de expressões regulares, os números que representam coordenadas iniciais e finais são separados e o *script* realiza a inserção no banco de dados utilizando o comando 'INSERT' da linguagem SQL.

Após a identificar todas as linhas de interesse e realizar as inserções, o *script* finaliza sua execução não precisando chegar até o fim do arquivo, pois toda a informação de interesse já foi analisada.

4.8.2.9.5 Phobius

O *parser* desenvolvido para o Phobius tem a função de extrair dados do seu relatório e inserir na tabela TRANSMEMBRANE. O arquivo de saída do Phobius apresenta o seguinte formato:

```
ID   Smp_031390
FT   SIGNAL         1      13
FT   DOMAIN         1       2       N-REGION
FT   DOMAIN         3       9       H-REGION
FT   DOMAIN        10      13       C-REGION
FT   DOMAIN        14     125       NON CYTOPLASMIC
//
ID   Smp_116550
FT   DOMAIN         1     106       NON CYTOPLASMIC
//
ID   Smp_145040
FT   DOMAIN         1      43       CYTOPLASMIC
FT   TRANSMEM      44      69
FT   DOMAIN        70      74       NON CYTOPLASMIC
FT   TRANSMEM      75      97
FT   DOMAIN        98     108       CYTOPLASMIC
FT   TRANSMEM     109     128
FT   DOMAIN       129     260       NON CYTOPLASMIC
FT   TRANSMEM     261     283
FT   DOMAIN       284     303       CYTOPLASMIC
```

No arquivo de saída o identificador 'ID' precede o nome da seqüência analisada. Essa expressão é utilizada pelo *parser* para identificar a linha que contém o nome da seqüência. Esse nome será utilizado para fazer referência à seqüência da tabela IUP, sendo, portanto chave estrangeira na tabela TRANSMEMBRANE. Abaixo dessa linha, estão descritas as diversas características preditas pelo Phobius. Lendo o arquivo linha-a-linha e utilizando expressões regulares, o *parser* identifica cada uma das características, suas coordenadas e seus atributos e realiza a inserção no banco de dados utilizando o comando 'INSERT' da linguagem SQL. Os caracteres '//' são utilizados para identificar a separação entre as predições das diversas proteínas. Em resumo, são extraídos do relatório todos os campos marcados em laranja.

Esse *parser* recebe obrigatoriamente dois parâmetros. São eles:

- nome do banco de dados no qual os resultados serão inseridos
- nome do arquivo de entrada (resultado da execução do Phobius)

A conexão com o SGBD MySQL é feita utilizando-se o módulo Mysql da linguagem Perl. A captura de parâmetros passados na linha de comando do shell é feita utilizando-se os módulos ' Getopt::Long' e ' IO::File'.

Não existem parâmetros de sensibilidade para serem considerados, tais como tamanho mínimo de uma região predita ou probabilidades, tais decisões são tomadas internamente pelo preditor.

4.8.2.9.6 Pepstats

O *parser* desenvolvido para o Pepstats tem a função de extrair dados do seu relatório e inserir na tabela IUP. O arquivo de saída do Pepstats apresenta o seguinte formato:

```
PEPSTATS of Smp_000150 from 1 to 730
Molecular weight = 82115.35         Residues = 730
Average Residue Weight  = 112.487   Charge   = 11.0
Isoelectric Point = 7.9619
A280 Molar Extinction Coefficient  = 66140
A280 Extinction Coefficient 1mg/ml = 0.81
Improbability of expression in inclusion bodies = 0.904
```

Residue	Number	Mole%	DayhoffStat
A = Ala	30	4.110	0.478
B = Asx	0	0.000	0.000
C = Cys	18	2.466	0.850
D = Asp	48	6.575	1.196
E = Glu	42	5.753	0.959

Property	Residues	Number	Mole%
Tiny	(A+C+G+S+T)	209	28.630
Small	(A+B+C+D+G+N+P+S+T+V)	393	53.836
Aliphatic	(A+I+L+V)	159	21.781
Aromatic	(F+H+W+Y)	64	8.767
Non-polar	(A+C+F+G+I+L+M+P+V+W+Y)	339	46.438
Polar	(D+E+H+K+N+Q+R+S+T+Z)	391	53.562
Charged	(B+D+E+H+K+R+Z)	198	27.123
Basic	(H+K+R)	108	14.795
Acidic	(B+D+E+Z)	90	12.329

No arquivo de saída o identificador 'PEPSTATS of' precede o nome da seqüência analisada. Abaixo dessa linha, estão descritas as diversas características calculadas pelo Pepstats. Lendo o arquivo linha-a-linha e utilizando expressões regulares, o *parser* identifica cada uma das características e seus atributos, e realiza a atualização da tabela IUP no banco de dados, utilizando o comando 'UPDATE' da linguagem SQL. Em resumo, são extraídos do relatório todos os campos marcados em laranja.

Esse *parser* recebe obrigatoriamente dois parâmetros. São eles:

- nome do banco de dados no qual os resultados serão inseridos
- nome do arquivo de entrada (resultado da execução do Pepstats)

A conexão com o SGBD MySQL é feita utilizando-se o módulo Mysql da linguagem Perl. A captura de parâmetros passados na linha de comando do shell é feita utilizando-se os módulos ' Getopt::Long' e ' IO::File'.

Não existem parâmetros de sensibilidade para serem considerados, tais como tamanho mínimo de uma região predita ou probabilidades, tais decisões são tomadas internamente pelo programa.

4.8.2.9.7 TargetP

O *parser* desenvolvido para o TargetP tem a função de extrair dados do seu relatório e inserir na tabela IUP.

O TargetP aceita como entrada um arquivo fasta (contendo somente uma seqüência), e por esse motivo, sua execução para todas as seqüências do proteoma predito é realizada através de um *script* bash. Isso implica na necessidade de se executar o *parser* do TargetP dentro de um laço, para todos os arquivos gerados. O laço foi executado com o comando 'for' do 'shell' bash.

O *script* bash de execução do TargetP (descrito no item 3.6), nomeia os arquivos de saída do preditor pela concatenação do ID da seqüência seguido pelo expressão '--TargetP.out'. Portanto, o primeiro passo do *parser* do TargetP é a obtenção do nome da seqüência, que é obtida através do nome arquivo de entrada fornecido como parâmetro. Os dados referentes às predições realizadas pelo programa são retirados do relatório que apresenta o seguinte formato:

```
### targetp v1.1 prediction results ###################################

Number of query sequences:  1

Cleavage site predictions included.

Using NON-PLANT networks.


Name                  Len            mTP     SP  other  Loc  RC  TPlen

--------------------------------------------------------------------

Smp_181440.1          248          0.301  0.636  0.022   S    4     25

--------------------------------------------------------------------

cutoff                             0.000  0.000  0.000
```

Lendo o arquivo linha-a-linha e utilizando expressões regulares, o *parser* identifica cada uma das características e seus atributos: a) *score* da predição de localização mitocondrial; b) *score* da predição de peptídeo sinal; c) *score* da predição de outras localizações (incertas para o preditor); d) localização sub-celular

mais provável; e) *score* da localização mais provável; e f) tamanho do peptídeo sinal predito. Após obtidos todos os valores, o *script* realiza a atualização da tabela IUP no banco de dados, utilizando o comando 'UPDATE' da linguagem SQL. Em resumo, são extraídos do relatório todos os campos marcados em laranja. Esse *parser* recebe obrigatoriamente dois parâmetros. São eles:

- nome do banco de dados no qual os resultados serão inseridos
- nome do arquivo de entrada (resultado da execução do TargetP)

A conexão com o SGBD MySQL é feita utilizando-se o módulo Mysql da linguagem Perl. A captura de parâmetros passados na linha de comando do shell é feita utilizando-se os módulos ' Getopt::Long' e ' IO::File'.

Não existem parâmetros de sensibilidade para serem considerados, tais como tamanho mínimo de uma região predita ou probabilidades, tais decisões são tomadas internamente pelo programa.

4.8.2.9.8 Anotação funcional

A anotação funcional das IUPs foi realizada com a utilização de algoritmos de busca por similaridade de seqüências. As seqüências das proteínas de *S. mansoni* que foram preditas como IUPs foram alinhadas contra as seqüências de referencia do banco de dados *Gene Ontology* (GO). Um *parser* desenvolvido pelo Dr. José Marcos Ribeiro (batizado de ‘Blaster’) foi utilizado para gerar um arquivo de saída contendo a relação entre os identificadores das IUPs e os termos anotadores do GO. Para analisar o arquivo de saída gerado pelo ‘Blaster’, desenvolvemos um *parser* escrito em linguagem Perl.

O arquivo de saída gerado pelo ‘Blaster’ apresenta o seguinte formato:

```
Smp_000030.1 Caenorhabditis elegans - reproduction - embryonic development ending in birth or egg hatching - nematode larval development =hyperlink(".\links\GO\Smp_000030.1-GO.txt",0) protein binding||binding binding protein binding GO:0005515 1E-159 proteasome complex||outer membrane\-bounded periplasmic space||external encapsulating structure part external encapsulating structure part outer membrane\-bounded periplasmic space GO:0000502 1E-159 anaphase\-promoting complex\-dependent proteasomal ubiquitin\-dependent protein catabolic process||proteasomal ubiquitin\-dependent protein catabolic process||ubiquitin\-dependent protein catabolic
```

```
process||modification\-dependent protein catabolic process||proteolysis
involved in cellular protein catabolic process||cellular protein catabolic
process||protein catabolic process||biopolymer catabolic
process||macromolecule catabolic process||macromolecule metabolic
process||metabolic process binding protein binding GO:0031145 1E-
159
```

Todos os dados referentes a uma proteína vêem em uma única linha, separados por uma tabulação. Essas tabulações são utilizadas pelo *parser* para separar os campos. Com os campos devidamente separados, o *parser* realiza a atualização da tabela IUP no banco de dados, utilizando o comando 'UPDATE' da linguagem SQL. Em resumo, são extraídos do relatório todos os campos marcados em laranja.

Lendo o arquivo linha-a-linha, todas as proteínas presentes no relatório são inseridas. Esse *parser* recebe obrigatoriamente dois parâmetros. São eles:

- nome do banco de dados no qual os resultados serão inseridos
- nome do arquivo de entrada (resultado da execução do ‘parser blast’)

A conexão com o SGBD MySQL é feita utilizando-se o módulo Mysql da linguagem Perl. A captura de parâmetros passados na linha de comando do shell é feita utilizando-se os módulos ' Getopt::Long' e ' IO::File'.

Não existem parâmetros de sensibilidade para serem considerados, tais como tamanho mínimo de uma região predita ou probabilidades, tais decisões são tomadas internamente pelo programa.

4.9 Análise do desempenho de predição do *pipeline*

4.9.1 Conjunto de seqüências controle

O conjunto de seqüências controle utilizados na análise de desempenho de predição foram obtidas do banco de dados DisProt, versão 4.9 (http://www.DisProt.org/data/version_4.9/DisProt_fasta_v4.9.txt) em formato fasta.

4.9.2 Pré-processamento das seqüências controle

O pré-processamento das seqüências controle foi realizado como descrito no item 3.2 uma vez que, para viabilizar a correta comparação dos experimentos, as

seqüências controle deveriam estar sujeitas ao mesmo tratamento e condições que foram submetidas às seqüências de *S. mansoni*.

4.9.3 Predição de desordem estrutural para as seqüências controle

A predição de desordem estrutural para as seqüências controle foi realizada com os mesmos preditores e parâmetros descritos no item 3.3.2. Foram fornecidos como arquivo de entrada para os preditores, os arquivos gerados no pré-processamento, descrito na seção anterior.

4.9.4 Integração dos dados

4.9.4.1 Criação do banco de dados relacional específico para análise do desempenho de predição

A criação do banco de dados relacional, específico para a análise do desempenho de predição dos quatro preditores selecionados foi realizada como descrito no item 3.8.1. O DER gerado encontra-se descrito no anexo (Anexo 2).

4.9.4.2 Inserção das anotações do DisProt no banco de dados relacional

Para a inserção das anotações de desordem estrutural do DisProt, no banco de dados desenvolvido, utilizamos o *script* 'parser_DisProt_fasta.perl'. Os parâmetros utilizados foram:

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>
- -i = <nome do arquivo texto contendo as seqüências do DISPROT no formato fasta>

4.9.4.3 Remoção da redundância de anotação do DisProt

Para remover a redundância das anotações de desordem estrutural presentes no DisProt, utilizamos o scritp 'DisProt_nr.perl'. Os parâmetros utilizados

foram:

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>

4.9.4.4 Inserção das predições no banco de dados relacional

Para inserção das predições do programa DisEMBL realizadas para as seqüências controle, utilizamos o *script* 'parser_disembl_evaluation.perl'. Esse *script* representa uma versão simplificada do *script* 'parser_disembl.perl', pois realiza a inserção de um número menor de dados na tabela. Os parâmetros utilizados para esse *script* foram:

- -l = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -d = <nome do banco de dados criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo DisEMBL>

Para inserção das predições do GlobPipe, utilizamos o 'parser_globpipe_evaluation.perl'. Os parâmetros utilizados para esse *script* foram:

- -l = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -d = <nome do banco de dados criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo GlobPipe>

Para inserção das predições do IUPred, utilizamos o *script* 'parser_iupred_evaluation.perl'.

Os parâmetros utilizados para esse *script* foram:

- -l = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -pr = 0.5 (probabilidade de que um aminoácido esteja em uma região desordenada)
- -d = <nome do banco de dados criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo IUPred>

Para inserção das predições do VSL2, utilizamos o *script*

'parser_vsl2b_evaluation.perl'.

Os parâmetros utilizados para esse *script* foram:

- -l = 40
- -d = <nome do banco de dados criado no MySQL>
- -i = <nome do arquivo texto contendo as predições feitas pelo VSL2>

4.9.5 Construção do gráfico ROC

Para a construção dos gráficos ROC, utilizamos o *script* 'ROC_analisys.perl'. Esse *script* gera um arquivo texto, com os dados necessários a construção do gráfico. Os parâmetros utilizados foram:

- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL >
- -p = <senha do usuário criado no MySQL>

Adicionalmente a saída gerada pelo *script* foi redirecionada para um arquivo texto com extensão '.csv', utilizando o operador '>' da interface Shell Linux. Esse arquivo de extensão '.csv' foi posteriormente aberto em uma planilha eletrônica para que o gráfico de então o gráfico de dispersão foi construído.

4.9.6 Seleção de uma combinação de preditores de desordem estrutural

Para selecionarmos uma combinação de preditores dentre todas as possíveis apresentadas no gráfico ROC, estabelecemos três critérios, que foram avaliados para as 5 melhores combinações.

1. Cálculo da diferença no número de TPR de cada combinação para a combinação com o maior TPR.
2. Cálculo da diferença no número de FP de cada combinação para a combinação com o menor valor de Falsos Positivos.
3. Cálculo da diferença no número de FN de cada combinação para a combinação com o maior valor de FN.

Selecionamos a combinação de preditores seguindo uma ordem de prioridades:

1. Menor valor de FP.

2. Maior valor de FN.
3. Maior valor de TPR.

Aquela combinação que satisfaz o maior número de critérios (pelo menos dois dos critérios) é selecionada. Caso uma combinação não satisfaça pelo menos dois critérios, o critério 1 (menor valor de FP) será mantido, e o critério 2 (maior FN) será reconsiderado, aceitando-se o segundo maior valor de FN.

4.10 Integrando todas as etapas – construção do *pipeline*

A execução de todas as atividades descritas é realizada de maneira automática utilizando-se o scritp 'trigger.perl'. Os parâmetros utilizados para execução automática foram:

- -i = <nome do arquivo texto contendo as seqüências preditas - proteoma>
- -ls = 100 (comprimento mínimo das seqüências)
- -met = T (seleciona seqüências que iniciam com Metionina)
- -func = T (separa as proteínas em hipotéticas e com função predita)
- -a = T (seleciona seqüências sem erros de anotação)
- -m = Schistosoma_mansoni (nome do organismo do qual as seqüências derivam)
- -d = <nome do banco de dados criado no MySQL>
- -u = <nome do usuário criado no MySQL>
- -r = <senha do usuário root do MySQL>
- -p = <senha do usuário criado no MySQL>
- -o = <prefixo para todos os arquivos gerados automaticamente pelo *pipeline*>
- -ld = 40 (comprimento mínimo (em aa) de cada região de desordem)
- -pr = 0.5 (probabilidade de que um aminoácido esteja em uma região desordenada)

Após a execução do *pipeline*, é gerado um banco de dados chamado contendo os resultados de todas as predições realizadas. Esse banco contém uma tabela com o consenso das predições de desordem estrutural dado pela combinação de preditores selecionados na etapa de análise de desempenho de predição. Além

dessa tabela, um arquivo multi-fasta também é gerado com a seqüência e a anotação de desordem estrutural para todas as proteínas identificadas como IUP.

4.11 Proteômica

4.11.1 Obtenção do extrato protéico

Partimos de uma quantidade inicial de 400µg de vermes adultos (macho/fêmea) cepa LE de *Schistosoma mansoni*, obtidos pela perfusão de camundongos previamente infectados.

Os vermes foram ressuspendidos em 1.600µL de tampão A (tampão de fosfato de sódio 10 mM, pH 7.0, NaCl 50 mM, DTT 50 mM, 1 x coquetel de inibidores de protease Roche (*Complete Mini Easypack Sample*) e ortovanadato de sódio 0.1 mM), homogeneizados utilizando pistões de homogeneização (*Kontes pellet pestle*) e centrifugados a 16.000g por 30 minutos em temperatura ambiente.

Após a centrifugação, realizamos a quantificação de proteínas no sobrenadante, e obtivemos uma concentração de 3µg/mL. Esse sobrenadante foi diluído 3 vezes no mesmo tampão A, para uma concentração final de 1µg/mL.

Esse extrato protéico foi então aquecido por 1 hora a 98°C. Após o período de aquecimento, o extrato protéico foi colocado em gelo por 15 minutos e depois centrifugado a 16.000g por 15 minutos em temperatura ambiente.

Proteínas solúveis presentes no sobrenadante foram então precipitadas com acetona (*overnight*) a -20°C. O volume de acetona acrescentado foi 4 vezes o volume do sobrenadante. Após o período de precipitação, centrifugamos a amostra a 13.000g por 15 minutos a 4°C.

Depois da centrifugação, o *pellet* foi ressuspendido em 560µL de tampão IEF (Uréia 8M, Tiouréia 2M, CHAPS 4%) sem azul de bromofenol para permitir a quantificação de proteínas pelo método de Bradford. A quantificação resultou em um extrato contendo 1µg/µL.

4.11.2 Eletroforese – Gel 1D/Gel 2D

Após a dosagem de proteínas, 1µg de proteínas de cada extrato (extrato protéico bruto e sobrenadantes descritos no item anterior) foram submetidos à separação eletroforética unidimensional em gel de poliacrilamida 12%, como descrito por Ausebel em 1995 (Ausubel, Brent *et al.*, 1995), a fim de avaliar a qualidade dos

extratos obtidos. Tendo sido constatada a sua qualidade, as proteínas foram separadas por 2DE.

Na primeira dimensão foi feita a focalização isoelétrica das proteínas. Para tal, ao volume equivalente a 15µg (para fitas de IPG de 7cm) da amostra de proteínas totais de *S. mansoni* foi acrescentado tampão de rehidratação (Uréia 8M, Tiouréia 2M, CHAPS 4%, Azul de bromofenol 0,5%, DTT 65mM e anfólitos 1x correspondente ao gradiente de pH da fita de IPG utilizada) para um volume final de 125µL. As amostras ficaram sob agitação por uma hora e foram centrifugadas a 16000 x *g*, 20-25^{o}C por 30min para retirada de material não solubilizado. Em seguida, sobrenadante foi aplicado sobre as fitas de IPG com gradiente de pH 3-10 e pH 5-8 não linear (NL), para fitas de IPG de 7cm. Após 10min, foram gotejados 750µL de óleo mineral nas fitas de IPG. As fitas foram submetidas à rehidratação e focalização isoelétrica no equipamento Protean IEF Cell (BIO-RAD) a 50µA/gel e a uma temperatura de 20^{o}C. As condições de rehidratação e focalização isoelétrica foram: rehidratação passiva por 4h, rehidratação ativa a 50V por 12h e focalização isoelétrica a 500V por 30min, 1000V por 30min, 4000V por 1h e 4000V até acumular 16000V/h. Após o término da focalização isoelétrica, as fitas foram retiradas do aparelho, tiveram o excesso de óleo mineral removido e foram congeladas a -70^{o}C.

Na segunda dimensão, as proteínas foram separadas por peso molecular em gel de poliacrilamida 12% (SDS-PAGE). As fitas de IPG de 7cm foram inicialmente equilibradas por 10min, sob agitação lenta, em 2,5mL, respectivamente, de tampão de equilíbrio I (Tris-HCl 50mM pH8.8, Uréia 6M, Glicerol 30%, SDS 2%, Azul de Bromofenol 0,5% e DTT 130mM), e posteriormente por 15min, também sob agitação lenta, em tampão de equilíbrio II (Tris-HCl 50mM pH8.8, Uréia 6M, Glicerol 30%, SDS 2%, Azul de Bromofenol 0,5% e Iodocetamida 135mM). O padrão de peso molecular (Broad Range, BIO-RAD) foi aplicado em um pedaço de papel filtro, colocado sobre o gel de poliacrilamida e selado com agarose 0,5% contendo azul de bromofenol. As fitas de IPG foram lavadas em tampão de corrida (Tris 25mM, Glicina 192mM, SDS 0,1%) antes de serem colocadas sobre os géis de poliacrilamida 12%. Da mesma forma que o padrão de peso molecular, elas foram seladas ao gel com agarose 0,5% contendo azul de bromofenol para facilitar o acompanhamento da corrida eletroforética. O sistema utilizados para a eletroforese nos géis de 7cm foi o Mini-Protean II (BIO-RAD) não refrigerado. A eletroforese dos géis de 7cm foi realizada a 50V por aproximadamente 10min e a 100V até o corante atingir a porção

inferior do gel.

Os géis foram corados utilizando-se nitrato de prata. De acordo com o protocolo para coloração por prata, os géis devem ser fixados em duas soluções: Etanol 40% e Ácido acético 10%; e Etanol 20%. Depois de fixados, os géis são lavados em água triplamente filtrada (último elemento filtrante de 0,22 mícron). A sensibilização ocorre em seguida, utilizando Diotidio de sódio 0,3 g/L. O passo seguinte é a coloração com Nitrato de prata 2 g/L. Em seguida os géis são novamente lavados com água trilplamente filtrada. A revelação dos géis ocorre em seguida utilizando apenas uma solução: Carbonato de Potássio 30 g/L, Formaldeído 250µL e Tiossulfeto de sódio 10mg/L. O bloqueio da revelação utiliza Tris 40 g/L e Ácido acético 2%.

Os géis corados foram escaneados utilizando um densitrômetro GS-800 (BIO-RAD), a uma resolução de 300dpi, e em seguida foram armazenados a 4^{o}C em solução de etanol 20%.

5 RESULTADOS

5.1 Pré-processamento do proteoma predito de S*. Mansoni*

Uma das etapas essências dos *pipelines* computacionais de análise de seqüências está relacionado ao emprego de vários critérios de corte visando gerar um conjunto de dados robusto em termos de consistência e conteúdo da informação necessária para o processamento.

Dentro dessa premissa, a primeira etapa da metodologia desenvolvida nesse trabalho esteve centrada no pré-processamento das seqüências protéicas preditas para o genoma de *S. mansoni*.

Os passos desse pré-processamento são aditivos, ou seja, o passo dois só foi realizado após a realização do passo um, e assim sucessivamente. Dessa forma, garante-se que o conjunto final de proteínas selecionadas satisfaça todos os critérios estabelecidos.

Assim sendo, a filtragem e seleção do conjunto de proteínas inicial estiveram centradas nos seguintes critérios:

a) Proteínas menores que 100 aminoácidos foram descartas uma vez que geram predições estatisticamente não confiáveis;

b) Proteínas que não apresentavam uma metionina como resíduo inicial foram descartadas uma vez que podem representar erro de predição gênica ou uma proteína truncada; e

c) Seqüências que apresentam caracteres ilegais, ou seja, caracteres não definidos na tabela de códigos da IUPAC - '*International Union of Pure and Applied Chemistry*' (http://www.dna.affrc.go.jp/misc/MPsrch/InfoIUPAC.html) também foram descartadas.

Adicionalmente, na fase de pré-processamento, é realizada uma separação de proteínas em dois grupos: proteínas com alguma função predita por similaridade de seqüência e proteínas hipotéticas (sem produto predito).

Partindo do proteoma predito para *S. mansoni* versão 4.0e com 13.175 proteínas, após a realização pré-processamento chegamos aos resultados explicitados no gráfico 1.

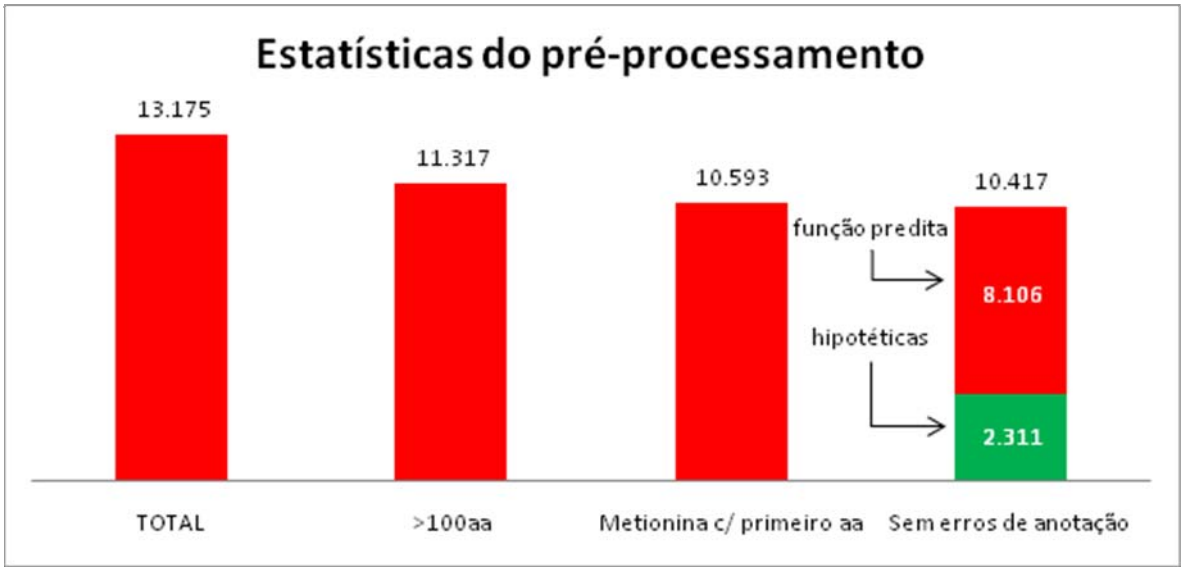


Gráfico 1: Etapa inicial do *pipeline* desenvolvido: Pré-processamento computacional do proteoma predito de *S. mansoni*. Um total geral de 13.175 proteínas passou pela etapa de pré-processamento. Aproximadamente 85,9% (11.317/13.175) foram classificadas como maiores que 100 aminoácidos sendo que 93.6% desse total (10.593/11.317) apresentavam uma metionina como resíduo inicial. 176 seqüências forma classificadas como contendo erros de anotação e foram removidas. De acordo com anotação original, desse conjunto 2.311 são proteínas hipotéticas e 8.106 apresentam alguma função descrita.

O aproveitamento final foi de 79.06% (10.147/13.175), e esse conjunto de seqüências foi utilizado nas etapas subseqüentes de caracterização.

5.2 Análise do desempenho de predição do *pipeline*

Como comentamos na parte introdutória desse trabalho um grande desafio na identificação computacional de IUPs esta relacionada à inexistência de um consenso na definição de desordem estrutural protéica.

Assim sendo, cada algoritmo preditor implementa sua própria definição fato esse que gera certa disparidade nas predições das diferentes abordagens além da impossibilidade de comparação de resultados.

Como exemplo disso podemos citar o programa DisEMBL (item 1.3.4.8) que implementa em suas redes neurais três definições. A primeira baseada no fator de temperatura dos carbonos α, a segunda na grande flexibilidade das regiões de coils e a terceira na falta de coordenadas para alguns resíduos nos relatórios PDB (Linding, Jensen *et al.*, 2003).

Por outro lado, o programa IUPred (ítem 1.3.4.10), baseia-se na interação físico-química entre os aminoácidos vizinhos para determinar a estabilidade estrutural de um segmento protéico (Dosztányi, Csizmók *et al.*, 2005) enquanto que

programa GlobPlot (item 1.3.4.9), como descrito no item 1.3.4.9, aposta no cálculo de propensões (Linding, Russell *et al.*, 2003).

Dentro do contexto acima exposto, uma das propostas desse trabalho, esteve centrada na predição integrativa de desordem estrutural que é sugerida como uma metodologia capaz de avaliar os resultados de diferentes preditores e gerar uma melhor predição.

Tendo esse objetivo como meta, utilizamos uma abordagem conhecida como gráficos ROC (*Receiver Operating Characteristics*) para avaliar a sensibilidade e especificidade dos diferentes preditores e identificar a melhor combinação entre eles. Para tanto utilizamos um conjunto controle contendo seqüências experimentalmente identificadas como desestruturadas. A seguir apresentamos as etapas desse processo.

5.2.1 Seqüências controle

Parte fundamental do processo de avaliação de sensibilidade e especificidade relaciona-se a obtenção de um bom conjunto de seqüências controle. Assim sendo, selecionamos um conjunto de proteínas desestruturadas validadas por diferentes metodologias experimentais.

Para a obtenção dessas seqüências utilizamos o banco de dados DisProt (http://www.DisProt.org/) que teve sua versão 4.9 (com 523 seqüências) instalada em nossos servidores.

O referido banco foi então submetido ao mesmo pré-processamento aplicado ao proteoma predito de *S. mansoni*. No final, 18,7% (98/523) das seqüências foram descartadas (46 menores que 100 aminoácidos e 52 seqüências por não apresentarem Metionina como resíduo inicial).

5.2.2 Predições de desordem estrutural para as seqüências controle

Após a construção do arquivo contendo as 425 seqüências controle utilizamos os preditores descritos no item 3.3.1 para realização das predições de desordem estrutural. A parametrização de cada um desses algoritmos seguiu os critérios definidos para análise das seqüências de *S. mansoni.*

5.2.3 Integração das predições para análise

Após a finalização do processo de predição estrutural para o conjunto de dados controle obtivemos um total geral de 2.532 predições. Essas predições refletem os resultados produzidos por quatro diferentes preditores (DisEMBL, GlobPIPE, IUPred e VSL2B) empregando um total de seis definições de desordem estrutural diferentes.

Tendo como objetivo a integração dessa informação e posterior análise visando avaliar a sensibilidade e especificidade de cada metodologia, um banco de dados relacional foi desenvolvido (vide descrição do banco no item abaixo).

5.2.3.1 Banco de dados relacional

Em essência, um banco de dados relacional (SGBD- Sistema Gerenciador de Banco de Dados) organiza e armazena dados em forma de tabelas. O gerenciamento dos dados é feito de acordo com o modelo relacional. Como descrito no item anterior, desenvolvemos um banco de dados relacional para integrar as predições realizadas, e possibilitar as análises comparativas de desempenho entre os preditores.

Esse banco de dados em essência armazena a informação relativa aos diferentes preditores utilizados, as coordenadas de localização de cada região de desordem estrutural predita dentro da seqüência protéica e conseqüentemente ao número de predições contido em cada proteína.

No total para o MR (Modelo Relacional) desenvolvido foram idealizadas quatro tabelas. A tabela central desse MR é chamada ‘dataset’. Nessa tabela ficam armazenados os identificadores únicos de cada proteína que serão referenciados nas outras três tabelas que compões o banco de dados. Além dos identificadores, essa tabela armazena também o tamanho (em número de aminoácidos) dessas seqüências.

As informações relativas às regiões de desordem de cada uma das proteínas referenciadas na tabela ‘dataset’ ficam armazenadas na tabela ‘disorder’. Essa tabela armazena a coordenada inicial e a coordenada final de cada região de desordem, o tamanho dessas regiões (em número de aminoácidos) e a referência para o identificador da proteína a qual pertencem. Aos dados de localização de desordem estrutural associamos o termo anotador: desordem estrutural.

A tabela ‘prediction’ por sua vez tem como função armazenar a coordenada inicial e a coordenada final de cada predição realizada. Além disso, armazena também o nome da metodologia utilizada em cada predição, o tamanho de cada uma dessas regiões e o identificador da seqüência a qual pertencem.

A tabela ‘disorder’, que em essência contem toda a anotação de desordem estrutural predita ao conjunto de dados, é uma tabela redundante. Isso acontece porque podem existir, para um mesmo trecho de seqüência protéica, evidências experimentais geradas por diferentes metodologias.

Essa redundância não é tratada no DisProt, mas representa uma condição essencial para nossas análises. Assim sendo, para armazenar a anotação de desordem estrutural sem redundância, o MR construído conta com uma tabela chamada ‘disorder_nr’. As estruturas das tabelas ‘disorder’ e ‘disorder_nr’ são idênticas com exceção do fato, como mencionado anteriormente, dessa última albergar dados não redundantes.

A seguir descrevemos a metodologia desenvolvida para remoção da redundância do DisProt assim como os resultados obtidos.

5.2.3.2 Remoção da redundância da anotação do DisProt

O banco de dados “*The Database of Protein Disorder*” (DisProt) representa um dos principais bancos de dados curados dedicados ao armazenamento de informações sobre proteínas que não apresentam uma estrutura 3D definida em seus estados nativos preditos. Criado como o resultado de um esforço colaborativo entre o Centro de Biologia Computacional e Bioinformática da Escola de Medicina da Universidade de Indiana e o Centro de Ciência e Tecnologia da Informação da Universidade da Universidade de Temple nos Estados Unidos, o DisProt armazena seqüências protéicas com anotações de desordem estrutural realizadas a partir de evidências experimentais.

Uma lista e breve descrição de todos os métodos de detecção utilizados para obtenção dos dados experimentais de desordem estrutural contemplados no DisProt pode ser encontrada em http://www.DisProt.org/view_detection.php.

Assim sendo, como uma dada proteína pode ter sido alvo de diferentes metodologias para obtenção da evidencia experimental de desordem estrutural, fica evidente a razão pela qual o banco é redundante. No contexto desse trabalho, o

termo redundante (para o banco de dados DisProt) referencia a existência de mais de uma evidência experimental para um mesmo trecho de seqüência protéica.

A existência desse tipo de redundância no banco representa um problema quando utilizamos essas seqüências como controle em um estudo comparativo do desempenho de diferentes preditores de desordem estrutural.

Para podermos utilizar as seqüências do DisProt como banco de dados controle em nosso experimento *in silico*, precisamos tratar essa redundância de modo que durante o processo de analise cada trecho de seqüência protéica com evidência experimental seja considerado uma única vez.

Com o objetivo de resolver essa redundância, desenvolvemos um *script* em linguagem Perl chamado 'disprot_nr.perl'.

A estratégia aplicada por esse *script* é relativamente simples com relação ao conceito empregado. Considerando uma dada proteína, o *script* armazena as coordenadas iniciais e finais de todos os trechos da seqüência dessa proteína que possuam evidências experimentais de desordem. Similarmente a idéia empregada na construção de *contigs* (seqüências consenso obtidas do alinhamento de inúmeros fragmentos de seqüências), um vetor consenso único de coordenadas é gerado a partir dos vetores que representam as diferentes predições. Esse vetor representa o consenso de várias evidências experimentais e define uma única região na proteína cujas coordenadas podem então ser utilizadas nos cálculos (Figura 8).

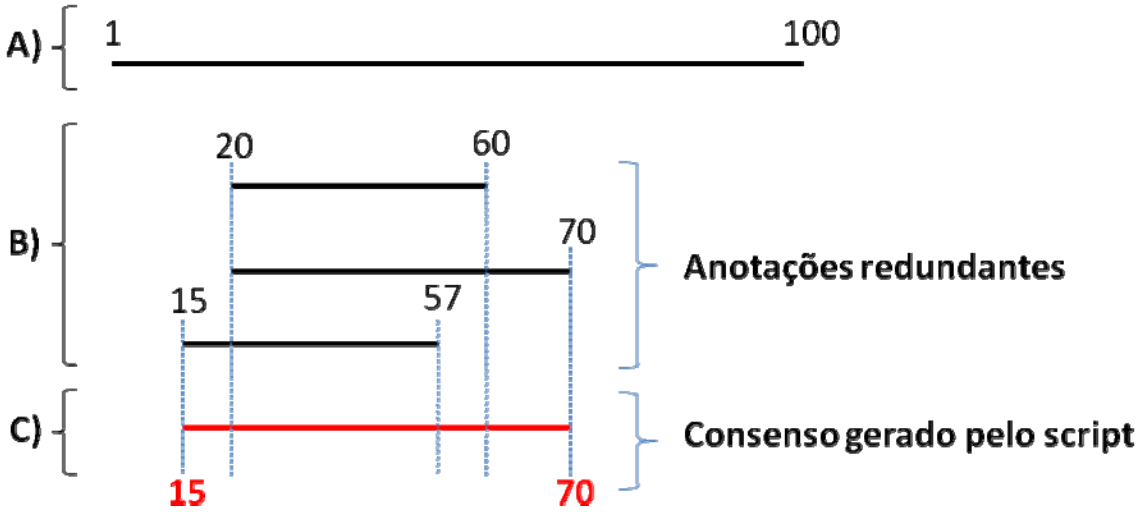


Figura 8: Estratégia do *script* `disprot_nr.perl`. Representação esquemática da estratégia empregada para remoção da redundância de predições experimentais do banco de dados DisProt. A) seqüência protéica hipotética de 100 aminoácidos; B) o conjunto de predições experimentais para um dado trecho dessa seqüência; C) o vetor consenso de predições cujas coordenadas são inseridas banco de dados.

Depois de removida a redundância, as coordenadas do vetor consenso de evidencias experimentais foram inseridas no banco de dados em uma nova tabela chamada ‘disorder_nr’. As coordenadas presentes nessa tabela foram então utilizadas nas comparações com as predições realizadas para as seqüências controle.

Na tabela 1 apresentamos os resultados obtidos durante o processo de remoção da redundância dos dados do DisProt.

Tabela 1: Remoção da Redundância do Banco de Dados DisProt versão 4.9.

	DisProt (Redundante)	**DisProt (Não Redundante)**
No. de Predições de Regiões Desordenadas	958	257
Tamanho médio (aa)	59,65	160,9

5.2.3.3 Inserção das predições no banco de dados relacional

Para realizar a inserção automática das predições de desordem no banco de dados relacional (vide item 4.2.3.1), se fez necessário o desenvolvimento de *scripts* específicos além a adaptação dos *parsers* descritos no item 3.8.2.

Como as seqüências do DisProt foram obtidas em formato fasta, desenvolvemos um *parser*, escrito em linguagem Perl chamado ‘parser_DisProt_fasta.perl’. Esse *script* extrai o identificador de cada seqüência (ID), o tamanho de cada seqüência e as coordenadas de cada uma das regiões de desordem, e posteriormente insere toda essa informação no banco de dados.

A modificação essencial dos *parsers* descritos no item 3.8.2 foi à remoção de uma série de comandos de inserção no banco de dados. Tais comandos acrescentam informações às tabelas que são fundamentais para a caracterização das seqüências no que diz respeito à desordem estrutural, mas são desnecessárias no contexto de avaliação do desempenho de predição. De maneira geral, os *parsers* foram simplificados e adaptados para reconhecerem os nomes das tabelas no banco de dados de seqüências controle.

Além do *script* ‘parser_DisProt_fasta.perl’ essas adaptações deram origem a quatro novos *scripts*: a) ‘parser_disembl_evaluation.perl’ b)

‘parser_globpipe_evaluation.perl’ c) ‘parser_iupred_evaluation.perl’ e d) ‘parser_vsl2b_evaluation.perl’.

Assim, similarmente ao descrito no item 4.2.3.1, o banco de dados desenvolvido armazena não somente o resultado das predições de desordem, mas também informação acerca da anotação de desordem estrutural das seqüências controle (seqüências do DisProt).

5.2.4 Cálculo de sensibilidade e especificidade

Como descrito na parte introdutória desse trabalho os gráficos ROC representam uma excelente abordagem que permite analisar a taxa de acertos e de erros de um algoritmo classificador.

A etapa fundamental na utilização dessa técnica é uma definição clara e inequívoca das situações que representam acertos e erros do algoritmo classificador.

Assim definimos os seguintes critérios de acerto e erro:

a) Acerto: todas as vezes em que um preditor gera um resultado, definido por uma coordenada inicial e uma coordenada final, totalmente inserido (10% de tolerância) na região predita por evidencias experimentais (TP – *True Positive* (Verdadeiro Positivo));
b) Acerto: todas as vezes em que o preditor não gera um resultado para uma região ordenada (TN – *True Negative* (Verdadeiro Negativo));
c) Erro: todas as vezes em que o preditor gera um resultado para uma região ordenada (FP – *False Positive* (Falso Positivo)); e
d) Erro: todas as vezes em que o preditor não gera um resultado para uma região predita por evidências experimentais (FN – *False Negative* (Falso Negativo)).

A figura 9 descreve esquematicamente a utilização do sistema de classificação TP, TN, FP e FN descrito acima.

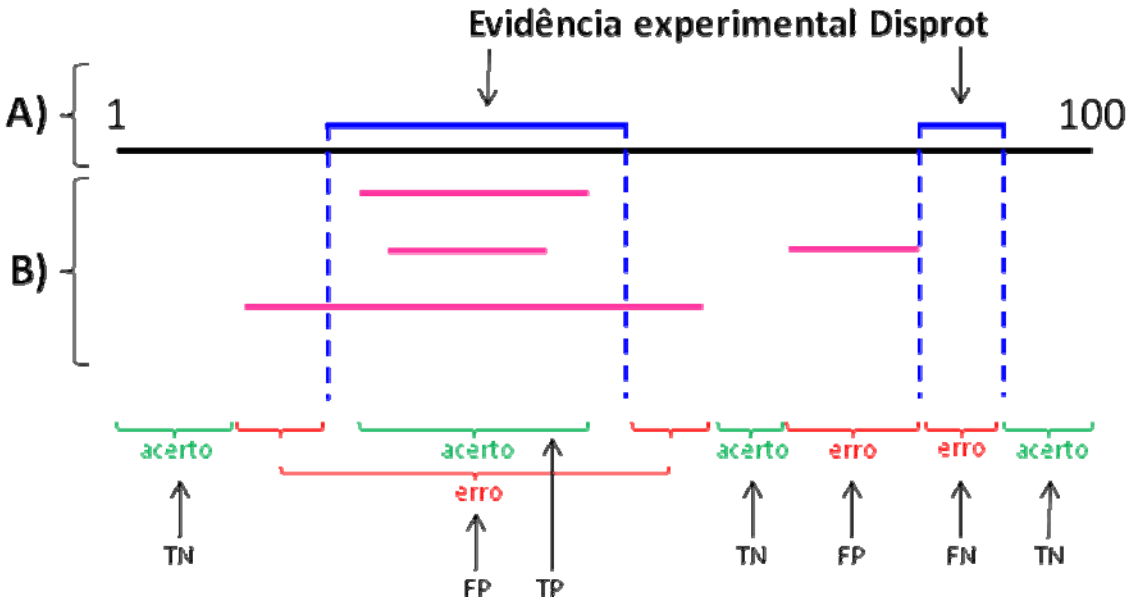


Figura 9: Representação Esquemática do Sistema de Classificação empregado. A) seqüência protéica hipotética de 100 aminoácidos; As linhas azuis representam a posição de regiões desordenadas para qual existe evidência experimental; B) predições realizadas por um preditor de desordem estrutural.

Após definidos os conceitos de Verdadeiro Positivo (TP), Falso Positivo (FP), Verdadeiro Negativo (TN) e Falso Negativo (FN) fica fácil calcular a freqüência de cada uma dessas situações nas classificações realizadas por dado preditor. No exemplo apresentado na figura 9, podemos facilmente quantificar esses classificadores (TP=1,TN=3, FP=2 e FN=1).

Como visto no item 1.5, esses quatro valores podem ser combinados para a construção de uma matriz de confusão. A matriz de confusão nada mais é do que a combinação desses quatro valores em uma tabela.

Uma vez construída a matriz de confusão, é possível se calcular a Taxa de Verdadeiros Positivos (ou TPR do inglês *True Positive Rate*) e a Taxa de Falsos Positivos (ou FPR do inglês *False Positive Rate*).

O TPR representa o número de vezes que o algoritmo realizou classificações corretas, seja julgando corretamente uma região desordenada (TP - predição de desordem) ou julgando corretamente uma região ordenada (TN - ausência de predição).

O FPR representa o número de vezes que o algoritmo realizou classificações erradas, seja julgando como desordenada uma região ordenada (FP - predição de desordem em trecho ordenado), ou julgando como ordenada uma região desordenada (FN - ausência de predição em trecho desordenado).

O gráfico ROC apresenta o valor TPR no eixo Y e o valor FPR no eixo X

estabelecendo uma relação entre a taxa de acertos e a taxa de erros do algoritmo.

Na figura 9 apresentamos a título de exemplo os erros e acertos de um algoritmo para uma dada proteína. Para extrapolarmos essa análise para todas as proteínas de um determinado conjunto de dados devemos calcular os valores TP, FP, TN e FN para cada uma das proteínas individualmente e somar esses valores de tal forma a produzir um valor acumulado para cada uma das variáveis.

Utilizando essa abordagem, uma matriz de confusão é gerada apresentando os acertos e erros do preditor na classificação de todas as regiões de desordem presentes no conjunto de proteínas.

Pela inviabilidade da análise manual do banco de seqüências controle que utilizamos nesse trabalho, desenvolvemos um *script* escrito em linguagem Perl, chamado 'ROC_analisys.perl'. Esse programa quantifica os acertos e erros de cada preditor, ou seja, o *script* calcula TP, FP, TN e FN para todos os preditores e, além disso, leva em conta as predições realizadas para todas as proteínas do conjunto de seqüências controle.

Como explicitado na figura 9, os valores da matriz de confusão são obtidos comparando dois conjuntos de coordenadas: a) as coordenadas das predições realizadas pelos diferentes algoritmos; e b) as coordenadas das regiões de desordem anotadas no banco de dados DisProt.

Como descrito no item 4.2.3.2, toda a informação necessária para essas comparações está presente no banco de dados relacional desenvolvido (item 4.2.3.1). Portanto, o *script* foi desenvolvido para obter os dados necessários à construção da matriz desse banco de dados relacional.

Após a construção de uma matriz de confusão para cada preditor, os valores TPR e FPR são calculados e podem ser apresentados em um gráfico cartesiano (gráfico ROC). Esse gráfico cartesiano é uma representação visual da relação de acertos e erros de cada uma das seis metodologias utilizadas nesse trabalho.

Como nosso interesse não era avaliar exclusivamente o desempenho das predições de cada metodologia individualmente, mas também a combinação dessas metodologias, efetuamos os cálculos de TPR e FPR para cada uma dessas combinações.

Utilizando tal estratégia pudemos construir um gráfico ROC que permitiu a identificação da combinação de metodologias com o melhor desempenho de predição (Figura 10).

Embora o princípio dos cálculos seja o mesmo, a construção da matriz de confusão para uma combinação de preditores envolve alguns passos adicionais.

Como a informação relacionada a essa comparação não está presente no banco de dados de maneira direta, o *script* 'ROC_analisys.perl' também avalia o consenso de suas predições para então construir uma matriz de confusão.

Para construir a matriz de confusão para uma combinação de cinco metodologias, o *script* avalia a existência de predições em cada uma das regiões de desordem anotada. Caso haja pelo menos uma predição válida (condizente com as coordenadas anotadas) para uma dada região anotada, um TP é computado. Mesmo que haja mais de uma predição válida para uma mesma região anotada, apenas um TP será considerado. De maneira análoga, erros também serão tratados dessa maneira, ou seja, mesmo que haja mais de uma predição de desordem para uma região ordenada, apenas um FP será considerado. A figura 10 ilustra esse processo com 5 acertos (2 TP + 3 TN) e 2 erros (2 FP).

No final do processo analítico, uma matriz de confusão específica para cada metodologia é gerada. Adicionalmente uma matriz de confusão para cada combinação de metodologias também é produzida. Cada uma dessas matrizes dará origem a um ponto no gráfico ROC.

Como parte do processamento, o *script* 'ROC_analisys.perl' também gera sem repetição todas as combinações possíveis de metodologias. O número de diferentes combinações é calculado pela fórmula descrita abaixo:

$$C = \sum_{k=1}^{n} \frac{n!}{k!(n-k)!}$$

Na fórmula descrita acima C representa o número de combinações possíveis e n é o número de metodologias sendo avaliadas. Calculando-se as combinações possíveis para seis metodologias empregadas nesse estudo, chega-se a um total de 63 combinações.

Após a criação das matrizes de confusão, os valores TPR e FPR são calculados e impressos em um arquivo texto no formato CSV (*Comma Separated Value File Format*). Esse arquivo pode ser importado em um programa de planilhas eletrônicas, e então o gráfico ROC é construído.

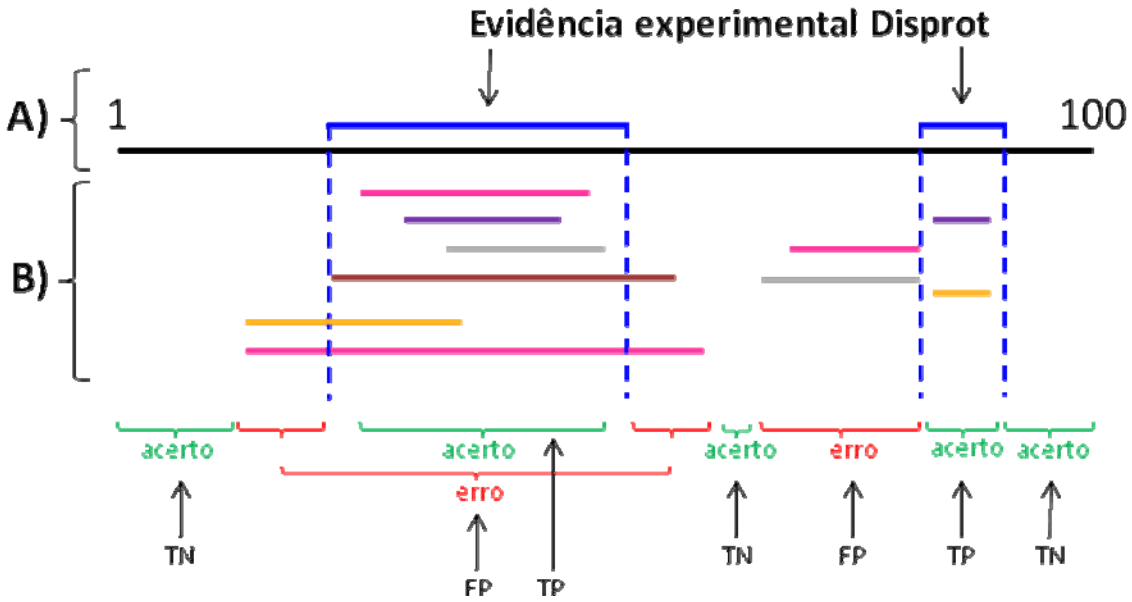


Figura 10: Consenso de predições de IUPs. A) seqüência protéica hipotética de 100 aminoácidos; As linhas azuis representam a posição de regiões desordenadas para qual existe evidência experimental; B) predições realizadas por diferentes preditores de desordem estrutural. No exemplo, são representadas cinco metodologias.

5.2.5 Gráfico ROC

O gráfico abaixo (Gráfico 2) apresenta o desempenho de cada uma das combinações de metodologias no espaço ROC. As atividades desenvolvidas para a construção desse gráfico são descritas do item 4.2.1 até o item 4.2.4.

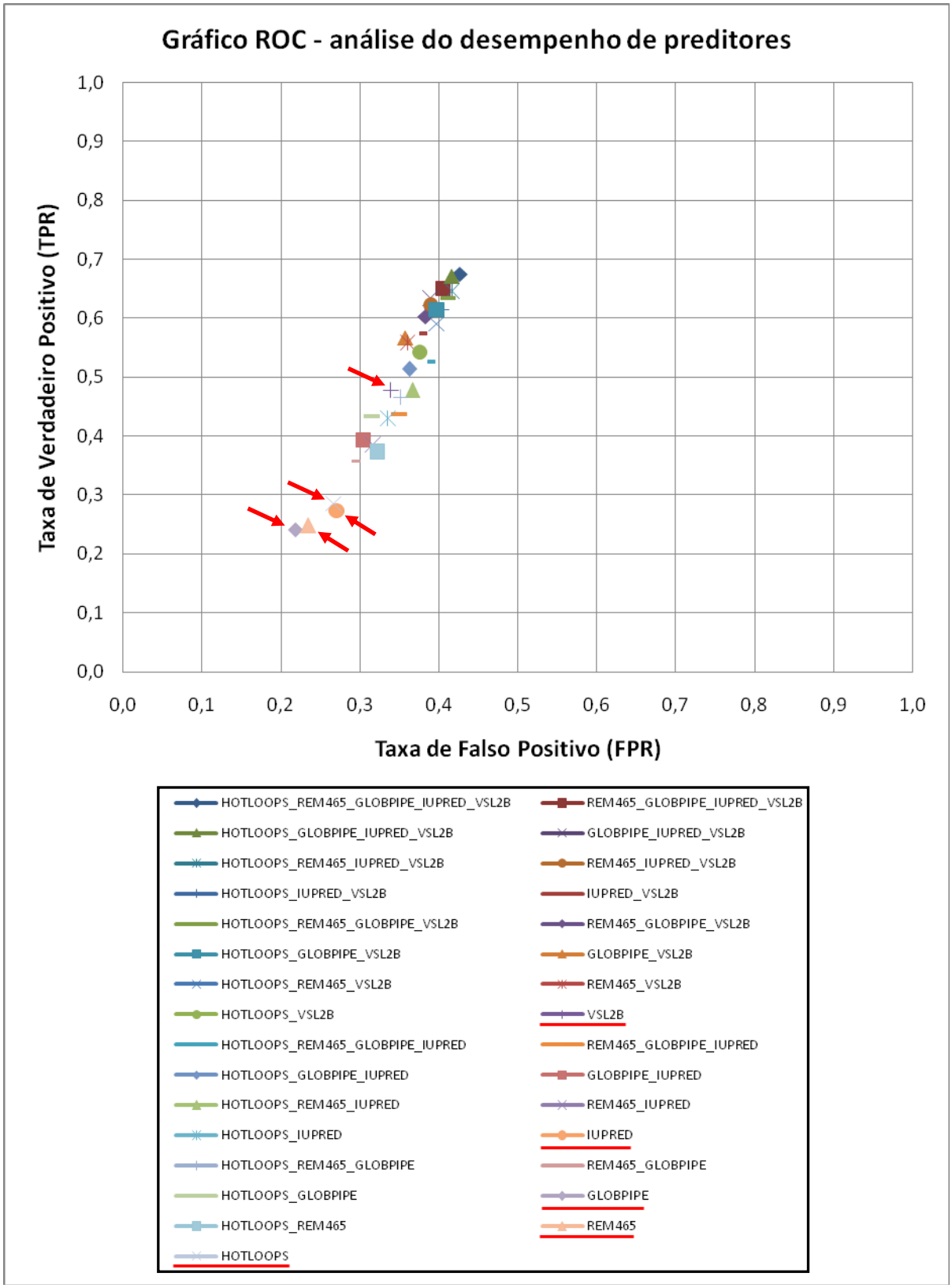


Gráfico 2: Gráfico ROC. Apresenta o desempenho de predição das 63 combinações possíveis de metodologias de predição de desordem estrutural. Os valores no eixo Y representam a Taxa de Verdadeiros Positivos (TPR). Esses valores denotam o número de vezes em que uma classificação foi feita corretamente (vide item 5.2.4). Os valores no eixo X representam a Taxa de Falsos Positivos (FPR). Esses valores denotam o número de vezes em que uma classificação foi feita erroneamente (vide item 5.2.4). Na legenda do gráfico, cada item representa uma metodologia de predição, que pode aparecer individualmente (indicados pelas setas vermelhas e grifados na legenda) ou de maneira combinada.

5.2.6 Seleção de uma combinação de preditores de desordem estrutural

Aplicando os critérios descritos no item 3.9.6 às cinco melhores combinações de preditores apresentadas no gráfico ROC da seção anterior, construímos a tabela abaixo:

COMBINAÇÃO DE METODOLOGIAS DE PREDIÇÃO	FPR	TPR	DIFERENÇA TPR	FP	DIFERENÇA FP	TP	TN	FN	DIFERENÇA FN
REM465/GLOBPIPE/IUPRED/VSL2B	0,402	0,670	**0,028**	455	**0**	167	675	82	**1**
HOTLOOPS/REM465/GLOBPIPE/VSL2B	0,408	0,666	0,032	484	29	166	701	83	0
HOTLOOPS/GLOBPIPE/IUPRED/VSL2B	0,413	0,694	0,004	513	58	173	729	76	7
HOTLOOPS/REM465/IUPRED/VSL2B	0,413	0,674	0,024	514	59	168	730	81	2
HOTLOOPS/REM465/GLOBPIPE/IUPRED/VSL2B	0,423	0,698	0,000	597	142	174	813	75	8

Tabela 2: Seleção de cinco melhores preditores no gráfico ROC. A primeira combinação de preditores apresenta o menor valor de FP e o segundo maior valor de FN, o que pode ser visto facilmente nas colunas de diferença de FP e diferença de FN.

Na tabela 2, vemos que a primeira combinação de preditores (destacada em verde) apresenta o menor valor de falsos positivos (FP) e o segundo maior valor de falsos negativos (FN). Esses valores refletem o caráter mais conservador dessa combinação de preditores.

De certa forma, ao invés de realizar uma classificação errônea, essa combinação de preditores tem uma tendência em não classificar uma RLD (Região Longa de Desordem) como desordenada. Apesar de não apresentar o maior valor de TPR (*True Positive Rate*), dentre as cinco combinações com melhor desempenho apresentadas no gráfico ROC (item 4.2.6), julgamos essa combinação conservadora como ideal para nossas análises.

5.3 Predição de desordem estrutural

Durante a etapa de predição de desordem estrutural foram analisadas 10.417 seqüências protéicas de *S. mansoni* em termos de quatro algoritmos e seis definições de desordem diferentes.

No final do processo obtivemos 53.809 predições de RLDs (Região Longa de Desordem) que vem relacionadas às diferentes definições de desordem no gráfico 3.

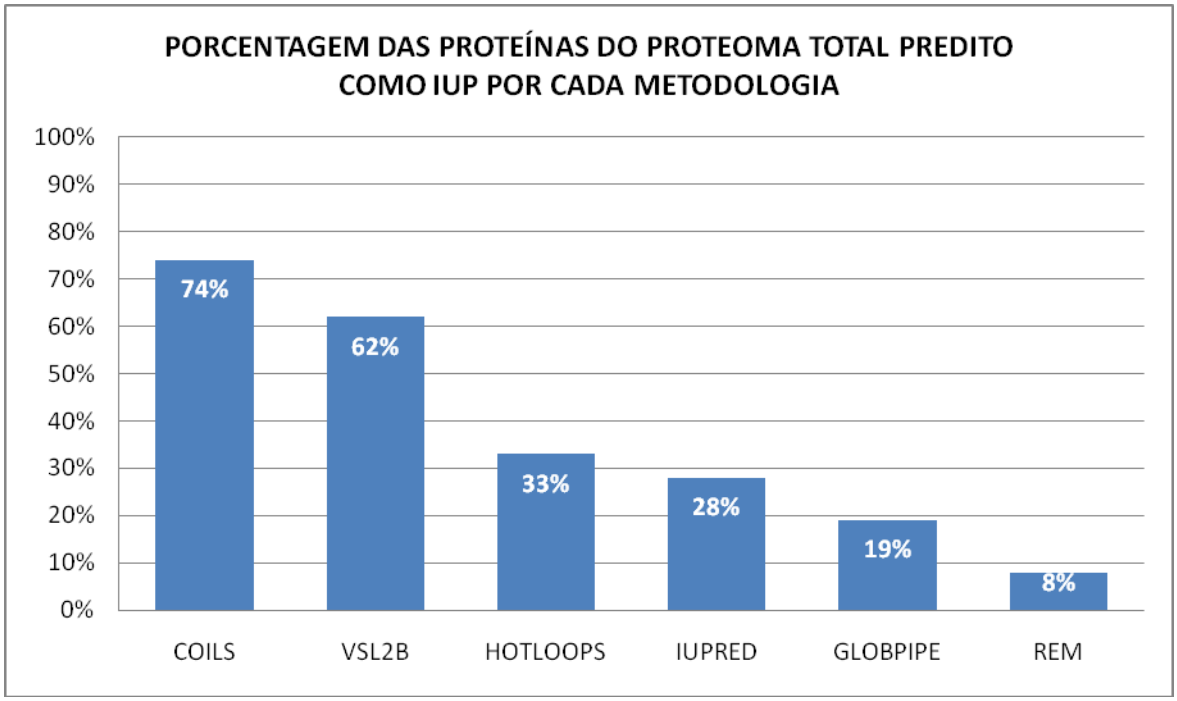


Gráfico 3: Desordem Estrutural Protéica no proteoma de *S. mansoni*. Foram analisadas 10.417 seqüências resultantes do pré-processamento. As colunas do gráfico representam cada uma das seis metodologias de predição de desordem estrutural utilizadas. São elas: COILS, VSL2B, HOTLOOPS, IUPred, GlobPipe e Remark465.

A tabela 3 contém os resultados individuais das predições de desordem estrutural realizadas por cada metodologia. A tabela apresenta também o número de proteínas identificadas com pelo menos uma RLD, o número total de RLDs preditas por cada metodologia e a extensão média das RLDs.

Tabela 3: Predição de desordem estrutural por metodologia. Porcentagens indicam a proporção das 10.417 preditas como IUP por cada metodologia.
.

	COILS	HOTLOOPS	REM465	GLOBPIPE	IUPRED	VSL2B
Proteínas com pelo menos uma RLD	7.752 (74%)	3.405 (33%)	840 (8%)	1.940 (19%)	2.930 (28%)	6.503 (62%)
Número de RLDs preditas	23.445	5.458	1.090	3.120	5.571	15.125
Extensão das RLDs (aa)	73,3	59,75	56,5	59,7	85	107,56

É importante enfatizar que a partir desse momento utilizamos a combinação de preditores de melhor desempenho (selecionada no item 4.2.7), assim sendo, 3.499 proteínas foram analisadas gerando um total de 7.373 predições.

As IUPs preditas apresentam porções de sua extensão como estruturalmente desordenadas, raramente apresentam toda a sua extensão desordenada. Por essa razão, calculamos a porcentagem de desordem apresentada por cada proteína. Essa porcentagem se define pela soma da extensão desordenada de cada proteína dividida pelo tamanho da proteína (em número de AA).

O gráfico 4 apresenta a porcentagem de IUPs que possuem de 0.4 a 100% de sua extensão desordenadas.

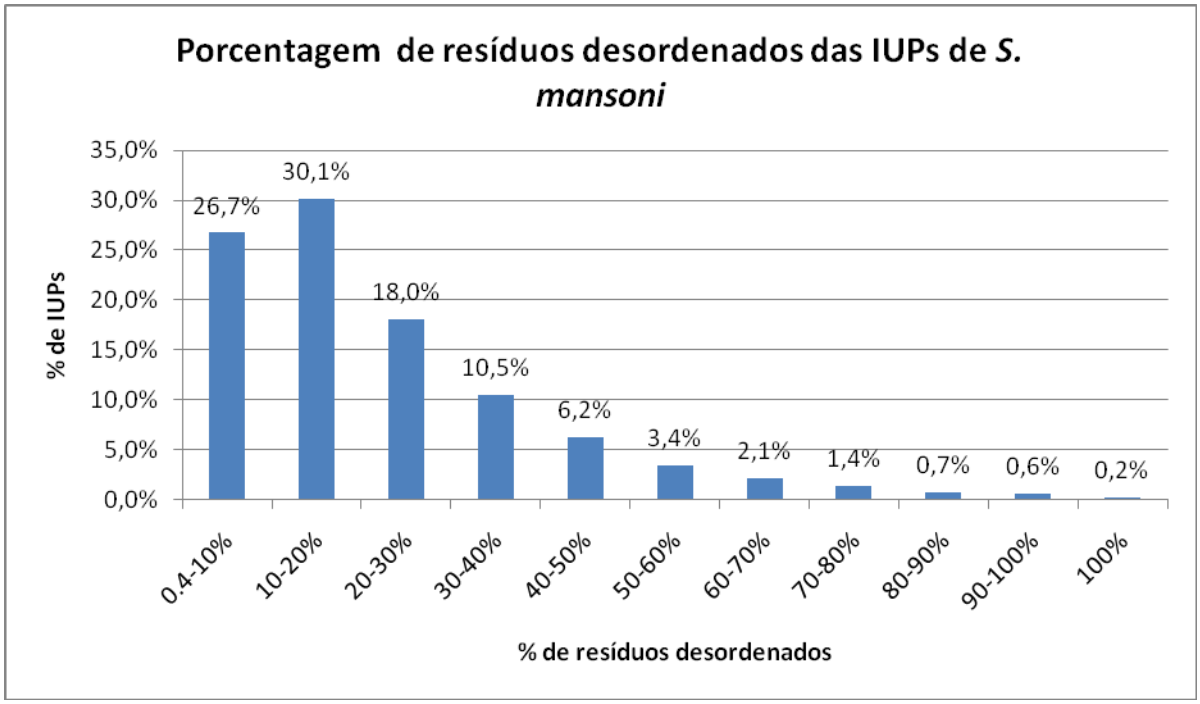


Gráfico 4: Porcentagem de resíduos desordenados das IUPs de *S. mansoni.*

Somente esse conjunto de dados (3.499 IUPs) foi considerado nas análises de caracterização subseqüentes e descritas a seguir.

5.4 Caracterização das IUPs

Como discutido na parte introdutória desse trabalho (item 1.3.4) optamos pela utilização de preditores de desordem estrutural protéica que empregam diferentes metodologias para identificação de trechos protéicos apresentando desordem estrutural. Em função da definição distinta de desordem estrutural adotada por cada uma dessas abordagens, resultados incongruentes podem surgir.

Um ponto de inconsistência que merece destaque está relacionado ao

estabelecimento das coordenadas das regiões limítrofes do trecho de desordem estrutural predito. De fato, esse item é tão relevante que Ferron e colaboradores (Ferron, Longhi *et al.*, 2006) sugerem que a delimitação real de uma região de desordem só possa ser feita acuradamente quando auxiliada pela comparação de coordenadas de outras predições, tais como de domínios ou motivos funcionais.

Por essa razão, realizamos a predição de diversas outras características não diretamente associadas à desordem estrutural que auxiliariam a caracterização das seqüências analisadas e também na definição dos limites das regiões desordenadas.

Durante essa etapa de anotação funcional, caracterizamos 3.499 IUPs para o genoma de *S. mansoni* em termos da presença de domínios funcionais e transmembrana e localização sub-celular. Realizamos também a predição de características físico-químicas e uma anotação funcional com base nos termos anotadores de função do Gene Ontology (GO), para as IUPs identificadas e também para as proteínas globulares do proteoma predito de *S. mansoni*.

Para caracterização dos domínios funcionais realizamos buscas por similaridade de seqüencia via RPS-BLAST (Reverse PSI-BLAST) contra uma cópia local do banco de dados “Conserved Domain Database” (CDD) http://www.ncbi.nlm.nih.gov/ Structure/ cdd/cdd_help.shtml.

Estabelecemos um valor de corte de *Expect (E) Value* de $1.0x10^{-6}$ e dentro desse critério pudemos atribuir domínios a 70% (2.456/3.499) ficando o restante das proteínas 30% (1.043/3.499) classificadas como hipotéticas. De um total de 7.373 regiões de desordem estrutural, 5,5% (410/7.373) tiveram suas coordenadas limítrofes ajustadas pela utilização dessa abordagem.

Visando refinar a anotação estrutural e funcional agregada ao proteoma predito de *S. mansoni* realizamos várias outras análises.

Com relação aos domínios transmembrana, pela utilização do programa Phobius foram mapeadas 268 predições referentes a 74 IUPs. O gráfico 5 apresenta a porcentagem de proteínas com apenas um domínio transmembrana predito e as respectivas porcentagens das proteínas multi-domínios transmembrana.

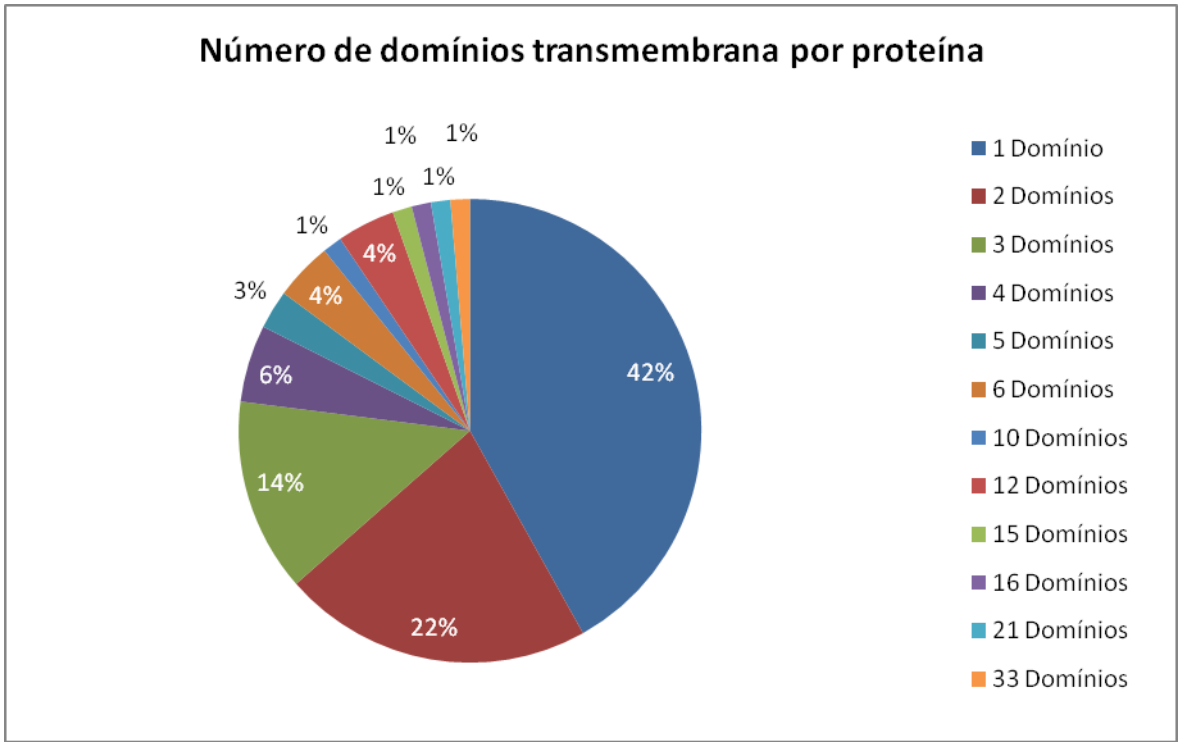


Gráfico 5: Número de domínios transmembrana nas IUPs. Porcentagem de IUPs que contém somente um domínio transmembrana, dois domínios e assim por diante. Números segundo predições do programa Phobius. No total, foram realizadas 268 predições de domínios transmembrana.

Para predição da localização sub-celular das IUPs preditas utilizamos o programa TargetP (item 3.6). Estabelecemos um valor de corte de RC (*Reliability Class*) de 2 (escala de 1 a 5, sendo o valor 1 o mais confiável) e dentro desse critério foram mapeadas 4.816 predições de localização sub-celular. A distribuição das IUPs segundo sua localização é apresentada no gráfico 6.

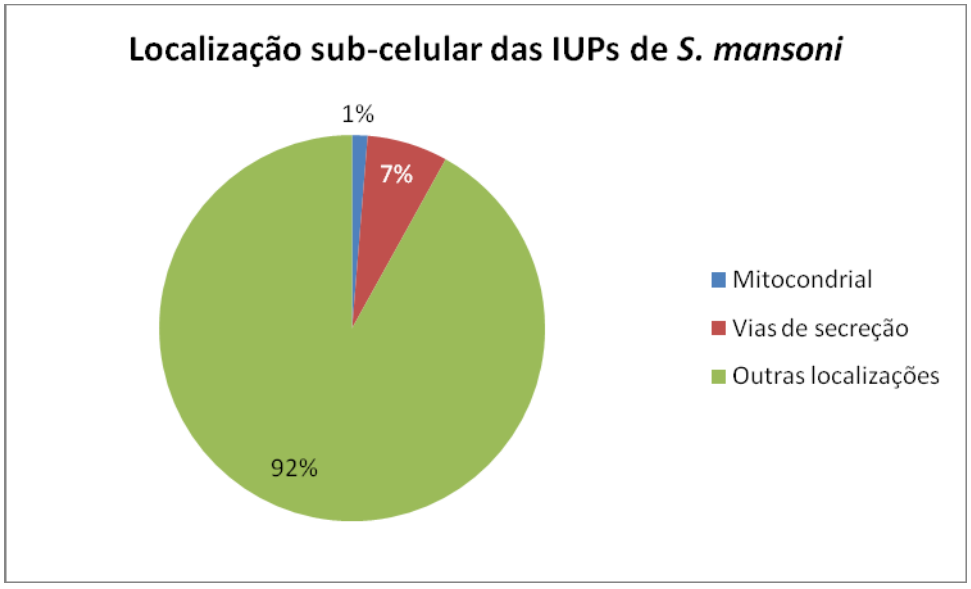


Gráfico 6: Distribuição das 3.499 IUPs segundo sua localização sub-celular. Números segundo predições do programa TargetP. No total foram realizadas 4.816 predições de localização sub-celula. RC (Reliability Class) <= 2.

Fizemos uma análise da distribuição do comprimento (em número de aa) das IUPs com relação ao comprimento das proteínas globulares (aquelas que não apresentaram predição de desordem pela combinação de metodologia selecionada). As faixas de comprimentos avaliadas são apresentadas no gráfico abaixo (Gráfico 7).

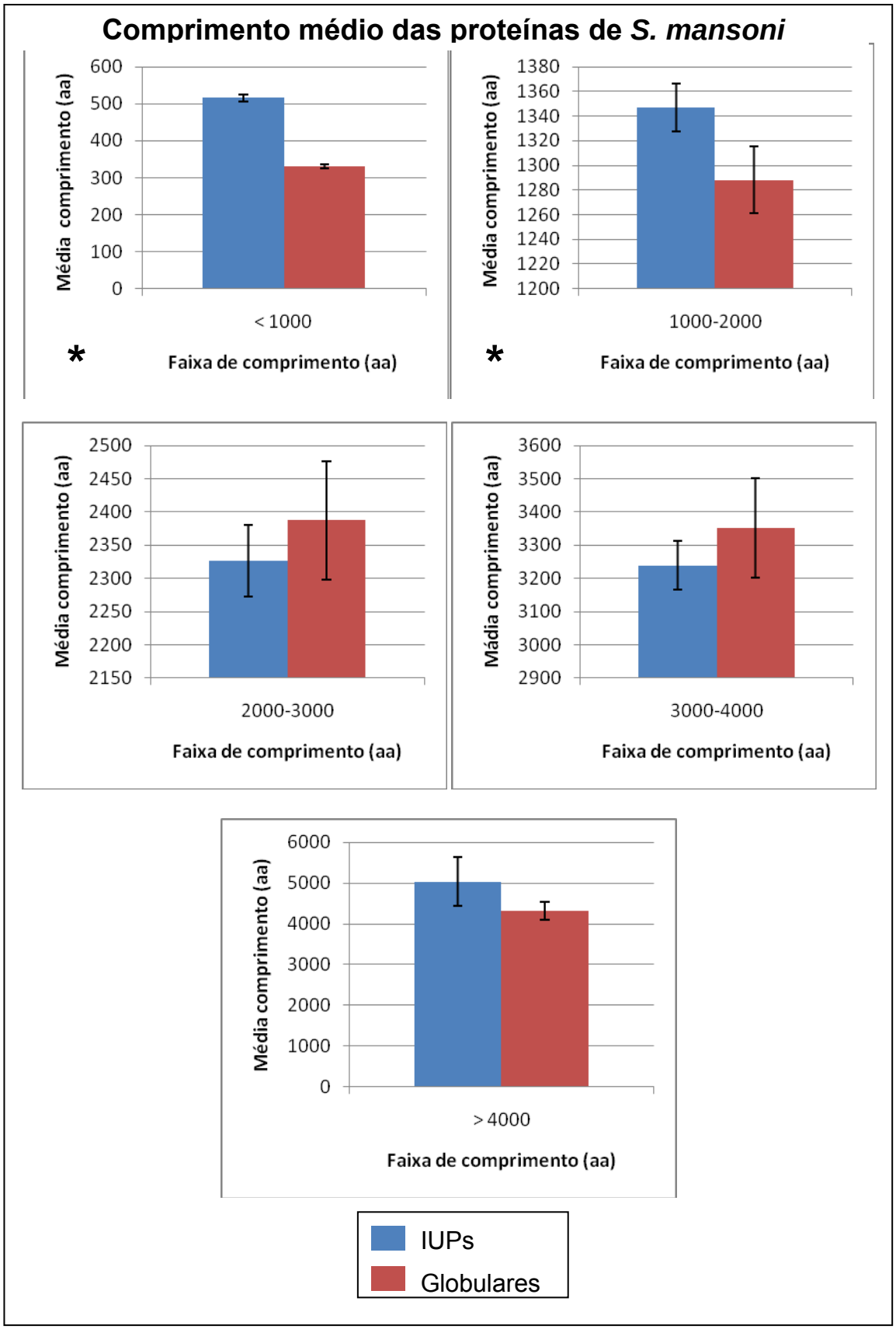


Gráfico 7: Comprimento médio das proteínas de *S. mansoni*. (*) Faixas de comprimento de proteínas pras quais há diferença estatisticamente significantiva entre IUP e proteínas globulares, com p-valor < 0,05 para teste T de Student.

Finalizamos a anotação estrutural das 3.499 IUPs de *S. mansoni* realizando uma caracterização de propriedades físico-químicas e comparando-as com as propriedades físico-químicas das proteínas globulares de *S. mansoni*. A distribuição do ponto isoelétrico, de carga elétrica e peso molecular, calculadas pelo programa PepStats (item 3.5) são apresentadas nos gráficos abaixo (Gráfico 8 até o Gráfico 10).

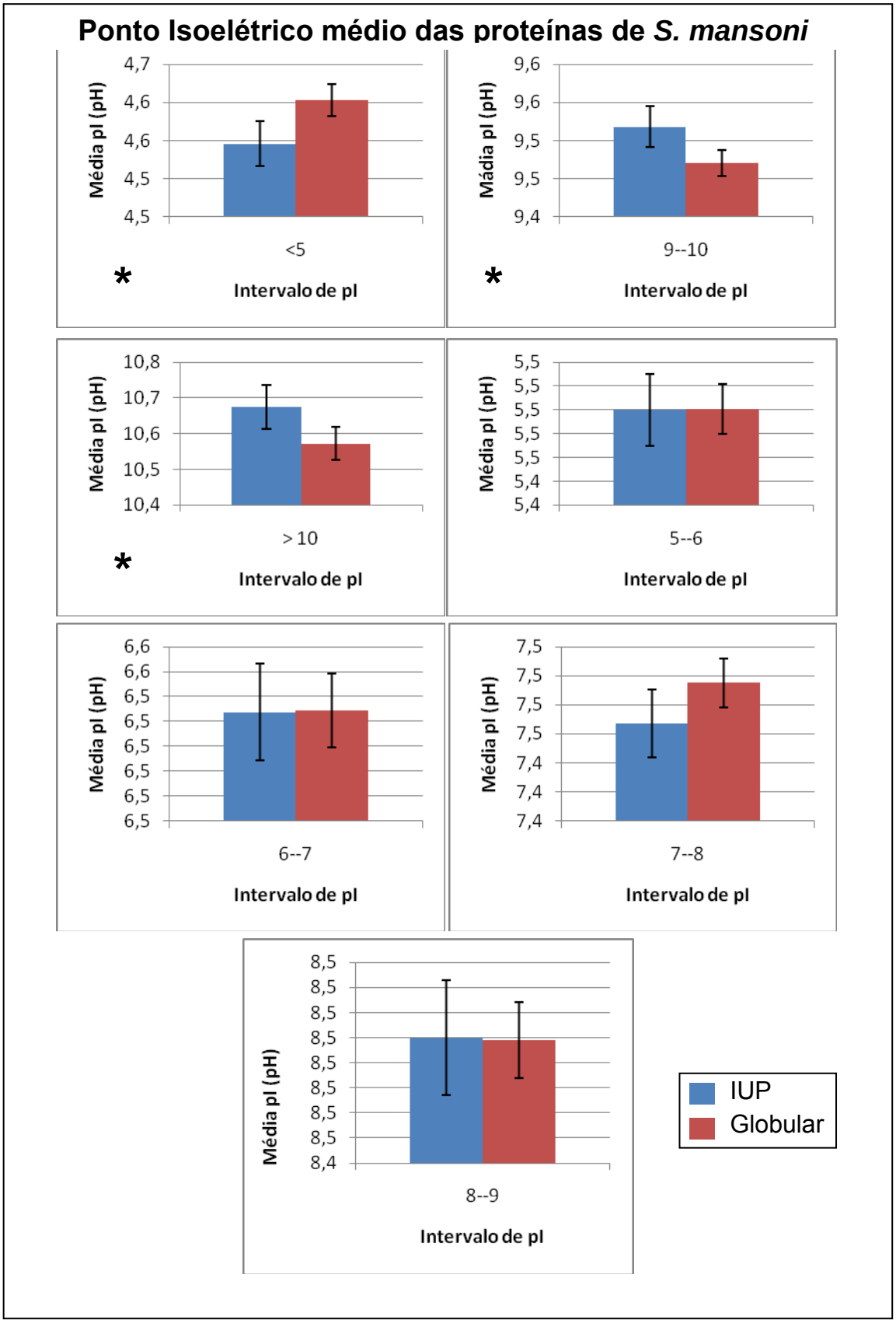


Gráfico 8: Ponto isoelétrico médio das proteínas de *S. mansoni*. (*) Diferença nas médias entre IUPs e proteínas globulares estatisticamente significativa, com p-valor < 0,05 para o Teste T de Student.

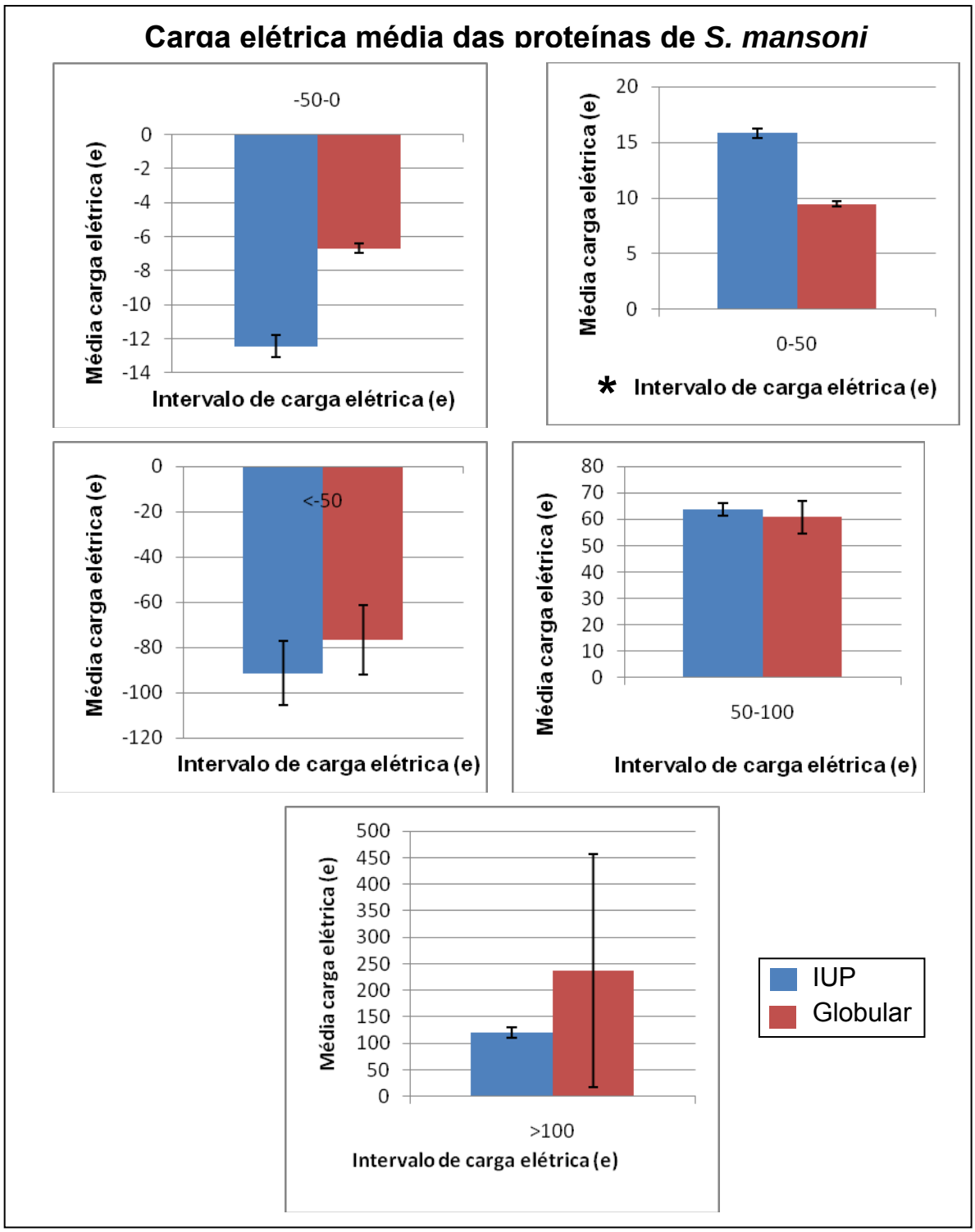


Gráfico 9: Carga elétrica média das proteínas de *S. mansoni*. (*) Diferença nas médias entre IUPs e proteínas globulares estatisticamente significativa, com p-valor < 0,05 para o Teste T de Student.

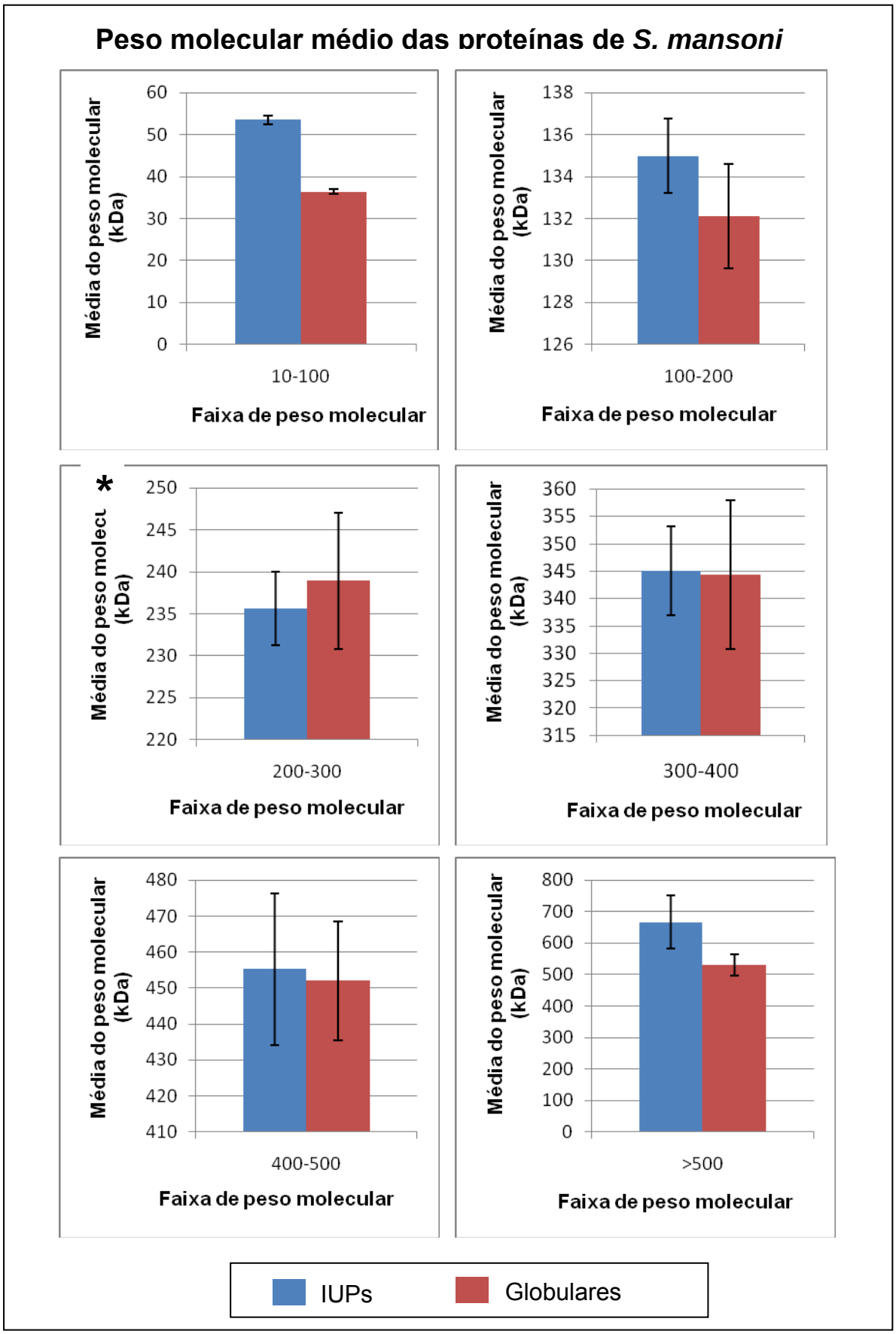


Gráfico 10: Peso molecular médio das proteínas de *S. mansoni*. (*) Diferença na média entre IUPs e proteínas globulares estatisticamente significativa, com p-valor < 0,05 para o Teste T de Student.

Concluindo a etapa de caracterização das proteínas, realizamos a anotação funcional das 3.499 seqüências de IUPs e das 6.918 seqüências de proteínas globulares segundo os termos anotadores do *Gene Ontology* (GO). Os termos anotadores foram associados às seqüências através de similaridade de seqüência, utilizando o algoritmo BLAST (Altschul, Gish *et al.*, 1990) versus o banco de dados do *Gene Ontology*. A classificação funcional para as IUPs é apresentada do gráfico 11 até o gráfico 13. A classificação funcional para as proteínas globulares é apresentada do gráfico 14 até o gráfico 16.

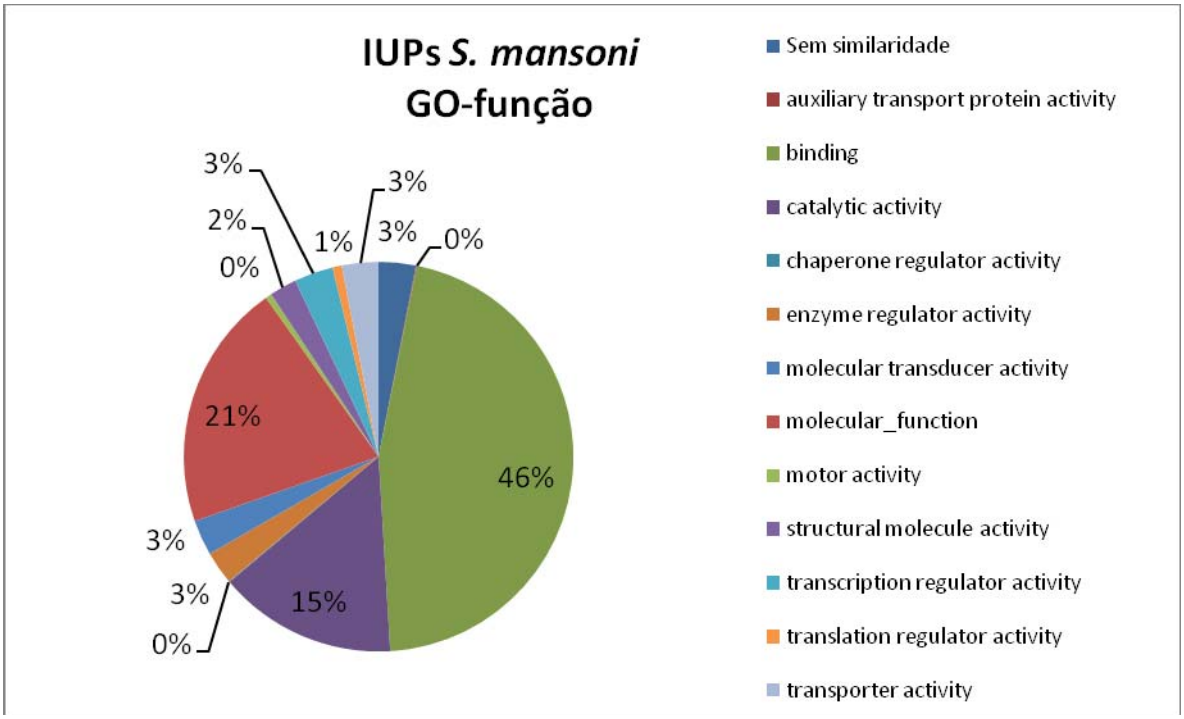


Gráfico 11: Anotação funcional das IUPs, categoria função - GO. Funções associadas às IUPs utilizando Blast contra o banco de dados do vocabulário de classificação funcional GO.

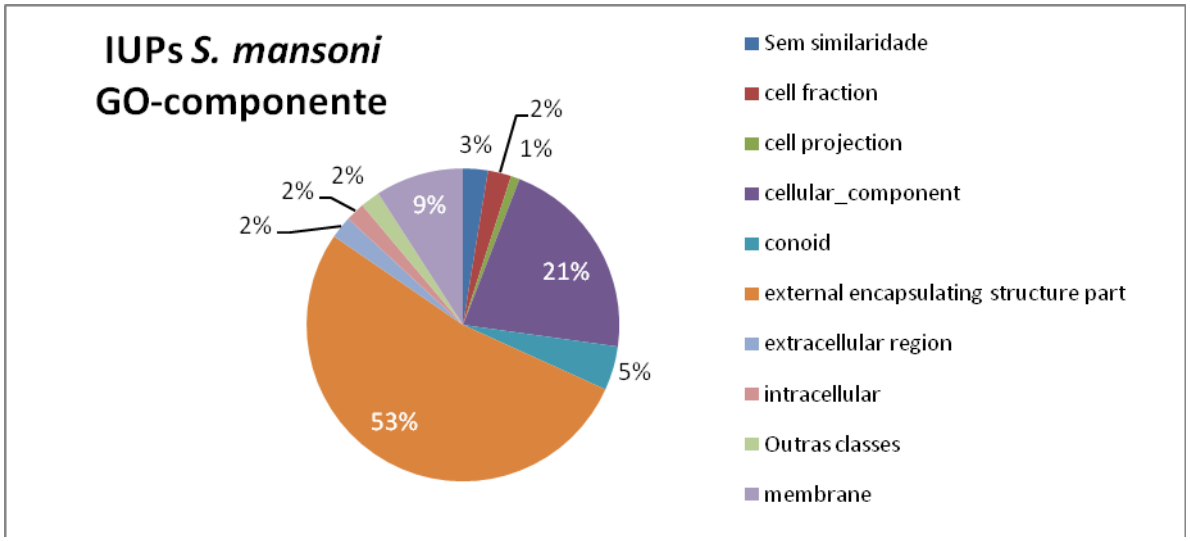


Gráfico 12: Anotação funcional das IUPs, categoria componente - GO. Componentes celulares associados às IUPs utilizando Blast contra o banco de dados do vocabulário de classificação funcional GO.

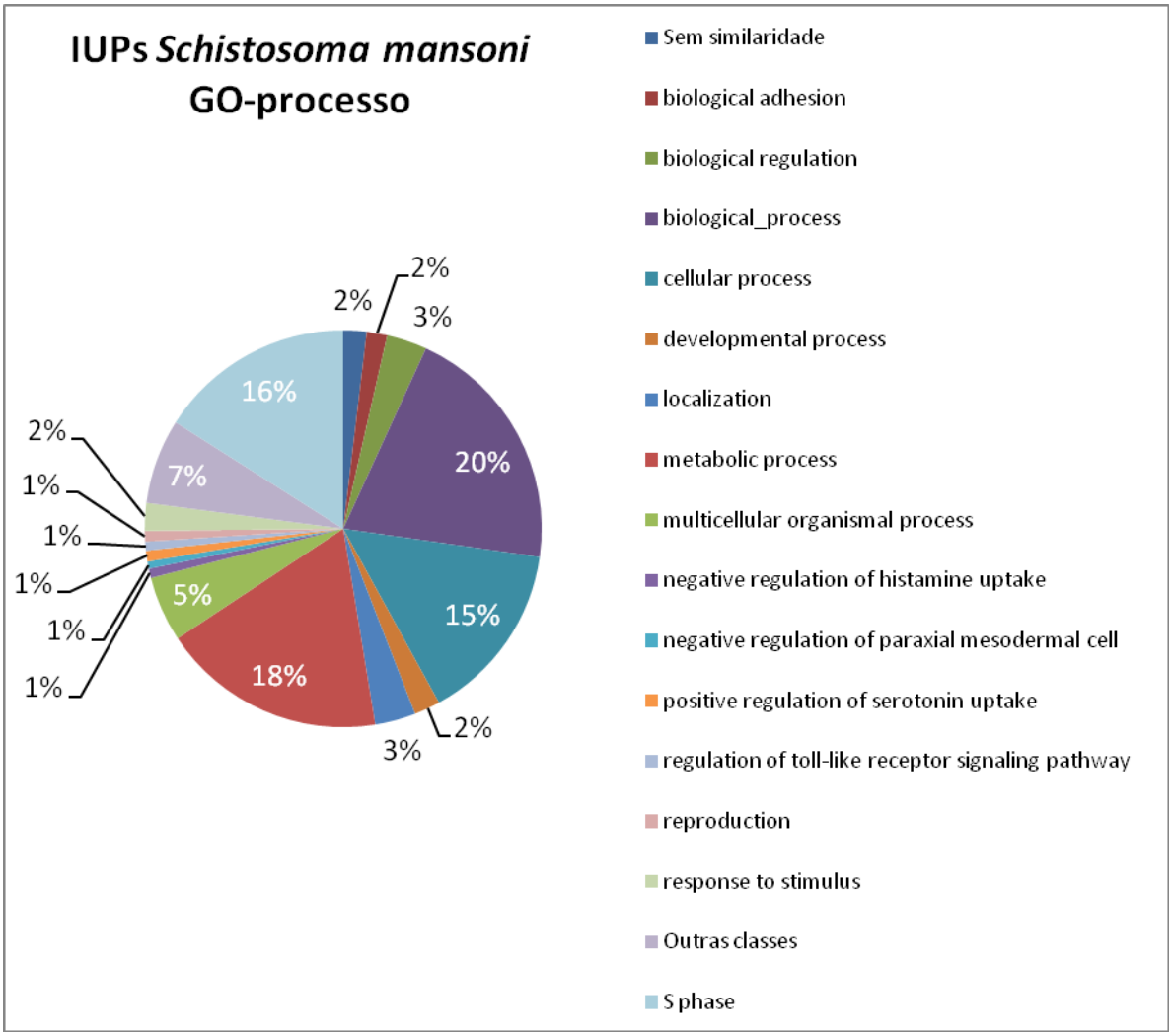


Gráfico 13: Anotação funcional das IUPs, categoria processo - GO. Processos associadas às IUPs utilizando Blast contra o banco de dados do vocabulário de classificação funcional GO.

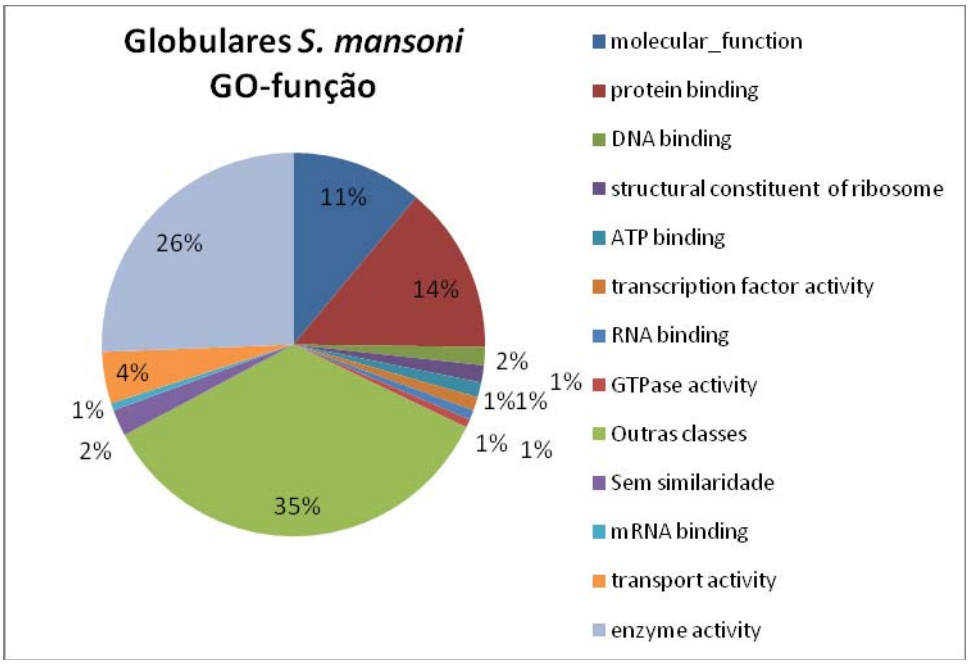


Gráfico 14: Anotação funcional das proteínas globulares, categoria função - GO. Funções associadas às proteínas globulares utilizando Blast contra o banco de dados do vocabulário de classificação funcional GO.

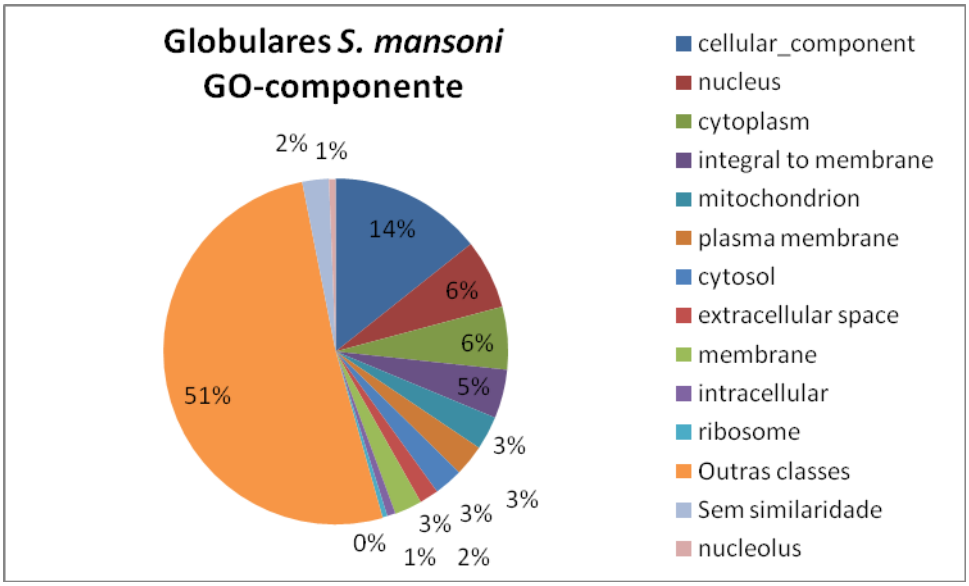


Gráfico 15: Anotação funcional das proteínas globulares, categoria componente - GO. Componentes celulares associados às proteínas globulares utilizando Blast contra o banco de dados do vocabulário de classificação funcional GO.

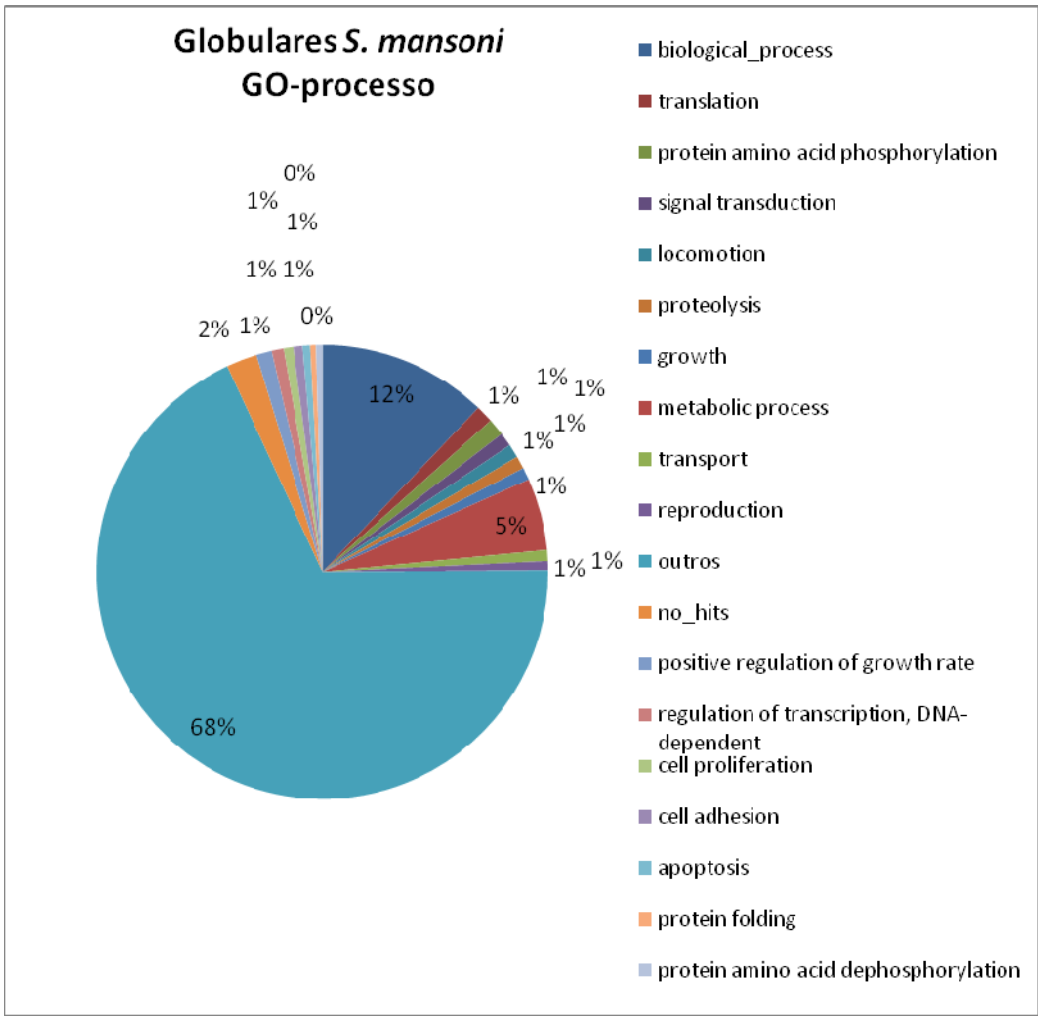


Gráfico 16: Anotação funcional das proteínas globulares, categoria processo - GO. Processos associadas às proteínas globulares utilizando Blast contra o banco de dados do vocabulário de classificação funcional GO.

5.5 Banco de dados relacional

O MER (Modelo de Entidades e Relacionamentos) é um modelo abstrato cuja finalidade é descrever de maneira conceitual os dados a serem utilizados em um sistema de informações.

Durante o desenvolvimento dessa dissertação foi idealizado um MER visando o armazenamento de dados da anotação estrutural e funcional referentes à seqüência de cada proteína, além dos resultados de todas as predições realizadas. O MER desenvolvido foi implementado na forma de um MR (Modelo Relacional) no MySQL.

O projeto contempla o armazenamento somente das proteínas identificadas como IUPs. Aquelas presentes no conjunto de dados inicial, porém que não contem nenhuma região desordenada não são armazenadas. O MR implementado conta com cinco entidades. Cada uma dessas entidades foi mapeada para uma tabela no MR (Modelo Relacional).

Todos os atributos únicos para cada seqüência aparecem na tabela IUP. Atributos que aparecem mais de uma vez relacionados à mesma seqüência, estão alocados em tabelas distintas.

A tabela IUP armazena dados básicos da seqüência, tais como: id (identificação única e exclusiva da seqüência), organismo, seqüência, descrição, tamanho da seqüência e função predita (sim/não). Além desses, características como, anotação funcional da seqüência (baseada no vocabulário de classificação funcional GO (*Gene Ontology*)), localização celular (baseada nas predições do programa TargetP) e propriedades físico-químicas (predições baseadas no programa PepStats) também são armazenadas.

A tabela DISORDER armazena todos os dados referentes a cada uma das predições de regiões de desordem para as seqüências da tabela IUP.

A tabela DISORDER_nr armazena todos os dados referentes ao consenso das predições realizadas pelas combinações de preditores.

A tabela TRANSMEMBRANE armazena dados sobre domínios transmembrana preditos paras as seqüências da tabela IUP pelo programa Phobius (Käll, Krogh *et al.*, 2004).

Por último, a tabela STATISTICS, armazena dados referentes ao número e a distribuição das regiões de desordem preditas ao longo da extensão da seqüência de cada proteína.

Um DER (Diagrama de Entidades e Relacionamentos), uma representação gráfica do MER desenvolvido, está disponível no anexo (Anexo 1).

Uma descrição detalhada sobre cada um dos campos das cinco tabelas do MR se encontra em uma tabela em anexo (Anexo 3).

5.6 Inserção das predições no banco de dados relacional

Uma vez estabelecido o MR a etapa subseqüente esteve centrada na correta implementação do banco de dados relacional no SGBD (Sistema Gerenciador de Banco de Dados) MySQL.

Parte fundamental desse processo consiste no tratamento dos arquivos gerados pelos diferentes algoritmos visando à conversão do formato original de saída de cada programa para um formato compatível e adequado a inserção da informação no banco de dados criado.

Para o desenvolvimento dos *parsers* (denominação geral dos *scripts* que analisam, extraem e formatam a informação contida em arquivos texto) utilizamos a linguagem de programação Perl (www.perl.org).

No total, para diferentes tarefas específicas, foram desenvolvidos oito *parsers*.

Como utilizamos quatro preditores de regiões de desordem estrutural, desenvolvemos quatro *scripts* diferentes. Esses quatro *scripts* são os mais importantes de todo o conjunto de *parsers*, pois além de extrair informações referentes às predições de regiões desestruturadas, também extraem os dados básicos tais como: ID (identificador único da proteína), descrição de produto predito, tamanho da seqüência e a própria seqüência do arquivo multi-fasta.

A inserção dessa informação relacionada a cada uma das proteínas analisada é realizada automaticamente no banco de dados por esses quatro *scritps*. É importante salientar que tal inserção somente acontece se a região de desordem estrutural predita for maior do que 40 aminoácidos.

Além desses quatro *parsers* principais, foram desenvolvidos outros quatro, que realizam a inserção das predições de caracterização das seqüências. São *parsers* para os programas TargetP, Phobius, PepStats e para o resultado do Blast contra o banco de dados do *Gene Ontology*.

O funcionamento de todos esses *parsers* é descrito em detalhes na seção de materiais e métodos desse documento (ítem 3.8.2.9.1 até item 3.8.2.9.8).

5.7 Integrando todas as etapas – construção do *pipeline*

Para integrar todas as etapas do projeto em uma única ferramenta automática, foi desenvolvido o *script* trigger.perl. Esse *script* gerencia a execução de todos os programas e *scripts* já descritos. Todos os parâmetros necessários à sua execução, e à execução dos outros programas e *scripts* são obtidos através da linha de comando, como descrito no item 3.12.

A primeira tarefa executada por esse *script* é a criação do banco de dados, descrito no item 4.4, no SGBD MySQL.

Após a criação do banco de dados, o *script* então cria a estrutura de diretórios que será utilizada para armazenar os arquivos gerados após a execução de cada etapa. São criados no total 13 diretórios, em uma estrutura hierárquica.

Após a criação dos diretórios, a etapa executada é o pré-processamento das seqüências assim como descrito no item 4.1. O resultado do pré-processamento é salvo no diretório PROTEOME.

Após o pré-processamento das seqüências, realiza-se então a separação das seqüências em arquivos individuais, etapa exigida por alguns preditores. A separação das seqüências é realizada pelo *script* cut_fasta-mod.pl (disponível como material suplementar no endereço eletrônico iup.cpqrr.fiocruz.br/IUPipeline), e os arquivos separados são salvos no diretório SPLIT, um subdiretório do diretório PROTEOME.

Alguns preditores exigem ainda a remoção da linha de cabeçalho dos arquivos fasta, etapa essa realizada após a divisão das seqüências em arquivos individuais, utilizando-se a ferramenta ‘sed’ do shell bash. As seqüências fasta sem cabeçalho, chamadas de arquivos *flat* são salvas no diretório FLAT_SPLIT, um subdiretório do diretório PROTEOME.

A partir de então as seqüências já podem ser utilizadas por todos os preditores. Seguem-se então as predições, utilizando-se parâmetros fornecidos na linha de comando, como descrito nas seções 3.3 até 3.8 e também no item 3.12. O resultado de cada predição é salvo em um diretório cujo nome remete ao nome do preditor executado. Após a execução de todos os preditores, o *script* trigger.perl realiza o *parsing* dos resultados, utilizando o respectivo *script* de *parsing* para cada preditor, e insere os resultados no banco de dados criado, com parâmetros descritos no item 3.8.

Após o término das inserções, o *script* trigger.perl utiliza mais dois *scripts* para

gerar o consenso das predições segundo a combinação de preditores selecionada (item 4.2.6). O *script* create_prediction_consensus.perl gera um consenso das predições realizadas pela combinação selecionada, e insere na tabela DISORDER_nr do banco de dados as coordenas, o tamanhos de cada região de desordem, a metodologia que realizou a predição (nesse caso composto pelo nome de todas as metodologias que compõem a combinação) e o identificador da proteína ao qual a predição pertence. As buscas subseqüentes no banco de dados então devem ser feitas baseando-se nas predições presentes na tabela DISORDER_nr.

Por fim, o último *script* acionado pelo *pipeline* é o create_fasta_consensus.perl. Esse *script* cria um arquivo fasta contendo a seqüências de todas as IUPs presentes na tabela DISORDER_nr, e acrescenta à linha de cabeçalho de cada seqüência as respectivas coordenas de regiões de desordem estrutural.

As etapas realizadas pelo *script* 'trigger.perl' que compõem o *pipeline* de identificação e caracterização de IUPs são representadas esquematicamente na figura abaixo (Figura 11).

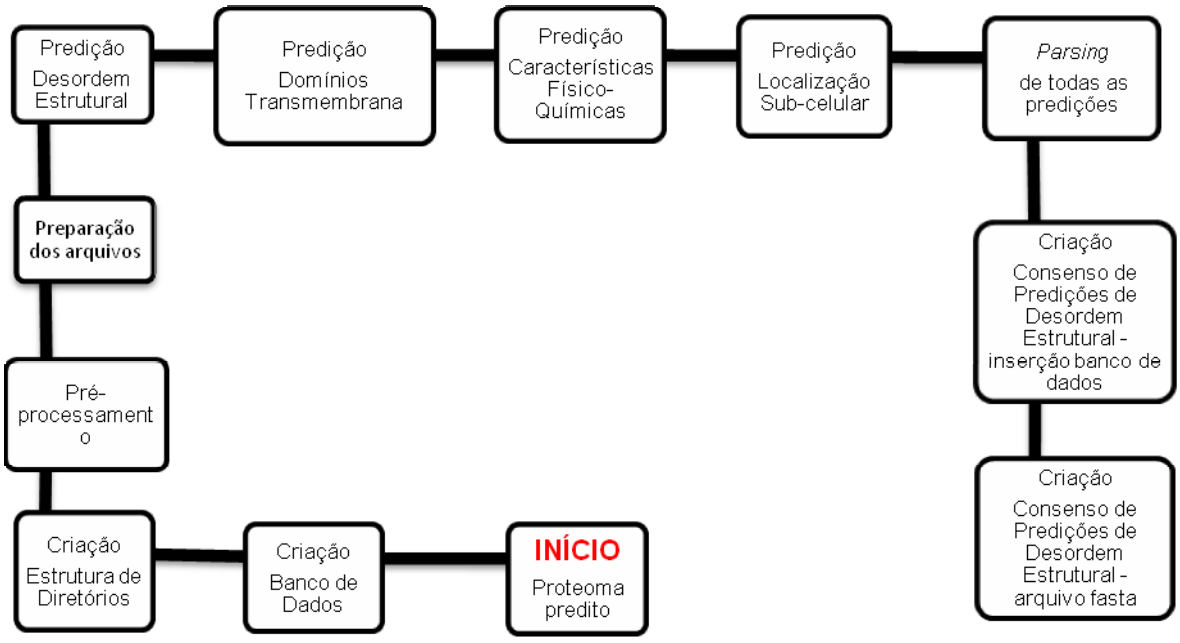


Figura 11: Automatização do *pipeline.*
Representação esquemática das etapas realizadas pelo *script* "trigger.perl".

5.8 Proteômica

Com o objetivo de aumentar a robustez do processo de identificação de proteínas com desordem estrutural, realizamos experimentos utilizando técnicas de proteômica.

Utilizamos um protocolo especial de preparação de amostra, que permite o enriquecimento da presença de IUPs no extrato protéico total de *Schistosoma mansoni*. Esse extrato protéico enriquecido com IUPs foi então utilizado na realização de uma eletroforese bidimensional em gel de poliacrilamida (Galea, Pagala *et al.*, 2006).

O processo de enriquecimento de IUPs no extrato protéico se da por um período de aquecimento da amostra por 1 hora a 98°C. Nessa condição, proteínas globulares se agregam. Após o aquecimento, a amostra é colocada em gelo por 15 minutos, e após esse período, é realizada uma centrifugação a 16.000g por 15 minutos a temperatura ambiente.

Com a centrifugação as proteínas globulares, agregadas durante o aquecimento, precipitam, formando um *pellet*. Com esse procedimento, se obtém um enriquecimento de IUPs no sobrenadante. As proteínas contidas no sobrenadante então são precipitadas com TCA/acetona e utilizadas nas fases seguintes do protocolo. Para descrição detalhada de todo o protocolo vide item 3.11.

Realizamos uma eletroforese unidimensional para avaliarmos a diferença na quantidade e no perfil de proteínas durante o processo de preparação da amostra para enriquecimento da quantidade de IUPs (Figura 12).

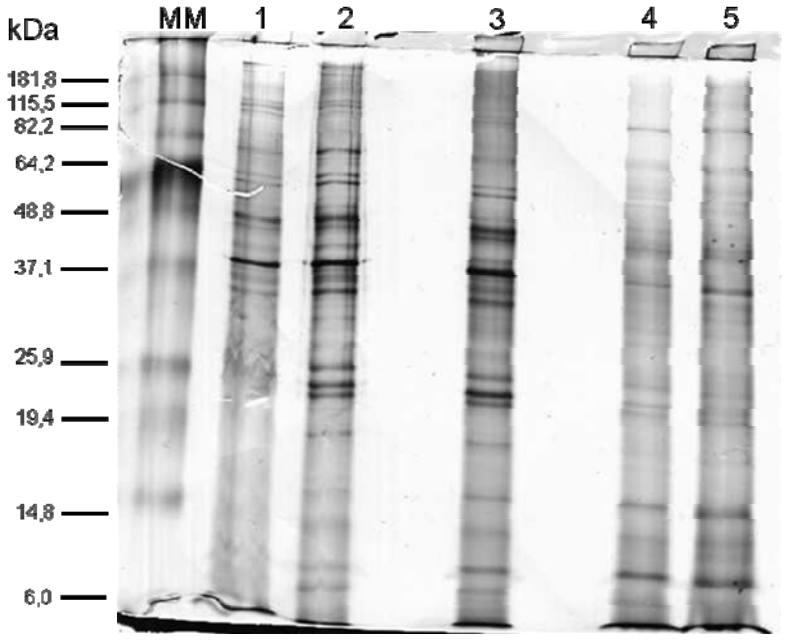


Figura 12: Eletroforese unidimensional. Amostras dos sobrenadantes obtidos durante a fase de preparação do extrato protéico enriquecido com IUPs. Foram aplicados 1µg em cada uma das canaletas: 1) Extrato bruto total de *S. mansoni*; 2) Extrato bruto total de *S. mansoni*; 3) IUPs antes da precipitação por acetona; 4) IUPs (0,5µg) ressuspendidas em IEF; e 5) IUPs ressuspendidas em IEF. Em um gel de poliacrilamida 12%, posteriormente corado pela prata. MM: Marcador de Massa Molecular Benchmark prestained protein ladder 10748-010 (Invitrogen).

Realizamos a focalização isoelétrica do extrato protéico em duas faixas distintas de pH: a) 5-8 e b) 3-10. Como resultado desse experimento, obtivemos os dois géis apresentados na figura abaixo (Figura 13).

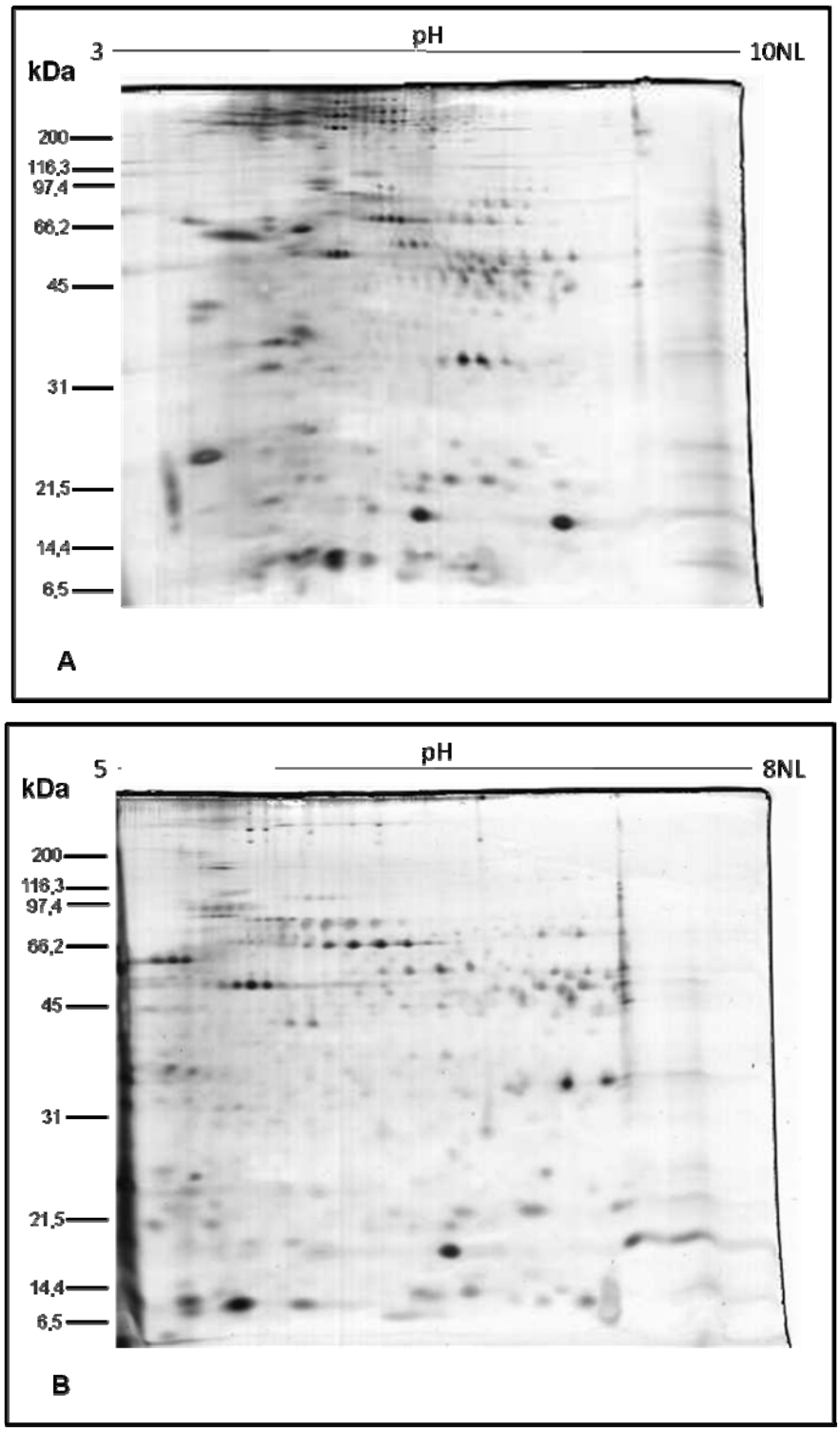


Figura 13: Géis bidimensionais do extrato protéico enriquecido com IUPs. Para a focalização isoelétrica, 15µg de proteínas do extrato protéico total de *S. mansoni* enriquecido com IUPs foi aplicado foram aplicadas em fitas de IPG de 7cm com gradiente de separação não linear de pH A) 3-10 e B) 5-8. Posteriormente as proteínas focalizadas foram separadas por eletroforese em géis de poliacrilamida 12% e coradas por Nitrato de Prata. Peso molecular em kDa (BroadRange;BIO-RAD).

6 DISCUSSÃO

6.1 Interesse prático na identificação de desordem estrutural, sua contribuição para a bioinformática

Apesar da existência de exemplos de dados experimentais relevantes classificando a existência de IUPs como proteínas associadas a diferentes doenças a importância da desordem protéica para função tem sido ignorada (Iakoucheva, Brown *et al.*, 2002). De fato, apesar da existência de relatos de função associada à desordem estruturais de mais de 50 anos, de nosso conhecimento, os principais livros texto atuais de bioquímica não apresentam um único exemplo que evidencie essa relação.

Vale ainda ressaltar, que mesmo vivendo na era genômica onde o seqüenciamento de genomas completos se tornou quase corriqueiro, os bancos de dados contendo anotação estrutural e funcional dos diferentes genomas contem pouca ou nenhuma informação ou mesmo predição em larga escala de desordem protéica.

Desse contexto surge uma pergunta: Qual seria o interesse prático na identificação de regiões de desordem?

Regiões desordenadas freqüentemente apresentam um viés composicional que pode levá-las a apresentar similaridade com proteínas cujas funções não são relacionadas. Assim a identificação de seqüências desordenadas se faz essencial para evitar alinhamentos espúrios com seqüências de proteínas globulares.

De fato, em 2001 Iyer e colaboradores (Iyer, Aravind *et al.*, 2001) demonstraram esse fenômeno. As proteínas ATF-2 e PIF3 haviam sido classificadas funcionalmente por similaridade de seqüência com dois domínios funcionais: HAT e PAS. Iyer e seu grupo utilizaram o programa SEG combinado com alinhamentos múltiplos e predição de estrutura secundária e demonstraram que os domínios HAT e PAS são globulares, mas apresentaram similaridade com uma seqüência desordenada. Esse fato gerou uma forte dúvida com relação a “suspeita homologia” encontrada. Após os estudos de Iyer e seu grupo, a classificação funcional dessas duas proteínas foi desconsiderada.

Desordem estrutural geralmente impede a cristalização de proteínas, e também impede a obtenção de dados interpretáveis em estudos de ressonância nuclear magnética.

A relação entre desordem estrutural e flexibilidade é bem conhecida por cristalógrafos. Quando uma proteína não cristaliza depois de repetidas tentativas, cristalógrafos geralmente removem as regiões hiper-variáveis da seqüência (um viés composicional associado à desordem estrutural), pois estes trechos da seqüência estão associados à alta-flexibilidade da molécula.

Por essa razão, biologistas estruturais utilizam a predição de desordem estrutural para auxiliar a cristalização de proteínas, o que conseqüentemente tem impacto direto na elucidação de estruturas tridimensionais protéicas (Friedberg, Jaroszewski *et al.*, 2004).

A busca por homologia baseada na similaridade de seqüência se vale do dogma estrutura função. Seqüências protéicas codificam estrutura, e estrutura codifica função. Devido a essa relação, utiliza-se a similaridade de seqüência para se inferir função para seqüência desconhecidas.

Entretanto, essa afirmativa não é válida para IUPs, e estratégias adequadas de identificação de ortólogos devem ser utilizadas para essas proteínas.

Em 2004, Kirsten e colaboradores (Rabitsch, Gregan *et al.*, 2004) estudando genes envolvidos no processo de meiose em *Drosophila melanogaster,* identificaram dois novos genes essências para esse processo celular: Sgo1 e Sgo2.

Esses dois genes apresentam uma arquitetura bastante peculiar, o que motivou os pesquisadores a desenvolver uma metodologia inédita para identificação de ortólogos. Eles possuem uma região de *coiled-coil* em sua extremidade N terminais, apresentam na região central um forte viés composicional (elevada freqüência de aminoácidos carregados e hidroxilados e baixa freqüência de aminoácidos hidrofóbicos), característica de regiões desordenadas (confirmada por pelo menos três preditores de desordem estrutural) e possuem motivos funcionais conservados na extremidade C terminal.

Devido à ausência de estruturas secundárias clássicas, a busca por homologia tradicional, baseada no dogma estrutura função descrita não poderia ser utilizada para essas proteínas. Ou seja, buscar ortólogos desses genes em outros organismos utilizando similaridade de seqüência não seria uma opção válida.

Por essa razão os autores introduziram uma metodologia eficiente na identificação de ortólogos de proteínas desordenadas.

Os autores realizaram a busca por genes que apresentassem a mesma arquitetura dos dois genes por eles identificados. Em outras palavras, buscaram

genes que apresentassem uma região de *coiled-coil* na extremidade N terminal, uma região central desordenada (com viés composicional semelhante ao apresentado pelos genes Sgo1 e Sgo2) e um motivo funcional conservado na extremidade C terminal.

A metodologia proposta por Kirsten se mostrou eficiente, e, portanto figura como uma importante contribuição a bioinformática, no que diz respeito a busca por homologia em trechos desordenados.

Motivos Eucarióticos Lineares, ou ELMs (do inglês *Ekariotic Linear Motifs*) são pequenos trechos de seqüência (de 3 a 10 aa) que constituem um motivo funcional. Estão envolvidos em sinalização celular, interação proteína-proteína, fosforilação, acetilação, glicosilação e numa infinidade de outras funções (mais detalhes sobre ELMs podem ser encontrados no endereço http://elm.eu.org/.

Mais de 70% desses motivos são tipicamente encontrados dentro de regiões desordenadas (Puntervoll, Linding *et al.*, 2003; Neduva, Linding *et al.*, 2005). São curtos trechos ordenados intercalados por longas regiões desordenadas. Obviamente, para a predição acurada de ELMs é preciso que haja uma noção precisa do início e do fim das regiões desordenadas. Para obter maior precisão, a predição de desordem estrutural deve ser feita considerando-se o consenso de vários preditores. Portanto, a identificação de regiões desordenadas representa uma ajuda crucial na identificação de ELMs.

Resumindo, além da evidente participação dessas proteínas nos estados de saúde e doença discutidos na parte introdutória desse trabalho, vimos a importante contribuição do estudo da desordem protéica para vários campos da bioinformática incluindo: a busca por similaridade de seqüências, a identificação de ortólogos, a identificação de motivos funcionais e a obtenção e análise de estruturas cristalográficas.

6.2 A combinação de diferentes metodologias melhora a predição de desordem

6.2.1 Gráfico ROC

Um dos aspectos centrais do trabalho desenvolvido esteve centrado na avaliação do desempenho das diferentes combinações dos inúmeros preditores de desordem estrutural. Nosso principal objetivo nessa etapa era responder a seguinte

questão: A integração das diferentes metodologias de predição poderia levar a uma melhor acurácia nas predições?

Um estudo prévio de 2006 de Ferron e colaboradores demonstrou que a predição de regiões de desordem baseada nos resultados de apenas um preditor é pobre e insuficiente para caracterizar o fenômeno (Ferron, Longhi *et al.*, 2006). No mesmo trabalho, os autores sugerem a possível utilização uma abordagem integrativa como maneira mais eficiente de análise. Assim sendo, adotando essa premissa, avaliamos detalhadamente o desempenho combinatorial de seis metodologias distintas de predição de desordem estrutural.

Como discutido anteriormente, o fato de não haver consenso sobre a definição de desordem estrutural levou ao desenvolvimento de preditores que empregam metodologias muito diversas para predição de desordem.

Assim sendo aplicamos um teste de desempenho de predições, ou seja, um teste que tem como objetivo avaliar a acurácia das predições. Essa avaliação feita através da analise de gráficos ROC (Receiver Operating Characteristic).

Integrando os resultados de diferentes preditores, combinamos os pontos fortes de cada metodologia. Avaliamos nos gráficos ROC gerados as taxas de verdadeiro positivo (TPR) e de falso positivo (FPR) em duas condições distintas: a) metodologias individuais; e b) metodologias combinadas.

Quando o desempenho de cada metodologia é avaliado individualmente, vemos que a taxa de verdadeiros positivos fica em torno de 20%. Por outro lado, para as combinações que integram o maior número de metodologias, a taxa de verdadeiro positivo fica próxima de 70%. Um aumento de aproximadamente 50% no número de acertos.

Já a taxa de falsos positivos para metodologias avaliadas individualmente fica entre 20% e 30%. Esse valor sobe pra pouco mais de 40% quando avaliamos combinações de 4 metodologias distintas. Isso representa um aumento de aproximadamente 20% na taxa de erro.

Vimos que a integração de diferentes metodologias apresenta um ganho no número de classificações corretas, entretanto apresenta também um aumento no número de classificações incorretas. Esse comportamento dificulta a obtenção de um resultado ideal de predição, que se caracteriza por uma TPR igual a um e FPR igual a zero. Apesar disso a metodologia desenvolvida apresenta um ganho número de predições acuradas evidente.

Nos itens 5.3.3 e 5.3.4 discutimos algumas perspectivas que podem reduzir a taxa de falsos positivos, aperfeiçoando ainda mais os processo de predição de desordem estrutural.

Um possível viés associado as nossas análises ROC está relacionado ao pequeno número de seqüências utilizadas como controle nesse experimento. Contudo é preciso levar em consideração que os dados de caracterização experimental para proteínas IUP ainda são escassos e que o DisProt atualmente representa o único repositório de seqüências com evidências experimentais para desordem estrutural disponível e de livre acesso. Mesmo sabendo da sua limitação com relação ao número de amostras, optamos por utilizar dados que possuem evidências experimentais para tornar as análises mais robustas.

O número original de seqüências do DisProt (523 em sua versão 4.9) foi ainda reduzido quando submetido ao nosso pré-processamento. No final desse processo, 18,7% (98/523) das seqüências foram descartadas por não satisfazerem os critérios estabelecidos (item 4.1) para as análises subseqüentes, restando assim um total de 425 seqüências. Entretanto, não poderíamos deixar de realizar essa etapa com as seqüências controle uma vez que o processo analítico empregado exige que todas as seqüências sejam expostas às mesmas condições experimentais de análise computacional.

Analisando as predições realizadas por cada uma das seis metodologias empregadas nesse estudo, observamos que o 'COILS' realiza um número de predições extremamente maior do que a média das outras metodologias (vide Gráfico 3). De fato, construímos alguns gráficos ROC incluindo o COILS em nossas análises, e notamos que o desempenho geral de predição era muito prejudicado (gráfico não apresentado).

O COILS identifica resíduos pertencentes a α hélices e folhas β, e classifica todos os outros resíduos (não pertencentes a nenhuma das estruturas citadas anteriormente) como desordenados (vide item 1.3.4.8). Por ter um critério de predição extremamente simplificado, essa metodologia apresenta um número de falsos positivos extremamente alto. Assim sendo, desconsideramos essa metodologia das análises subseqüentes.

Outra metodologia que ficou de fora dos resultados finais de nossas análises foi o HOTLOOPS. Essa metodologia foi considerada em nossas análises até a construção do gráfico ROC, e quando combinada com outras metodologias

apresentou bom desempenho. Entretanto, quando analisando as cinco combinações de metodologias com melhor desempenho de predição (Tabela 2), com o objetivo de selecionarmos apenas uma para as análises de caracterização subseqüentes, o HOTLOOPS foi descartado.

O critério utilizado para a seleção de apenas uma combinação de metodologias, dentre as cinco com melhor desempenho, foi o de escolher a combinação mais conservadora. Ou seja, aquela que apresenta o maior número de Falsos Negativos, em detrimento daquela que apresenta o maior número de Falsos Positivos. Acreditamos que o HOTLOOPS apresenta um comportamento menos conservador em suas classificações devido ao fato de ser derivado do COILS (Linding, Jensen *et al.*, 2003).

Embora tenhamos utilizado um conjunto limitado de seqüências controle, ainda pudemos observar um ganho de 50% no número de acertos obtidos quando integramos quatro metodologias, se comparados ao número de acertos individuais de cada metodologia. Assim podemos responder positivamente a nossa pergunta inicial. A integração de diferentes metodologias de predição de desordem estrutural realmente leva a uma melhor acurácia nas predições.

6.3.2 Peculiaridades das seqüências de aminoácidos das IUPs de S. mansoni

6.3.2.1 Caracterização estrutural e funcional

Uma vez tendo identificado o conteúdo de IUPs no genoma de *S. mansoni* uma etapa essencial para o estudo do papel biológico dessas proteínas é a anotação estrutural e funcional. É através da interpretação do contexto biológico, explicitado pela anotação, que podemos inferir hipóteses sob as diferentes vias nas quais elas se inserem.

Nesse contexto, várias camadas de informação foram inseridas a anotação existente no banco de dados do parasito, entre elas predições de domínios transmembrana, localização sub-celular, características físico-químicas e a utilização de um vocabulário controlado de termos descrevendo as características do produto gênico definido pelo Gene Ontology (http://www.geneontology.org/).

Iniciamos a etapa de caracterização das proteínas com uma análise da freqüência de aminoácidos promotores de desordem estrutural em regiões desordenadas de IUPs e em proteínas globulares.

No gráfico 17, os aminoácidos estão classificados em duas classes: a) resíduos que promovem desordem; e b) resíduos que não promovem desordem.

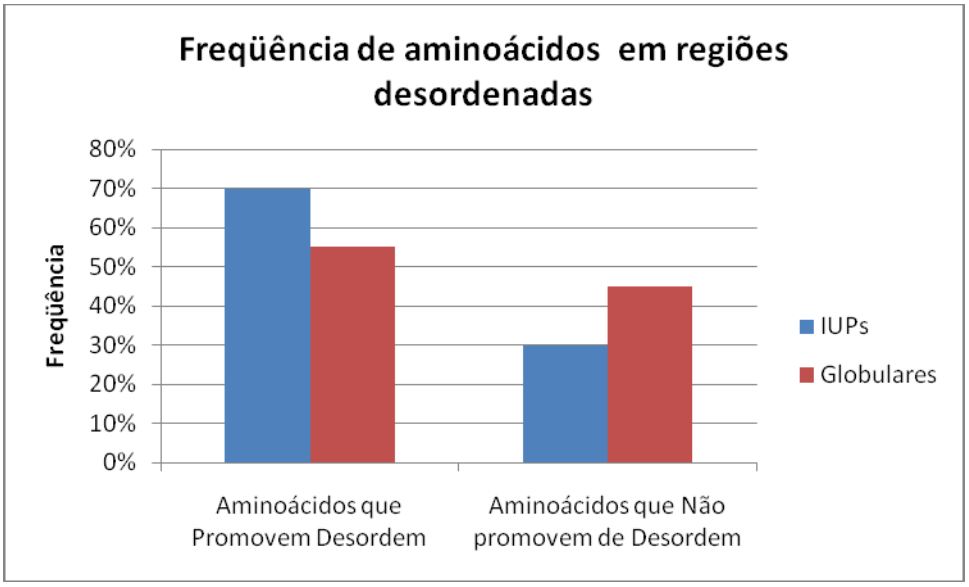


Gráfico 17: Freqüência de aminoácidos em IUPs e proteínas globulares. Freqüência observada de aminoácidos 'promotores de desordem' em regiões desordenadas de IUPs e regiões globulares. Informação sobre resíduos promotores de desordem, e 'não promotores de desordem' obtidas de trabalhos prévios (Ferron, Longhi *et al.*, 2006; Peng, Radivojac *et al.*, 2006; Radivojac, Iakoucheva *et al.*, 2007; Campen, Williams *et al.*, 2008).

No total foram avaliados 210.880 resíduos desordenados e 1.223.598 resíduos ordenados.

A diferença na freqüência de aminoácidos que promovem desordem com relação aos aminoácidos que não promovem desordem, entre as IUPs e as proteínas globulares se mostrou estatisticamente significativa, com p-valor < 0,05 para o teste Qui-quadrado.

Esse resultado corrobora o bom desempenho do nosso processo de identificação de IUPs, uma vez que trabalhos prévios demonstram que há uma marcante diferença na freqüência de determinados aminoácido em regiões desordenadas (Dunker, Oldfield *et al.*, 2008).

Prosseguindo o processo de anotação estrutural das seqüências, analisamos comparativamente, IUPs e proteínas globulares com relação as suas propriedades físico-químicas. Foram avaliadas as diferenças no comprimento médio, no peso molecular médio, no ponto isoelétrico médio e na carga elétrica média. Todos os aspectos citados foram estratificados, e as proteínas foram avaliadas em faixas ou intervalos de cada uma das características.

Os resultados das comparações que apresentaram diferenças estatisticamente significativas entre as IUPs e as proteínas globulares são apresentados do gráfico 7 até o gráfico 10.

Os testes estatísticos realizados (teste T de Student) para ao proteoma avaliado, indicam diferenças nas médias de comprimentos entre IUPs e proteínas globulares menores do que 1000 resíduos, e entre 1000 e 2000 resíduos.

Verificamos também que há uma diferença significativa no ponto isoelétrico médio das IUPs se comparado ao das proteínas globulares no intervalo de pI menor do que 5, no intervalo de pI entre 9 e 10, e no intervalo de pI maior do que 10. Os resultados indicam certa tendência das IUPs apresentarem pI mais ácido e mais básico se comparados as proteínas globulares. Em 2006, Galea e colaboradores (Galea, Pagala *et al.*, 2006) sugeriram que IUPs (identificadas em fibroblastos de camundongo) apresentam tendência a pI ácido. Os nossos resultados indicam que *S. mansoni*, além de apresentar o perfil ácido de pI para as IUPs, também apresenta um perfil básico. Análises complementares devem ser realizadas para investigar essa hipótese.

A carga elétrica entre -50 e 0, e entre 0 e 50 (e) também apresenta diferenças significativas entre IUPs e proteínas globulares. Trabalhos prévios apontam a presença de resíduos carregados como um dos fatores que contribui para a desordem estrutural protéica (Daughdrill, Pielak *et al.*, 2005; Dunker, Oldfield *et al.*, 2008).

Finalizando a análise das características físico-químicas, os testes estatísticos sugerem uma diferença significativa no peso molecular médio das IUPs se comparadas as proteínas globulares na faixa entre 10 e 100kDa.

Todas as outras faixas de comprimentos, intervalos de ponto isoelétrico, intervalos de carga elétrica e faixas de peso molecular não apresentam diferença significativa (gráficos apresentados no anexo 5).

Os resultados nos mostram que existem diferenças significativas nos valores médios das características preditas em poucas situações. Um aspecto que corrobora a nossa análise se relaciona ao fato de que 85% das IUPs apresentam menos de 40% de sua extensão com desordem estrutural (Gráfico 4). Isso significa que a maior parte da extensão dessas proteínas apresenta domínios globulares, e conseqüentemente, deve apresentar propriedades físico-químicas semelhantes às de uma proteína globular. Como nossas análises não foram estratificadas com

relação à extensão desordenada de cada proteína, temos em nosso conjunto de dados um grande número de proteínas que apresentam longos trechos globulares.

Assim como mencionado no inicio desse item, é através do conhecimento do contexto biológico de cada uma das IUPs que podemos inferir hipóteses sob as diferentes vias nas quais elas se inserem. Essa informação é obtida durante a etapa de anotação funcional.

A utilização de algoritmos de agrupamento de seqüências para associar função às proteínas com desordem estrutural não apresenta resultados satisfatórios, como foi demonstrado por Dunker e colaboradores em 2008 (57). Por essa razão, nesse trabalho, a associação de função as IUPs foi baseada na similaridade de seqüências (item 3.7), utilizando um vocabulário controlado de termos descrevendo as características do produto gênico definido pelo *Gene Ontology*.

Todas as seqüências de *S. mansoni* foram comparadas contra o banco de dados de referência do GO, e as classes com maior freqüência são apresentadas do gráfico 11 até o gráfico 16. Observamos nesses gráficos a presença de termos chaves, tais como: regulação, componente extracelular (matriz, membrana), componentes intracelulares (organelas), metabolismo e interação. Nossas conclusões derivam da contagem de quantos termos anotadores apresentam essas palavras chaves no gráfico.

O vocabulário estruturado do *Gene Ontology* classifica a anotação biológica em três grandes categorias: a) Componente celular; b) Processo biológico; e c) Função biológica.

O gráfico de componentes celulares (Gráfico 12) sugere que a classe mais abundante para as IUPs é '*external encapsulating structure part*', notamos também uma grande representatividade de proteínas classificadas como componentes de membrana ou extracelulares. Tais classes estão relacionadas a componentes extracelulares que encobrem toda a célula e que estão localizados na parte externa da membrana plasmática e que portanto devem estar associadas a vias de sinalização e transporte importante no parasito.

Já para as proteínas globulares do parasito, os componentes celulares mais abundantes são intracelulares, tais como: núcleo, citoplasma, mitocôndria e nucléolo (Gráfico 15). Isso representa uma grande diferença no perfil do repertório funcional de proteínas estruturadas e apresentando desestruturação.

De fato nossas constatações suportam alguns resultados de Feng e

colaboradores (22) que mostram que 50% das IUPs identificadas e analisadas em *Plasmodium falciparum* estão associadas à membrana. Foi demonstrado no mesmo estudo, que tais proteínas poderiam ter envolvimento nos processos de interação celular entre parasita e hospedeiro.

Outra característica interessante que surge da analise dos processos biológicos mais abundantes em IUPs é a grande quantidade de proteínas envolvidas em vias de regulação, adesão celular e resposta a estímulos (Gráfico 13). Essas observações surgerem o envolvimento amplo das IUPs nos processos de catálise e nos permitem especular a existência do envolvimento mais freqüente de IUPs em processos relacionados à regulação celular. Tais observações também corroboram estudos feitos em outros organismos (21, 53, 57).

Por outro lado as proteínas globulares de *S. mansoni* apresentaram uma proporção maior de processos biológicos relacionados a metabolismo se comparadas as IUPs (Gráfico 16) fato também corroborado por estudos similares realizados por Dunker e colaboradores em 2008 (57).

Finalizando a análise das três grandes categorias de classificação funcional descrita pelo GO, observamos as funções mais freqüentes.

Para as IUPs, são mais freqüentes os termos anotadores relacionados à regulação (Gráfico 11), assim como descrito para processos celulares mais freqüentes em IUPs. Já para as proteínas globulares, observamos uma expressiva proporção de classes funcionais relacionadas à interação não covalente, de uma molécula com um ou mais sítios ativos de outras moléculas (Gráfico 14).

Finalizando a caracterização das proteínas, realizamos 268 predições de domínios transmembrana para as IUPs preditas, o que compreende 74 proteínas. Dessas 74 proteínas, 42% apresentam somente um domínio transmembrana (Gráfico 5), e podem representar importantes moléculas com função de sinalização celular (Alberts, Johnson *et al.*, 2002).

Outro aspecto interessante apresentado no gráfico 5 está relacionado ao fato de que 58% das proteínas apresentam múltiplos domínios transmembrana o que traz uma camada extra de evidência do envolvimento dessas proteínas nos processos de sinalização e transporte (Alberts, Johnson *et al.*, 2002).

Análises complementares são necessárias para confirmar essas indagações, entretanto estas proteínas representam candidatos interessantes para análises futuras.

6.3.2.2 Proteômica

Com o objetivo de validar experimentalmente nosso pipeline de predição de desordem estrutural iniciamos uma etapa de caracterização experimental das proteínas com desordem estrutural presentes no proteoma de *S. mansoni*.

Para tanto escolhemos uma abordagem de estudo proteômico utilizando protocolos que enriquecem a presença de IUPs nas amostras.

Partindo de vermes adultos (machos e fêmeas) de *S. mansoni*, e seguindo o protocolo previamente descrito por Galea e colaboradores (21), realizamos duas eletroforeses bidimensionais em gel de acrilamida, para separação das IUPs (item 4.8).

Como parte do processo de enriquecimento de IUPs no extrato protéico, realizamos uma eletroforese unidimensional (Figura 12). Nesse gel podemos verificar a redução na quantidade de proteínas total, e uma diferença no perfil das proteínas presentes no extrato protéico bruto e no extrato protéico já enriquecido com IUPs.

Durante essa fase de validação experimental realizamos duas eletroforeses bidimencional em duas faixas de pH: 5-8 e 3-10 (Figura 13) que mostram o resultado esperado de enriquecimento das IUPs no gel.

Os dois géis apresentam quantidades significativas de *spots* que estão sendo processados para posterior identificação por espectrometria de massa.

Após a identificação, acrescentaremos ao nosso banco de dados mais uma camada de anotação relacionada à desordem estrutural no proteoma de *S. mansoni*.

6.3.3 Erros associados às metodologias de predição de desordem estrutural

A discrepância mais evidente nos resultados de diferentes metodologias de predição para uma mesma proteína é a definição das coordenadas limítrofes de uma região desordenada (Ferron, Longhi *et al.*, 2006).

A falta de consenso dos preditores com relação às coordenadas limítrofes das regiões de desordem tem um impacto importante no desempenho integrado de diferentes metodologias.

Quando observamos o resultado da predição de dois ou mais preditores para uma mesma região de uma proteína, não podemos considerar como verdadeiro positivo a predição combinada se uma delas extrapola em muito a região

verdadeiramente desordenada. Isso faz com que a essa combinação seja penalizada, não recebendo pontuação no seu somatório de verdadeiros positivos, devido a discrepâncias com relação ao limite da região desordenada. Isso conseqüentemente tem uma implicação no valor de falso positivo, uma vez que existe predição de desordem para uma região estruturada.

Por essa razão, a taxa de falsos positivos de uma combinação de preditores fatalmente será mais alta se comparada à taxa de falsos positivos de uma metodologia individual.

Essas discrepâncias surgem em função das diferentes definições de desordem empregada em cada metodologia, do conjunto de dados utilizados para treinamento e dos métodos de avaliação de erros considerados no desenvolvimento de cada programa.

Além desses fatores relacionados à implementação e ao desenvolvimento de cada preditor, aspectos experimentais também contribuem para a falta de consenso existente entre os programas com relação à posição e ao tamanho de uma região desordenada. As técnicas experimentais utilizadas na identificação de desordem estrutural estão sujeitas a falhas. Artefatos presentes durante a etapa de cristalização das proteínas e a resolução da estrutura de um peptídeo em meio não aquoso representam dois exemplos comuns de fontes de erro.

Complementarmente algumas das propriedades das IUPs também podem dificultar a identificação de regiões desordenadas. A existência do processo conhecido com *induced folding* é um exemplo dessa constatação. Quando associadas a um ligante, regiões desordenadas podem assumir conformações temporárias que mascaram a identificação da desordem estrutural.

Um outro aspecto que merece destaque nesse contexto está relacionado a existência de uma pequena quantidade de proteínas com evidências experimentais de desordem estrutural disponíveis para treinamento dos preditores. Tal fato resulta em programas que somente apresentam bom desempenho quando realizam predições para proteínas semelhantes aquelas utilizadas no seu treinamento. Ou seja, os preditores de desordem estrutural apresentam vieses relacionados ao restrito conjunto de treinamento disponível para uso. Atrelado a esse fato, a existência de três tipos de desordem estrutural, batizadas por Vucetic e colaboradores de tipos ‘V’, ‘C’ e ‘S’ com diferentes particularidades impede que algoritmos treinados com um tipo reconheçam o outro. Vucetic demonstrou ainda a

existência de mais um complicador.

Em resumo podemos agrupar todos esses fatores em três itens que justificam o aumento na taxa de falsos positivos nas predições:

1. A diversidade de tipos de desordem estrutural presente nos organismos;
2. A existência de preditores que apresentam desempenhos diferentes para cada tipo de desordem; e
3. Em estudos de larga escala não é possível prever quais e em que proporção cada tipo de desordem irá aparecer.

No intuito de tentar minimizar a influência desses diferentes fatores na identificação das regiões limítrofes das regiões de desordem preditas utilizamos uma abordagem através da qual o mapeamento de predições assessórias de domínios e/ou motivos funcionais foram utilizadas no ajuste das coordenadas de desordem.

Nossos resultados demonstraram que 5,5% (410/7.373) das RLDs preditas tiveram suas coordenadas ajustadas por comparação com domínios conservados (vide item 4.4) fato que evidencia a abordagem como importante dentro do contexto geral de predição integrada de desordem funcional.

6.3.4 Aperfeiçoamento dos métodos de predição de desordem estrutural

Apesar de nossos resultados indicarem um ganho evidente (50% de aumento na taxa de verdadeiros positivos) na qualidade e acurácia das predições. Avaliamos como possível a implementação de uma nova etapa de aperfeiçoamento dos métodos integrativos de predição de desordem estrutural.

A seguir delineamos algumas idéias que em um futuro próximo poderão ser implementadas no *pipeline* desenvolvido nesse trabalho.

6.3.4.1 Conhecimento prévio de desvios de composição

A primeira perspectiva de aperfeiçoamento dos métodos de predição é a identificação de regiões com algum desvio composicional, mas que não esteja relacionado à desordem. Como exemplo de desvio de composição podemos citar: a) regiões de baixa complexidade; b) presença de peptídeos sinais; c) domínios transmembrana; d) zíperes de leucina e; e) coiled-coils. A identificação prévia desse

tipo de características na seqüência evita o risco de classificações errôneas como a predição de coiled-coils como região de desordem estrutural.

Ferron e colaboradores (2) demonstraram como utilizar esse conhecimento prévio em um experimento em 2006. Selecionaram uma proteína pequena, que apresenta trechos desordenados em suas extremidades N e C terminal (resíduos 1-8 e 58-76 respectivamente). Submeteram a seqüência a identificação de regiões de desordem por diversos preditores, e avaliaram os resultados de cada um deles individualmente. Os preditores apresentaram resultados divergentes. Alguns preditores apresentaram predições de desordem estrutural nas extremidades e também ao longo de toda a extensão da seqüência. Outros somente na região central, e outros sequer apresentaram predições de desordem.

Em análises computacionais tradicionais para identificação de desordem estrutural a obtenção de um resultado consenso sobre a correta posição das regiões de desordem seria bastante complicada. O conhecimento prévio da existência e da localização de um coiled-coil teria resolvido o impasse.

De fato, utilizando um preditor de coiled-coils e a técnica de HCA, os autores do estudo identificaram um coiled-coil que se estende por toda a região central da seqüência. Fica, portanto, demonstrada a contribuição que essas análises podem trazer as metodologias de predição de desordem estrutural.

6.3.4.2 Conhecimento prévio da existência de estruturas secundárias

Por definição, estruturas secundárias e regiões desordenadas não podem ser coincidentes, portanto, assim como a identificação de domínios, motivos ou desvios composicionais, a identificação prévia de estruturas secundárias pode contribuir para a correta identificação de desordem estrutural.

6.3.4.3 Alta variabilidade de seqüência em regiões desordenadas

Como descrito no item 1.3.3.2.4, regiões desordenadas apresentam maior variabilidade de seqüência se comparadas a regiões ordenadas. Essa informação é valiosa e pode ser utilizada no aperfeiçoamento das predições. Alinhamentos múltiplos de seqüência (MSA do inglês *Multiple Sequence Alignment*) representam a maneira mais adequada de se avaliar essa variabilidade. MSAs podem ser utilizados

para auxiliar a identificação de regiões de desordem e inclusive contribuir para a identificação de suas regiões limítrofes.

6.3.4.4 Automatização da técnica de HCA

A técnica de HCA tem sido sugerida como bastante eficiente para auxiliar a delimitação das coordenadas limítrofes de regiões desordenadas. Entretanto, a análise das coordenadas dos domínios hidrofóbicos tem que ser feita manualmente. Isso inviabiliza a utilização dessa técnica em escala proteômica.

A automatização da construção de gráficos HCA e da análise das coordenadas limítrofes seria uma contribuição importante para os processos computadorizados de predição de desordem estrutural.

6.3.4.5 Contribuição para o entendimento do fenômeno de desordem estrutural

É importante também analisarmos todos os aspectos citados sob o ponto de vista do *pipeline* desenvolvido nesse trabalho. Até o presente momento, não existe nenhuma metodologia automática de predição de desordem estrutural que integre diferentes preditores, empregando uma técnica sistemática e automatizada para avaliar qual a melhor combinação de programas.

A utilização de gráficos ROC possibilita a análise rápida e simples da melhor combinação de preditores para um determinado conjunto de proteínas desordenadas. As ferramentas criadas possibilitam a rápida inclusão de novas seqüências ao conjunto de teste, além de possibilitarem também a inclusão de novos preditores, sendo somente necessário o desenvolvimento de um *parser* caso este resultado se apresente em um formato específico não contemplado anteriormente. Assim é possível que se escolha a melhor combinação de preditores de acordo com a natureza das seqüências com que se está trabalhando. A integração de todas as predições de desordem estrutural em um banco de dados relacional permite inclusive que se façam análises complexas acerca das semelhanças e diferenças dos preditores para uma mesma proteína. Além disso, torna bastante simples a comparação e ajuste das coordenadas limítrofes de uma região de desordem estrutural pela comparação com as coordenadas de um domínio ou motivo funcional, de um sítio ativo ou de outras características funcionais e

estruturais contempladas no banco de dados relacional.

A metodologia desenvolvida, por integrar dados de diferentes naturezas acerca de cada uma das seqüências de estudo, torna possível a identificação de novos tipos de desordem estrutural (além dos três descritos por Vucetic (V, C e S) (Vucetic, Brown *et al.*, 2003). Além disso, a descoberta da existência de IUPs em diferentes organismos, tais como parasitos extracelulares como é o caso do *S. mansoni* pode elucidar o envolvimento dessas proteínas em processos biológicos ainda não relacionados com desordem estrutural.

Esses novos conhecimentos podem também contribuir com o aperfeiçoamento dos métodos computacionais de predição de desordem estrutural.

7 CONCLUSÕES

Os resultados obtidos ao final deste trabalho nos permitiram concluir que:

1. O banco de dados relacional desenvolvido permite a integração de predições baseadas em diferentes definições de desordem estrutural, além de albergar informações referentes à caracterização funcional e estrutural das IUPs identificadas.

2. A utilização de gráficos ROC combinada com a integração de dados proporcionada pelo banco de dado relacional se mostrou eficiente na seleção de uma combinação de preditores com maior acurácia (50%) na identificação de regiões desordenadas.

3. A integração das coordenadas de domínios funcionais conservados ao banco de dados permitiu o ajuste das coordenadas limítrofes de aproximadamente 5,5% das regiões desordenadas preditas.

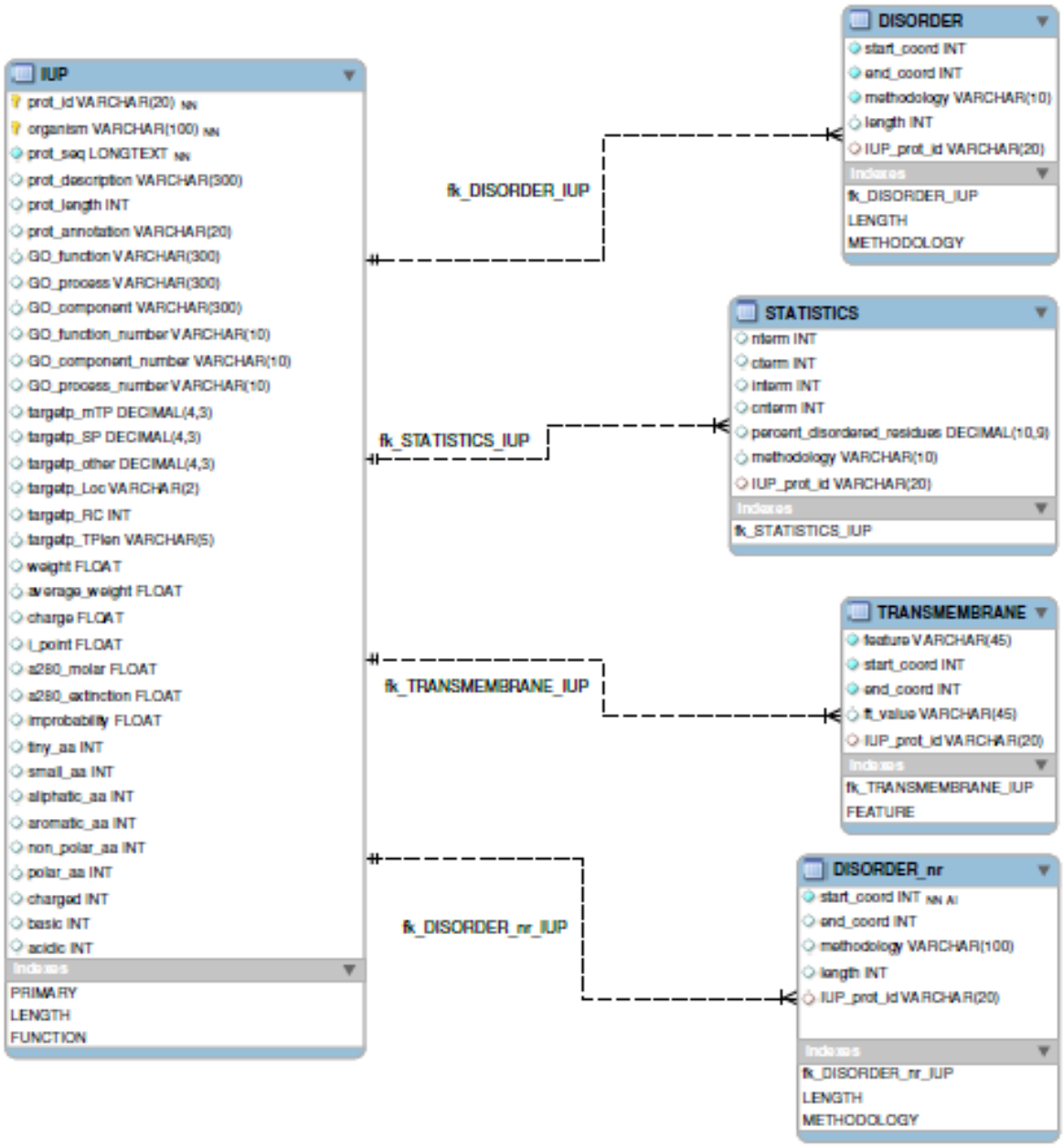
4. Aproximadamente 33,6% do proteoma predito de *S. mansoni* apresenta desordem estrutural.

5. As IUPs identificadas no proteoma predito de *S. mansoni* em sua maioria estão envolvidas em processos de regulação celular e componentes extracelulares.

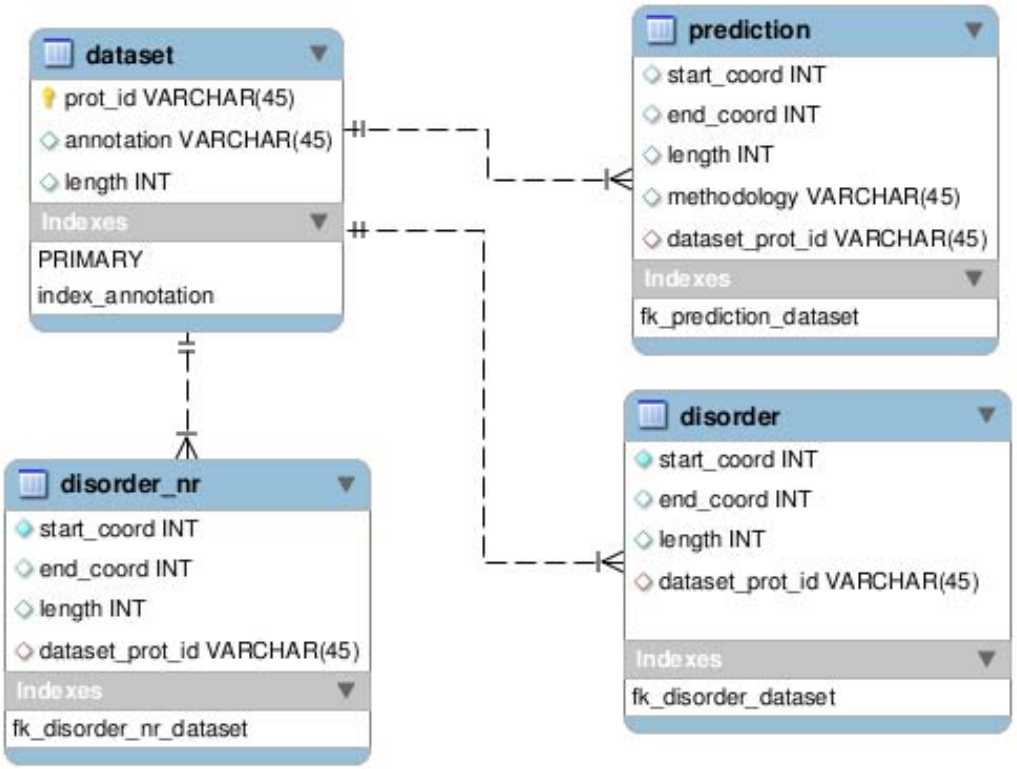
6. A integração de todas as etapas de predição e caracterização de IUPs em um pipeline automático representa uma contribuição importante para as metodologias computacionais de identificação e caracterização de desordem estrutural em proteomas preditos.

8 ANEXOS

8.1 Anexo 1: Diagrama Entidade Relacionamento – DER – *Pipeline*.



8.2 Anexo 2: Diagrama Entidade Relacionamento – DER – avaliação de desempenho dos preditores de desordem estrutural.



8.3 Anexo 3: Descrição de todos os atributos (campos) das cinco tabelas do DER (Anexo 1) desenvolvido para o pipeline.

Tabela IUP	
ATRIBUTO	**DESCRIÇÃO**
prot_id	Identificador (ID) da proteína pra qual foi predita a região de desordem.
organism	Nome do organismo analisado.
prot_seq	Seqüência de aminoácidos da proteína analisada.
prot_description	Descrição da proteína. Presente na anotação do genoma (arquivo fasta).
prot_length	Tamanho da proteína, em número de aminoácidos.
prot_annotation	Anotação da proteína como hipotética ou com função predita.
GO_function	Função mais geral (primeiro nível) predita pelo GO para a proteína.
GO_process	Processo mais geral (primeiro nível) predito pelo GO para a proteína.
GO_component	Componente mais geral (primeiro nível) predito pelo GO para a proteína.
GO_function_number	Identificador (número) da função predita.
GO_component_number	Identificador (número) do componente predito.
GO_process_number	Identificador (número) do processo predito.
targetp_mTP	Escore que representa a chance do peptídeo ser mitocondrial.
targetp_SP	Escore que representa a chance do peptídeo ser de sinalização (peptídeo sinal).
targetp_other	Escore que representa a chance do peptídeo ser localizado em qualquer outro compartimento celular que não seja mitocôndria. Representa também a chance da seqüência não representar um peptídeo sinal.
targetp_Loc	Predição da localização (*chloroplast, mitochondrion, secretory pathway, any other location e don't know*), baseado nos escores descritos acima.
targetp_RC	Indica a "certeza" da predição, é a diferença entre o maior escore e o segundo maior escore, quanto menor o valor mais segura é a predição.
targetp_TPlen	Tamanho da seqüência sinalizadora de localização celular.
weight	Peso molecular.
average_weight	Média do peso do resíduo.
charge	Carga elétrica da proteína.
i_point	Ponto isoelétrico.
a280_molar	Coeficiente de extinção molar (A280).
a280_extinction	Coeficiente de extinção 1mg/ml (A280).
improbability	Improbabilidade de expressão nos corpos de inclusão.
tiny_aa	Número de aminoácidos muito pequenos.
small_aa	Número de aminoácidos pequenos.
aliphatic_aa	Número de aminoácidos alifáticos.
aromatic_aa	Número de aminoácidos aromáticos.
non_polar_aa	Número de aminoácidos não polares.
polar_aa	Número de aminoácidos polares.

charged	Número de aminoácidos carregados.
basic	Número de aminoácidos básicos.
acidic	Número de aminoácidos ácidos.
Tabela DISORDER	
ATRIBUTO	**DESCRIÇÃO**
star_coord	Coordenada inicial da região de desordem predita.
end_coord	Coordenada final da região de desordem predita.
length	Tamanho da região de desordem predita.
IUP_prot_id	Identificador da proteína que foi predita a região de desordem.
methodology	Metodologia responsável pela predição da região de desordem.
Tabela DISORDER_nr	
ATRIBUTO	**DESCRIÇÃO**
star_coord	Coordenada inicial da nova região de desordem predita, sem redundância.
end_coord	Coordenada final da nova região de desordem predita, sem redundância.
length	Tamanho da região de desordem predita.
IUP_prot_id	Identificador da proteína que foi predita a região de desordem.
methodology	Metodologia responsável pela predição da região de desordem.
Tabela STATISTICS	
ATRIBUTO	**DESCRIÇÃO**
nterm	Quantidade de regiões de desordem predita na região N terminal da proteína.
cterm	Quantidade de regiões de desordem predita na região C terminal da proteína.
interm	Quantidade de regiões de desordem predita na região intermediária (entre C e N terminal) da proteína.
cnterm	Quantidade de regiões de desordem predita que abrange as regiões C e N terminal da proteína.
percent_disordered_residues	Porcentagem de resíduos desordenados em relação ao total de resíduos da proteína.
methodology	Metodologia responsável pela predição da região de desordem.
IUP_prot_id	Identificador da proteína que foi predita a região de desordem.
Tabela TRANSMEMBRANE	
ATRIBUTO	**DESCRIÇÃO**
feature	Tipo de domínio predito.
start_coord	Coordenada inicial do domínio predito.
end_coord	Coordenada final do domínio predito.
ft_value	Localização da região.
IUP_prot_id	Identificador da proteína analisada.

8.4 Anexo 4: Tabela com todos os preditores de desordem estrutural avaliados.

PREDITOR	ANO DE PUBLICAÇÃO	ARTIGO DE REFERÊNCIA	DISPONIBILIDADE PARA DOWNLOAD	COMENTÁRIOS	CITAÇÕES ISI*	SITUAÇÃO
PONDR	2001	acesso restrito	proprietário	Por ser proprietário, não é uma opção para esse projeto.	319	não adquirido
DISOPRED2	2004	ok	sim	Apresentou erro de compilação.	256	não instalado
DisEMBL	2003	ok	sim	ok	223	instalado e funcional
GlobPlot 2	2003	ok	sim	ok	202	instalado e funcional
FoldIndex©	2005	ok	somente web service	Como o download não é possível, não é uma opção.	126	-
VL2	2003	ok	sim	ok	119	instalado e funcional
IUPred	2005	ok	requisitar ao autor	ok	105	instalado e funcional
VL3, VL3H, VL3E	2003	ok	sim	ok	96	erro durante teste
VSL2	2005	ok	sim	ok	50	erro durante teste
PreLink	2005	ok	web	Como o download não é possível, não é uma opção para esse projeto.	46	-
SPRITZ	2006	ok	web	Como o download não é possível, não é uma opção para esse projeto.	23	-
FoldUnfold	2006	ok	web	Como o download não é possível, não é uma opção para esse projeto.	16	-
DISpro	2005	-	sim	Apresentou erro durante testes de predição.	-	erro durante teste
DRIPPRED	-	ok	requisitar ao autor	Problemas durante a instalação.	-	não instalado

9 REFERÊNCIAS BIBLIOGRÁFICAS

Alberts, Bruce et al. Molecular biology or the cell. 4. ed. New York: Garland Science, 2002. 1463 p. il. ISBN 0-8153-3218-1.

Altschul, S. et al. Basic local alignment search tool. J Mol Biol, v. 215, n. 3, p. 403-10, Oct 1990.

Anfinsen, C. Principles that govern the folding of protein chains. Science, v. 181, n. 96, p. 223-30, Jul 1973.

Anson, M.; Mirsky, A. On some general properties of proteins. J Gen Physiol, v. 9, n. 2, p. 169-179, Nov 1925.

Ausubel, F. M. *et al.* Current Protocols in Molecular Biology. New York Greene Publishing Associates and Wiley-Interscience, 1995.

Bergquist, N. Schistosomiasis: from risk assessment to control. Trends Parasitol, v. 18, n. 7, p. 309-14, Jul 2002.

Berriman, M. *et al.* The genome of the blood fluke Schistosoma mansoni. Nature, v. 460, n. 7253, p. 352-8, Jul 2009.

Brown, C. *et al.* Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol, v. 55, n. 1, p. 104-10, Jul 2002.

Campen, A. *et al.* TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett, v. 15, n. 9, p. 956-63, 2008.

Chervitz, S. *et al.* Comparison of the complete protein sets of worm and yeast: orthology and divergence. Science, v. 282, n. 5396, p. 2022-8, Dec 1998.

Chitsulo, L. *et al.* The global status of schistosomiasis and its control. Acta Trop, v. 77, n. 1, p. 41-51, Oct 2000.

Cioli, D. *et al.* Determination of ED50 values for praziquantel in praziquantel-resistant and -susceptible Schistosoma mansoni isolates. Int J Parasitol, v. 34, n. 8, p. 979-87, Jul 2004.

Csizmók, V. *et al.* Towards proteomic approaches for the identification of structural disorder. Curr Protein Pept Sci, v. 8, n. 2, p. 173-9, Apr 2007.

_____. A novel two-dimensional electrophoresis technique for the identification of intrinsically unstructured proteins. Mol Cell Proteomics, v. 5, n. 2, p. 265-73, Feb 2006.

Daughdrill, G. W. *et al.* Natively Disordered Proteins. In: Buchner J, Kiefhaber T. Protein Folding Handbook. 1st ed. Weinheim: Wiley-VCH, 2005. P. 275-357.

Davis, a. H. Schistosomiasis. Epidemiology and the Community Control of Disease. In: Derek R. Warm Climate Countries. 2nd ed. Edinburgh: Churchill Livingstone,

1984. P. 389-412.

de Silva, N. *et al.* Anthelmintics. A comparative review of their clinical pharmacology. Drugs, v. 53, n. 5, p. 769-88, May 1997.

Deléage, G.; Roux, B. An algorithm for protein secondary structure prediction based on class prediction. Protein Eng, v. 1, n. 4, p. 289-94.

Demchenko, A. Recognition between flexible protein molecules: induced and assisted folding. J Mol Recognit, v. 14, n. 1, p. 42-61.

Doenhoff, M. *et al.* Praziquantel: mechanisms of action, resistance and new derivatives for schistosomiasis. Curr Opin Infect Dis, v. 21, n. 6, p. 659-67, Dec 2008.

Dosztányi, Z. *et al.* The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol, v. 347, n. 4, p. 827-39, Apr 2005.

Dunker, A. *et al.* Intrinsic disorder and protein function. Biochemistry, v. 41, n. 21, p. 6573-82, May 2002.

_____. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. Pac Symp Biocomput, p. 473-84, 1998.

_____. Intrinsically disordered protein. J Mol Graph Model, v. 19, n. 1, p. 26-59, 2001.

_____. The unfoldomics decade: an update on intrinsically disordered proteins. BMC Genomics, v. 9 Suppl 2, p. S1, 2008.

Dyson, H.; Wright, P. Coupling of folding and binding for unstructured proteins. Curr Opin Struct Biol, v. 12, n. 1, p. 54-60, Feb 2002.

Dyson, M. *et al.* Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression. BMC Biotechnol, v. 4, p. 32, Dec 2004.

Elmasri, R.; NAVATHE, S. B. Sistema de Banco de Dados - Fundamentos e Aplicações. 4ª ed. São Paulo: Pearson Education do Brasil Ltda, 2005.

Emanuelsson, O. *et al.* Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol, v. 300, n. 4, p. 1005-16, Jul 2000.

Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. Palo Alto: Hewlett-Packard Labs, 2004.

Feng, Z. *et al.* Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. Mol Biochem Parasitol, v. 150, n. 2, p. 256-67, Dec 2006.

Ferron, F. *et al.* A practical overview of protein disorder prediction methods. Proteins,

v. 65, n. 1, p. 1-14, Oct 2006.

Fischer, E. Einfluss der Configuration auf die Wirkung der Enzyme. Ber Dtsch Chem Ges, v. Vol. 27, n. 3, p. 2985-2993, 1894.

Friedberg, I. *et al.* The interplay of fold recognition and experimental structure determination in structural genomics. Curr Opin Struct Biol, v. 14, n. 3, p. 307-12, Jun 2004.

Gaboriaud, C. *et al.* Hydrophobic cluster analysis: an efficient new way to compare and analyse amino acid sequences. FEBS Lett, v. 224, n. 1, p. 149-55, Nov 1987.

Galea, C. *et al.* Proteomic studies of the intrinsically unstructured mammalian proteome. J Proteome Res, v. 5, n. 10, p. 2839-48, Oct 2006.

Gibas, C.; Jambeck, P. Desenvolvendo Bioinformática. Rio de Janeiro: Editora Campus, 2001.

Green, D. M.; Swets, J. M. Signal detection theory and psychophysics. New York: John Wiley and Sons Inc., 1966.

Gryseels, B. *et al.* Are poor responses to praziquantel for the treatment of Schistosoma mansoni infections in Senegal due to resistance? An overview of the evidence. Trop Med Int Health, v. 6, n. 11, p. 864-73, Nov 2001.

_____. Human schistosomiasis. *Lancet*, v. 368, n. 9541, p. 1106-18, Sep 2006.

Hartwell, L.; Kastan, M. Cell cycle control and cancer. Science, v. 266, n. 5192, p. 1821-8, Dec 1994.

Hokke, C. *et al.* Glycomics-driven discoveries in schistosome research. Exp Parasitol, v. 117, n. 3, p. 275-83, Nov 2007.

Hubbard, S. *et al.* Modeling studies of the change in conformation required for cleavage of limited proteolytic sites. Protein Sci, v. 3, n. 5, p. 757-68, May 1994.

Iakoucheva, L. *et al.* Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol, v. 323, n. 3, p. 573-84, Oct 2002.

Iyer, L. *et al.* Quoderat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. Genome Biol, v. 2, n. 12, p. RESEARCH0051, 2001.

Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, v. 22, n. 12, p. 2577-637, Dec 1983.

Katz, N.; Peixoto, S. Critical analysis of the estimated number of Schistosomiasis mansoni carriers in Brazil. Rev Soc Bras Med Trop, v. 33, n. 3, p. 303-8, 2000 May-Jun 2000.

Kendrew, J. *et al.* A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature, v. 181, n. 4610, p. 662-6, Mar 1958.

_____. Structure of myoglobin: A three-dimensional Fourier synthesis at 2 A. resolution. Nature, v. 185, n. 4711, p. 422-7, Feb 1960.

Koonin, E. V.; Galperin, M. Sequence-evolution-function: computational approaches in comparative genomics. Norwell: Kluwer Academic Publishers, 2003.

Käll, L. *et al.* A combined transmembrane topology and signal peptide prediction method. J Mol Biol, v. 338, n. 5, p. 1027-36, May 2004.

Linding, R. *et al.* Protein disorder prediction: implications for structural proteomics. Structure, v. 11, n. 11, p. 1453-9, Nov 2003.

_____. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res, v. 31, n. 13, p. 3701-8, Jul 2003.

Io Conte, L. *et al.* SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res, v. 30, n. 1, p. 264-7, Jan 2002.

Loukas, A. *et al.* Schistosome membrane proteins as vaccines. Int J Parasitol, v. 37, n. 3-4, p. 257-63, Mar 2007.

Loverde, P. *et al.* Schistosoma mansoni genome project: an update. Parasitol Int, v. 53, n. 2, p. 183-92, Jun 2004.

Neduva, V. *et al.* Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol, v. 3, n. 12, p. e405, Dec 2005.

Obradovic, Z. *et al.* Exploiting heterogeneous sequence properties improves prediction of protein disorder. Proteins, v. 61 Suppl 7, p. 176-82, 2005.

Obuchowski, N. Receiver operating characteristic curves and their use in radiology. Radiology, v. 229, n. 1, p. 3-8, Oct 2003.

Oliveira, C. H. P. D. SQL - Curso Prático. São Paulo: Editora Novatec, 2002.

Pauling, L. *et al.* The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. Proc Natl Acad Sci U S A, v. 37, n. 4, p. 205-11, Apr 1951.

Peng, K. *et al.* Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics, v. 7, p. 208, 2006.

Pepe, M. S. The statistical evaluation of medical tests for classification and prediction. New York: Oxford, 2003.

Puntervoll, P. *et al.* ELM server: A new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res, v. 31, n. 13, p. 3625-30, Jul 2003.

Rabitsch, K. *et al.* Two fission yeast homologs of Drosophila Mei-S332 are required for chromosome segregation during meiosis I and II. Curr Biol, v. 14, n. 4, p. 287-301, Feb 2004.

Radivojac, P. *et al.* Intrinsic disorder and functional proteomics. Biophys J, v. 92, n. 5, p. 1439-56, Mar 2007.

Rice, P. *et al.* EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet, v. 16, n. 6, p. 276-7, Jun 2000.

Shaiu, W. *et al.* The hydrophilic, protease-sensitive terminal domains of eucaryotic DNA topoisomerases have essential intracellular functions. Pac Symp Biocomput, p. 578-89, 1999.

Spackman, K. A. Signal detection theory: Valuable tools for evaluating inductive learning. Proceedings of the Sixth International Workshop on Machine Learning; 1989 Jun 26-27; Ithaca, New York, United States. San Mateo: California; Morgan Kaufmann Publishers Inc; 1989.

Spolar, R.; RECORD, M. J. Coupling of local folding to site-specific binding of proteins to DNA. Science, v. 263, n. 5148, p. 777-84, Feb 1994.

Steinmann, P. *et al.* Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis, v. 6, n. 7, p. 411-25, Jul 2006.

Utzinger, J. *et al.* The potential of artemether for the control of schistosomiasis. Int J Parasitol, v. 31, n. 14, p. 1549-62, Dec 2001.

Vucetic, S. *et al.* Flavors of protein disorder. Proteins, v. 52, n. 4, p. 573-84, Sep 2003.

Wilson, R.; Coulson, P. Schistosome vaccines: a critical appraisal. Mem Inst Oswaldo Cruz, v. 101 Suppl 1, p. 13-20, Sep 2006.

Wilson, R. *et al.* From genomes to vaccines via the proteome. Mem Inst Oswaldo Cruz, v. 99, n. 5 Suppl 1, p. 45-50, 2004.

Wu, H. Studies on denaturation of proteins. XIII. A theory of denaturation. 1931. Adv Protein Chem, v. 46, p. 6-26; discussion 1-5, 1995.

Yuan, Z. *et al.* Prediction of protein B-factor profiles. Proteins, v. 58, n. 4, p. 905-12, Mar 2005.

Zerlotini, A. *et al.* SchistoDB: a Schistosoma mansoni genome resource. Nucleic Acids Res, v. 37, n. Database issue, p. D579-82, Jan 2009.

Zweig, M.; Campbell, G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem, v. 39, n. 4, p. 561-77, Apr 1993.