Ministério da Saúde

Fundação Oswaldo Cruz

Centro de Pesquisas René Rachou

Programa de Pós-graduação em Ciências da Saúde

Análise computacional baseada no desenvolvimento de um *pipeline* de técnicas *ab initio* para predição de desordem estrutural protéica em genomas de tripanosomatídeos

por

Patrícia de Cássia Ruy

Belo Horizonte

Janeiro/2011

Ministério da Saúde

Fundação Oswaldo Cruz

Centro de Pesquisas René Rachou

Programa de Pós-graduação em Ciências da Saúde

Análise computacional baseada no desenvolvimento de um *pipeline* de técnicas *ab initio* para predição de desordem estrutural protéica em genomas de tripanosomatídeos

por

Patrícia de Cássia Ruy

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

Orientação: Dr. Jeronimo Conceição Ruiz

Belo Horizonte

Janeiro/2011

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

R985a 2011

Ruy, Patrícia de Cássia.

Análise computacional baseada no desenvolvimento de um pipeline de técnicas ab initio para predição de desordem estrutural protéica em genomas de tripanosomatídeos / Patrícia de Cássia Ruy. – Belo Horizonte, 2011.

xviii, 85 f: il; 210 x 297mm.

Bibliografia: f. 96 - 103

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. Doença de Chagas/genética 2. *Trypanosoma cruzi*/genética 3. Leishmaniose/genética 4. Leishmania/genética 5. Mapeamento de Interação de Proteínas/métodos I. Título. II. Ruiz, Jeronimo Conceição (Orientação).

CDD – 22. ed. – 616.936 3

Ministério da Saúde

Fundação Oswaldo Cruz

Centro de Pesquisas René Rachou

Programa de Pós-graduação em Ciências da Saúde

Análise computacional baseada no desenvolvimento de um *pipeline* de técnicas *ab initio* para predição de desordem estrutural protéica em genomas de tripanosomatídeos

por

Patrícia de Cássia Ruy

Foi avaliada pela banca examinadora composta pelos seguintes membros:

Prof. Dr. Jeronimo Conceição Ruiz (Presidente)

Prof. Dra. Silvane Maria Fonseca Murta

Prof. Dr. Christian Macagnan Probst

Suplente: Prof. Dra. Célia Maria Ferreira Gontijo

Dissertação defendida e aprovada em: 27/01/2011.

COLABORADORES

Faculdade de Medicina de Ribeirão Preto – USP/Ribeirão Preto
Dra. Angela Kaysel Cruz
Dr. Juliano Toledo

SUPORTE FINANCEIRO

Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq – nº 550399/2009-7).

Não sei...Se a vida é curta

Ou longa demais pra nós,

Mas sei que nada do que vivemos

Tem sentido, se não tocamos o coração das pessoas...

Cora Coralina

À Selma, Claudemir, Priscila e Pâmela por serem meu porto seguro.

AGRADECIMENTOS

Ao Dr. Jeronimo Conceição Ruiz, pela orientação científica, pela amizade de tantos anos e por ser o responsável pelo nascimento da paixão que tenho em bioinformática.

À Dra. Ângela Kaysel Cruz e ao Dr. Juliano Toledo, pela colaboração nas análises experimentais.

A todos os colegas do Laboratório de Parasitologia Celular e Molecular, pelo apoio e amizade.

À minha mãe e meu pai, pelo apoio e amor incondicionais.

Às minhas irmãs, por me fazerem acreditar que quando estamos juntas podemos tudo.

Ao meu avô Francisco e a minha avó Maria, por serem meus exemplos de vida.

À minha tia Lucéia, pelo apoio gigantesco e por agora ser minha segunda mãe.

À minha tia Cá, pelas conversas sempre muito produtivas.

Aos meus tios Elias e Guta, pelo apoio e pelas risadas.

Ao Raul, pelo apoio, paciência, amor, pelos dias e noites trabalhando e pelo conforto em saber que juntos conseguimos.

Ao Antonio e a Tatiana por me fazerem saber que não estava sozinha.

À Suélen, por ser minha irmã por escolha e sempre estar pronta para o que der e vier.

À Loislene, que mesmo a distância sempre teve muita paciência em me ouvir e ajudar.

Ao Antonio e a Daniela pela ajuda na leitura da dissertação.

Ao CNPq pela bolsa de estudos.

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.

SUMÁRIO

LISTA DE FIGURAS XII
LISTA DE TABELAS XII
LISTA DE GRÁFICOS XIII
LISTA DE ABREVIATURAS E SÍMBOLOS XV
LISTA DE DEFINIÇÕES XVI
RESUMO XVII
ABSTRACT XVIII
1 INTRODUÇÃO 19
1.1 IUPs (*Intrinsically Unstructured Proteins*) 19
1.2 Papel biológico das IUPs 21
1.3 Terminologia associada ao fenômeno 23
1.4 Predições de IUPs 24
1.5 A importância de um *pipeline* de execução e análise 26
1.6 Organismo modelo 27
1.6.1 Genomas dos organismos estudados 29
2 JUSTIFICATIVA 31
3 OBJETIVOS 33
3.1 Objetivo geral 33
3.2 Objetivos específicos 33
4 MATERIAIS E MÉTODOS 34
4.1 Download das seqüências 34
4.2 Linguagem de programação 34
4.3 *Pipeline* de IUPs 35
4.3.1 Visão geral 35
4.3.2 Funcionalidades 35
4.3.2.1 Criação da estrutura de diretórios 36
4.3.2.2 Pré-processamento das seqüências 37
4.3.2.3 Predição das IUPs 38
4.3.2.3.1 DisEMBL 38
4.3.2.3.2 IUPred 39
4.3.2.3.3 GlobPipe 39
4.3.2.3.4 VSL2B 40
4.3.2.4 Predição das regiões transmembranas e peptídeo sinal das IUPs preditas 40
4.3.2.5 Predição das características físico-químicas das IUPs preditas 40

4.3.2.6 Predição da localização subcelular das IUPs preditas..... 41
4.3.2.7 Predição da anotação funcional das IUPs predita..... 41
4.3.2.7.1 Gene Ontology..... 41
4.3.2.7.2 Blast2GO..... 42
4.3.2.7.3 Análise de enriquecimento funcional..... 43
4.3.2.8 Armazenamento dos resultados no banco de dados formado..... 43
4.3.2.9 Classificação hipotética ou predita..... 44
4.3.2.10 Criação das regiões consenso de predição e dos arquivos fasta contendo essas regiões formadas pelas diferentes combinações de preditores de IUPs..... 44
4.3.2.11 Calculo de estatísticas das regiões de IUPs preditas e armazenamento no banco de dados..... 45
4.3.2.12 Criação dos arquivos de relatório..... 45
4.3.2.12.1 Arquivo de análise descritiva..... 46
4.3.2.12.1.1 Calculo da freqüência de aminoácidos..... 47
4.3.2.12.2 Arquivo de análise de contingência..... 47
4.3.3 Estrutura necessária e execução do *pipeline* de IUPs..... 49
4.3.3.1 Pré-requisitos necessários a instalação do pipeline de IUPs.....49
4.3.4 Criação do banco de dados de IUPs..... 50
4.3.5 Criação do banco de dados GO_terms..... 55
4.3.6 Parâmetros obrigatórios na execução do *pipeline* de IUPs..... 55
4.3.7 Melhor combinação de preditores de desordem estrutural..... 56
4.3.8 Arquivo READ-ME..... 57
4.4 Testes..... 57
4.5 Limitações do *pipeline* de IUPs..... 58
4.6 Análise Experimental..... 58
5 RESULTADOS..... 60
5.1 Banco de dados de IUPs..... 60
5.2 Arquivos de análise descritiva..... 60
5.2.1 Pré-processamento..... 61
5.2.2 Identificação das IUPs..... 61
5.2.3 Caracterização das IUPs..... 62
5.2.3.1 Análise da freqüência de aminoácidos..... 63
5.2.3.2 Distribuição dos resíduos desordenados..... 66
5.2.3.3 Tamanho das regiões desordenadas..... 68
5.2.3.4 Predição da localização das IUPs..... 69
5.2.3.5 Número de domínios transmembranas das IUPs..... 69

5.2.3.6 Ancoramento das regiões desordenadas nas proteínas..... 71
5.2.3.7 Predição funcional das IUPs..... 71
5.2.3.8 Predição de características físico-químicas das IUPs..... 75
5.3 Arquivos de análise de contingência..... 76
5.4 Análise experimental..... 79
6 DISCUSSÃO..... 83
6.1 IUPs nos Tripanosomatídeos..... 83
6.2 *Pipelin*e de IUPs..... 84
6.3 Freqüência de aminoácidos..... 86
6.4 IUPs e suas funções nos tripanosomatídeos..... 87
6.5 Análise de contingência..... 89
6.6 Validação experimental das predições *in silico*..... 91
6.7 IUPs como alvos para o desenvolvimento de drogas..... 92
7 CONCLUSÕES..... 94
8 ANEXOS..... 95
8.1 Anexo 1 - README..... 95
9 REFERÊNCIAS BIBLIOGRÁFICAS..... 96

LISTA DE FIGURAS

Figura 1: Diferentes níveis de ordem e desordem.....20

Figura 2: Diagrama de contexto do *pipeline* de IUPs.....35

Figura 3: Diagrama entidade relacionamento do banco de dados de IUPs.....51

Figura 4: Análise do gel 2D de IUPs em *Leishmania ssp*.....81

LISTA DE TABELAS

Tabela 1: Informações dos genomas estudados.....30

Tabela 2: Tabela IUP do banco de dados de IUPs.....52

Tabela 3: Tabela DISORDER do banco de dados de IUPs.....52

Tabela 4: Tabela TRANSMEMBRANE do banco de dados de IUPs.....53

Tabela 5: Tabela GO_terms do banco de dados de IUPs.....53

Tabela 6: Tabela DISORDER_nr do banco de dados de IUPs.....54

Tabela 7: Tabela STATISTICS do banco de dados de IUPs.....54

Tabela 8: Parâmetros do *pipeline* de IUPs.....56

Tabela 9: Resultado do pré-processamento das seqüências no *pipeline* de IUPs para os cinco organismos estudados.....61

Tabela 10: Resultado do pré-processamento das seqüências no *pipeline* de IUPs para os cinco organismos estudados.....62

Tabela 11: Validação experimental *in silico* das IUPs preditas.....82

Tabela 12: Teste de validação do *pipeline* de IUPs.....84

LISTA DE GRÁFICOS

Gráfico 1: Número de publicações relacionadas ao termo desordem estrutural.....19

Gráfico 2: Número de resíduos desordenados.....62

Gráfico 3: Perfil de composição ordem/desordem de *L. braziliensis*.....63

Gráfico 4: Perfil de composição ordem/desordem de *L. major*.....64

Gráfico 5: Perfil de composição ordem/desordem de *L. infantum*.....64

Gráfico 6: Perfil de composição ordem/desordem de *T. brucei*.....65

Gráfico 7: Perfil de composição ordem/desordem de *T. cruzi*.....65

Gráfico 8: Distribuição de resíduos desordenados em *L. braziliensis*.....66

Gráfico 9: Distribuição de resíduos desordenados em *L. major*.....67

Gráfico 10: Distribuição de resíduos desordenados em *L. infantum*.....67

Gráfico 11: Distribuição de resíduos desordenados em *T. brucei*.....67

Gráfico 12: Distribuição de resíduos desordenados em *T. cruzi*.....68

Gráfico 13: Distribuição de resíduos desordenados em *L. braziliensis*.....68

Gráfico 14: Localização subcelular das IUPs em *L. braziliensis*.....69

Gráfico 15: Número de regiões transmembranas nas IUPs de *L. braziliensis*.....70

Gráfico 16: Número de regiões transmembranas nas IUPs de *T. brucei*.....70

Gráfico 17: Localização das regiões desordenadas nas proteínas.....71

Gráfico 18: 20 primeiros termos GO enriquecidos da categoria componente celular das IUPs de *L. braziliensis*.....72

Gráfico 19: 20 primeiros termos GO enriquecidos da categoria função molecular das IUPs de *L. braziliensis*.....73

Gráfico 20: 20 primeiros termos GO enriquecidos da categoria processo biológico das IUPs de *L. braziliensis*.....74

Gráfico 21: Carga das IUPs de *L. braziliensis*.....75

Gráfico 22: Ponto isoelétrico das IUPs de *L. braziliensis*.....75

Gráfico 23: Associação entre porcentagem de resíduos desordenados e a anotação da proteína como com função predita ou hipotética nas IUPs de *L. braziliensis*.....76

Gráfico 24: Associação entre porcentagem de resíduos desordenados com o ponto isoelétrico nas IUPs de *L. braziliensis*.....77

Gráfico 25: Associação entre porcentagem de resíduos desordenados com a localização predita para as IUPs de *L. braziliensis*.....78

Gráfico 26: Associação entre porcentagem de resíduos desordenados com a localização predita para as IUPs de *L. braziliensis*.....79

LISTA DE ABREVIATURAS E SÍMBOLOS

aa: aminoácidos

Da: abreviação da unidade de medida de peso molecular Dalton

DER: Diagrama de Entidades e Relacionamentos

GO: *Gene Ontology*

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

Kb: kilobase

pb: par de bases

MER: Modelo Entidade Relacionamento

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

PDB: *Protein Data Bank*

ROC: *Receiver Operating Characteristics*

SQL: *Structured Query Language*

XML: *eXtensible Markup Language*

Gly, Gli ou G: Glicina

Ala ou A: Alanina

Leu ou L: Leucina

Val ou V: Valina

Ile ou I: Isoleucina

Pro ou P: Prolina

Phe, Fen ou F: Fenilalanina

Ser ou S: Serina

Thr,The ou T: Treonina

Cys, Cis ou C: Cisteina

Tyr, Tir ou Y: Tirosina

Asn ou N: Asparagina

Gln ou Q: Glutamina

Asp ou D: Aspartato ou Ácido Aspártico

Glu ou E: Glutamato ou Ácido Glutâmico

Arg ou R: Arginina

Lys, Lis ou K: Lisina

His ou H: Histidina

Trp ou Tri: Triptofano

Met ou M: Metionina

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*.

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 é 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.

Script: uma série de instruções formais escritas para um interpretador.

RESUMO

Proteínas são compostas por uma ou mais cadeias de aminoácidos e exibem vários níveis de organização estrutural. Recentemente, uma classe de proteínas conhecidas como IUPs (*Intrinsically Unstructured Proteins*) foi descoberta e sua principal característica é a ausência de estrutura parcial ou total em seu estado nativo. Devido à sua adaptabilidade intrínseca, tais proteínas participam em muitos processos biológicos regulatórios incluindo o escape do sistema imune.

Utilizando a informação contida no proteoma predito de *Leishmania braziliensis, Leishmania major, Leishmania infantum, Trypanosoma cruzi* e *Trypanosoma brucei*, desenvolvemos um *pipeline* de análise computacional que tem como objetivo a identificação, caracterização e análise de IUPs. Nosso principal objetivo é investigar as correlações biológicas entre desordem estrutural e as interações parasito-hospedeiro. O *pipeline* emprega 6 metodologias de predição de desordem, integra informações obtidas através da anotação estrutural e funcional, predição subcelular e o cálculo de propriedades físico-químicas. Como cerne do *pipeline* de IUPs existe um banco de dados relacional modelado de forma a viabilizar a extração das relações entre as inúmeras variáveis analisadas e gerar os relatórios.

Nossos resultados demonstram que as espécies de *Leishmania* e *Trypanosoma* possuem aproximadamente 70% e 55% de IUPs, respectivamente. Nossos resultados indicam que nos tripanosomatídeos os aminoácidos promotores de ordem são: W, Y, F, V, I, L e C e os promotores de desordem são P, Q, E, R e S.

A anotação funcional revelou o enriquecimento dos termos GO: *transcription, ribonucleoproteins, RNA metabolic process, protein binding* e *ribonucleotide binding*. Outra característica é a associação entre o aumento da porcentagem de resíduos desordenados, a localização subcelular e o número de regiões transmembrana.

Como resultado da validação experimental feita através de técnicas de eletroforese bidimensional (2D) específicas para identificação de IUPs, 100% dos *spots* identificados foram preditos *in silico*.

Até o presente momento este trabalho representa a primeira tentativa de estabelecimento das correlações entre desordem estrutural protéica e função nos tripanosomatídeos. A execução do pipeline pode ser solicitada por instituições acadêmicas através de solicitação no sitio http://iup.cpqrr.fiocruz.br/iup-pipeline.

ABSTRACT

Proteins are composed of one or more chains of amino acids, and exhibit several levels of structure. Recently, a class of proteins called IUPs (Intrinsically Unstructured Proteins) has been discovered that do not fold into any particular configuration existing as dynamic ensembles in their native state. Due to their intrinsic adaptability, they participate in many regulatory biological processes including parasite immune escape.

Using the information from Leishmania braziliensis, Leishmania major, Leishmania infantum, Trypanosoma cruzi and Trypanosoma brucei proteomes we developed a pipeline aiming the identification, characterization and analysis of IUPs, our main goal is to establish the biological correlations between protein structural disorder and host-parasite interactions. The pipeline employs 6 methodologies of disorder prediction, integrates information obtained through the structural and functional annotation, subcellular prediction and physicochemical properties. As the core of the IUP pipeline there is a relational database modeled to enable the extraction of relations between the numerous variables and generate reports.

The results demonstrate that *Leishmania* and *Trypanosoma* species has approximately 70% and 55% of IUPs respectively. Our results indicate that in tripanosomatides IUPs have disorder promoter amino acids (P, Q, E, R and S) and order promoter amino acids (W, Y, F, V, I, L and C).

The functional annotation pointed the enrichment of the following GO terms: transcription, ribonucleoproteins, RNA metabolic process, protein binding and ribonucleotide binding, among others. Another characteristic is the association between the increase of disordered residues percent with the nuclear subcelullar localization and the lack of transmembrane regions.

We made an experimental validation through 2D electrophoresis that identifies IUPs and our results revealed that 100% of the identified spots were predicted *in silico*.

Since there is no pipeline or databases addressing this issue this IUP pipeline represents the first attempt to establish the correlations between protein function and structural disorder. The execution of pipeline can be requested by academic institutions at http://iup.cpqrr.fiocruz.br/iup-pipeline. .

1 INTRODUÇÃO

1.1 IUPs (*Intrinsically Unstructured Proteins*)

A visão tradicional da relação estrutura/função tem como um de seus princípios a concepção de que a função biológica de uma dada proteína é criticamente dependente de uma estrutura conformacional bem definida.

As bases experimentais que conduziram o estabelecimento da inter-relação entre o papel funcional de uma determinada proteína e a sua estrutura tridimensional ocorreu há mais de 100 anos e tem a proposta “chave-fechadura” feita por Emil Fischer (Fischer, 1894) como um desdobramento altamente plausível do conhecimento da época. Baseado na especificidade da ação enzimática, o conceito proposto por Fischer é amplamente aceito na biologia molecular, e estabelece que o centro ativo de uma enzima tem uma determinada estrutura que é capaz de interagir de forma complementar a estrutura do substrato. O substrato corresponderia à chave e o centro ativo à fechadura.

Apesar da existência de estudos que datam de mais de uma década (Lynch, Riseman *et al.*, 1987) (Barlow, Vidal *et al.*, 1988) demonstrando a existência de proteínas e domínios protéicos sem estrutura definida, que desempenhavam suas funções biológicas, mas que não eram capazes de manter uma conformação tridimensional estável em condições fisiológicas, somente recentemente a generalidade do fenômeno foi notada (Wright e Dyson, 1999), fato que pode ser constatado inclusive pelo expressivo aumento de publicações nos cinco últimos anos (Gráfico 1).

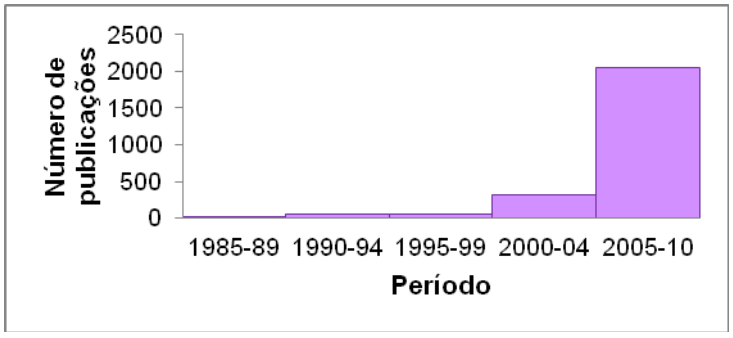


Gráfico 1: Número de publicações relacionadas ao termo desordem estrutural. Foram utilizados os seguintes termos em busca feita no Pubmed (http://www.ncbi.nlm.nih.gov/pmc/): *intrinsically disordered proteins*, *intrinsically unfolded proteins*, *intrinsically unstructured proteins* e *natively unfolded proteins*. Pesquisa feita em 21/10/2010.

A desordem estrutural pode se manifestar em vários contextos, afetando vários níveis da estrutura da proteína (Figura 1). De fato, estudos realizados por Obradovic e colaboradores (Obradovic, Peng *et al.*, 2003) com dados de estrutura de proteínas presentes no PBD (*Protein Data Bank*) (Berman, Bhat *et al.*, 2000) mostraram que somente 32% das estruturas cristalizadas são completamente desprovidas de desordem, o que indica que essas representam a exceção e não a regra. Alguns autores (Dunker e Obradovic, 2001) inclusive têm sugerido que devido à existência de desordem protéica intrínseca deveria haver uma reavaliação do paradigma proteína-estrutura-função.

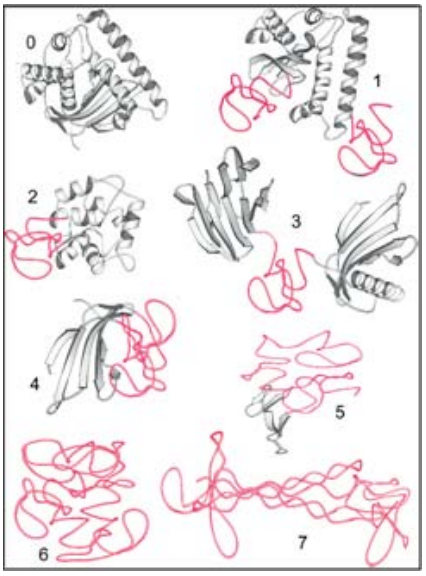


Figura 1: Diferentes níveis de ordem e desordem. (0) nenhuma desordem; (1) desordem N e C terminal; (2) desordem de ligação; (3) loop desordenado; (4) domínio desordenado; (5) proteína desordenada com alguns resíduos ordenados; (6) completamente desordenada, sobretudo dobrada e (7) completamente desordenada e estendida. (Figura adquirida de Uversky, 2005).

Outra característica particular das IUPs, que surge como conseqüência da sua "falta" de estrutura definida, está associada à elevada acessibilidade de suas cadeias polipeptídicas, propiciando o envolvimento dessa classe de proteínas em extensivas modificações pós-traducionais como fosforilações, acetilações, ubiquitinações, o que permite a modulação de sua função ou atividade biológica (Ishida, Hara *et al.*, 2002) (Blain e Massagué, 2002) (Tsvetkov, Yeh *et al.*, 1999).

Análises *in silico* feitas para genomas completos de eucariotos pela utilização de algoritmos do tipo *disorder predictor* indicam que 36 a 63% do produto dos genes codificadores de proteínas têm estruturas desordenadas na extensão total ou em

segmentos (maior que 40 resíduos) de suas seqüências (Oldfield, Cheng *et al.*, 2005) (Dunker, Obradovic *et al.*, 2000) (Uversky, Gillespie *et al.*, 2000). Para bactéria e *Archaea*, as predições de regiões desordenadas variam de 7 a 33% e 9 a 37%, respectivamente (Dunker, Obradovic *et al.*, 2000), o que mostra que as proteínas desordenadas estão mais presentes em organismos multicelulares do que nos organismos unicelulares. Esse comportamento pode ser explicado pela alta flexibilidade desse tipo de proteína que pode proporcionar uma melhor resposta a mudanças do ambiente do que proteínas rígidas (Dunker, Lawson *et al.*, 2001) e por uma maior necessidade de sinalização e regulação, que são funções associadas às IUPs (mostradas no item 1.2), em células nucleadas (Dunker e Obradovic, 2001) (Dyson, H. e Wright, P., 2002) (Iakoucheva, Brown *et al.*, 2002).

Devido à plasticidade ocasionada pela desordem estrutural, a proteína que contém trechos desordenados adota diferentes estruturas e isso acontece através da interação com diferentes ligantes, como por exemplo, a p53 que é uma proteína com aproximadamente 50% de desordem, interage com um grande número de ligantes diferentes (Russell e Gibson, 2008) (Oldfield, Meng *et al.*, 2008). Devido à capacidade de interação com múltiplos ligantes, vários autores têm sugerido que proteínas do tipo IUP são centrais (proteínas do tipo *hub*) em redes de interação proteína-proteína (Oldfield, Meng *et al.*, 2008) (Haynes e Iakoucheva, 2006) (Dunker, Cortese *et al.*, 2005) (Dyson e Wright, 2005).

Vale ainda ressaltar que para o desempenho das suas funções biológicas, as IUPs necessitam da interação/ligação com outras biomoléculas e esse processo envolve uma transição *disorder-to-order* que é fundamental para que elas possam atuar em diferentes vias de regulação e controle (Kriwacki, Hengst *et al.*, 1996) (Dyson, H. e Wright, P., 2002) (Lacy, Filippov *et al.*, 2004).

1.2 Papel biológico das IUPs

Apesar de pequenos polipeptídios bioativos e flexíveis em termos estruturais como os hormônios serem conhecidos há muito tempo (Boesch, Bundi *et al.*, 1978), o fato das IUPs estarem envolvidas em uma ampla gama de papéis biológicos somente recentemente foi reconhecido (Dyson e Wright, 2005).

O papel biológico das IUPs inclui regulação da divisão celular, transcrição e tradução, transdução de sinais, fosforilação protéica, atividade chaperona e regulação da montagem de grandes complexos protéicos como o do ribossomo (Dyson e Wright, 2005) (Dunker, Brown *et al.*, 2002) (Namba, 2001) (Tompa, 2002) (Tompa e Csermely, 2004) (Uversky, Permyakov *et al.*, 2002) (Wright e Dyson, 1999).

Em 2008 Russell e colaboradores (Russell e Gibson, 2008) argumentaram que as IUPs freqüentemente estão envolvidas em eventos de regulação ou controle mais proeminentes em organismos complexos e raramente aparecem envolvidas em processos *housekeeping* como metabolismo. Em outras palavras, proteínas em vias de sinalização celular contêm mais desordem estrutural do que proteínas envolvidas em processos metabólicos.

Além disso, estudos recentes demonstraram que 79% das proteínas humanas associadas ao câncer (HCAPs – *Human cancer-associated proteins*) e 47% de todas as proteínas eucarióticas no banco de dados SWISS-PROT associadas aos mecanismos de tumorigênese podem ser classificadas como IUPs, fato que evidencia o importante envolvimento dessas proteínas nos estados de saúde e doença (Galea, Wang *et al.*, 2008).

Outro ponto relevante está centrado na participação das IUPs na interação parasita/hospedeiro, onde a plasticidade estrutural das IUPs, a qual permite interações promíscuas, pode favorecer a sobrevivência do parasita, tanto pela inibição da geração de respostas efetivas de anticorpos de alta afinidade, quanto por facilitar a interação com moléculas do hospedeiro necessárias à ligação e invasão de células hospedeiras (Feng, Zhang *et al.*, 2006). Há uma alta incidência de desordem nas proteínas envolvidas com doenças, como por exemplo, o HIV do tipo 1 (agente causador da AIDS) que possui uma pequena proteína de RNA-biding denominada Tat (*transactivator of transcription*). Essa proteína é uma IUP e possui um papel central na regulação da replicação do HIV-1 (Shojania e O'neil, 2006).

Outro exemplo é a MSP2, uma das proteínas de superfície mais abundantes no merozoíto de *Plasmodium falciparum*, que está sendo usada para o desenvolvimento de vacina para a malária. A MSP2 é uma IUP envolvida na ligação do merozoíto aos eritrócitos do hospedeiro, sendo que o processo envolve uma transição para uma estrutura mais ordenada (Feng, Zhang *et al.*, 2006).

Outra doença a qual as IUPs estão envolvidas é o Alzheimer que possui uma proteína precursora da NAC (*non-A beta component of Alzheimer's disease amyloid plaque*) a NACP que é um membro de uma família altamente conservada termoestáveis específicas do cérebro que têm seu envolvimento sugerido na formação e/ou estabilização da sinapse que é uma representante da classe de proteínas de IUPs (Weinreb, Zhen *et al.*, 1996).

1.3 Terminologia associada ao fenômeno

Vários termos vêm sendo utilizados para descrever proteínas e/ou regiões de proteínas incapazes de formar estruturas 3D específicas, incluindo *partially folded* (Linderstrøm-Lang e Schellman, 1959), *flexible (Pullen, Jenkins et al., 1975)* e *mobile* (Cary, Moss *et al.*, 1978). Tais terminologias não são inteiramente adequadas uma vez que também podem ser associadas a proteínas normalmente estruturadas. Como exemplo da aplicação, sem critérios bem estabelecidos dessa terminologia, temos proteínas ordenadas que apresentam elevado fator-B (o fator-B de um cristal de proteína reflete a flutuação dos átomos sobre sua posição média e proporcionam uma importante informação sobre a dinâmica da proteína (Yuan, Bailey *et al.*, 2005)) e que freqüentemente são chamadas de *flexible* ou *mobile*.

Como conseqüência desse desencontro nas definições, surgiram outras terminologias para descrever o fenômeno da desordem estrutural, entre eles: *intrinsically unstructured* (Wright e Dyson, 1999), *intrinsically disordered* (Dunker, Lawson *et al.*, 2001) e *natively disordered* (Daughdrill, Pielak *et al.*, 2005), além de várias combinações desses diferentes termos estão presentes na literatura.

Nenhuma dessas terminologias foi adotada universalmente entre os diferentes pesquisadores e esse ainda é um campo efervescente de discussão e reflete o rápido crescimento dos estudos de desordem estrutural. Nessa dissertação, decidimos adotar a utilização do termo IUPs (*Intrinsically Unstructured Proteins*) para definir proteínas que são inteiramente ou possuem regiões (maior ou igual a 40 aminoácidos) desordenadas, ou seja, que possuem desordem estrutural.

1.4 Predições de IUPs

Em decorrência da importância funcional das IUPs várias metodologias computacionais e diferentes ferramentas de análise têm sido desenvolvidas objetivando a identificação sistemática do fenômeno de desordem estrutural em diferentes genomas (Bracken, Iakoucheva *et al.*, 2004).

Apesar da existência de inúmeros algoritmos, a identificação computacional de IUPs ainda representa um grande desafio. Além da ausência de uma definição consensual adequada sobre como identificar uma IUP, vários outros parâmetros físicos contribuem para o problema. Em particular a definição termodinâmica de desordem para uma cadeia polipetídica está associada a um estado estrutural do tipo de *random coil* que pode ser melhor entendido como um conjunto estrutural onde todos os peptídeos da cadeia polipeptídica adotam todos os graus de liberdade possíveis. Entretanto, mesmo sob condições extremas de desnaturação como uréia 8M, tal estado teórico não é observado nas proteínas solvatadas (Shortle e Ackerman, 2001) (Ackerman e Shortle, 2002) (Klein-Seetharaman, Oikawa *et al.*, 2002). Em solução tais proteínas parecem ainda manter certo grau de estrutura (Linding, Jensen *et al.*, 2003).

Por outro lado a desordem protéica é apenas indiretamente observada por uma variedade de métodos experimentais: a) cristalografia de Raio X com a desordem sendo indicada por resíduos perdidos dos mapas de densidade eletrônica; b) ressonância magnética nuclear com a indicação de desordem ocorrendo pela presença de picos agudos, pela ausência da característica NOEs (*Nuclear Overhauser effect*) de estrutura secundária ou por valores negativos para NOEs heteronucleares de ^{1}H-^{15}N; c) dicroísmo circular com a desordem sendo indicada pela baixa intensidade de ~ 210 até ~ 240 nm; d) digestão por protease com indicação de desordem por sítios de hipersensibilidade; e e) determinação de valores hidrodinâmicos, onde raios de Stoke atipicamente grandes para um determinado peso molecular indicam desordem estrutural (Romero, Obradovic *et al.*, 2001) (Smyth, Syme *et al.*, 2001) (Dunker, Lawson *et al.*, 2001). Cada um desses métodos detecta diferentes aspectos de desordem, resultando em diferentes definições de desordem protéica (Tompa, 2002).

Como características gerais de composição, as proteínas classificadas como IUPs possuem regiões de baixa complexidade (que são raramente encontradas em proteínas com estruturas 3D), de baixo conteúdo em aminoácidos hidrofóbicos e conseqüentemente elevada característica polar e de aminoácidos carregados (Romero, Obradovic *et al.*, 2001) (Vucetic, Brown *et al.*, 2003). Tais características são consistentes com a inabilidade dessas proteínas adotarem a conformação globular e têm levado ao desenvolvimento de diversos algoritmos capazes de predizer regiões desordenadas de proteínas (Romero, Obradovic *et al.*, 2001) (Uversky, Gillespie *et al.*, 2000) (Linding, Jensen *et al.*, 2003) (Ferron, Longhi *et al.*, 2006) (Dosztányi, Mészáros *et al.*, 2010) (Oldfield, Cheng *et al.*, 2005).

Além dessas características, outras peculiaridades têm se mostrado eficientes para a identificação de desordem estrutural protéica, como a preferência por determinados aminoácidos e alta variabilidade desses aminoácidos nas seqüências (Dunker, Lawson *et al.*, 2001) (Linding, Jensen *et al.*, 2003) (Brown, Takayama *et al.*, 2002). Dada a grande diversidade de características, abordagens computacionais eficientes ainda são relativamente escassas e apresentam algoritmos incapazes de predizer *in silico* todas as características acima descritas. Como decorrência dessa constatação fica claro que abordagens combinatoriais devem fornecer respostas mais completas ao problema da predição de IUPs (Ferron, Longhi *et al.*, 2006).

Além dos pontos expostos acima, estudos de NMR (*Nuclear Magnetic Resonance*) têm mostrado que o grau de desordem das IUPs varia enormemente indo desde a completa ausência de estrutura secundária ou terciária (Abercrombie, Kneale *et al.*, 1978) (Penkett, Redfield *et al.*, 1997) até a presença de estruturas secundárias parciais (Bai, Chung *et al.*, 2001) (Uversky, Permyakov *et al.*, 2002) (Wright e Dyson, 1999) o que dificulta ainda mais as predições.

Devido a essa complexidade na predição dessas regiões, em 2009, nosso grupo propôs a utilização de gráficos ROC (*Receiver Operating Characteristic*) (Green e Swets, 1966) para identificar a melhor combinação de preditores de desordem estrutural (Torrieri, 2009).

Assim, neste trabalho utilizamos a implementação de gráficos ROC desenvolvida anteriormente no estabelecimento de um *pipeline* para análise computacional empregando técnicas *ab initio* para predição de desordem estrutural protéica em genomas de tripanosomatídeos. Essa metodologia integra em um único

pacote predições de desordem estrutural, anotação funcional, localização sub-celular, características físico-químicas, regiões transmembrana e peptídeo sinal em um banco de dados relacional modelado para possibilitar a extração das complexas inter-relações entre todas essas variáveis. Pela utilização desse banco de dados relatórios de análise descritiva e de contingência são gerados automaticamente. O detalhamento das etapas desse *pipeline* é feito no item materiais e métodos desta dissertação.

1.5 A importância de um *pipeline* de execução e análise

Um ponto chave na pesquisa científica "moderna" é a integração da informação gerada visando o entendimento da biologia dos organismos estudados. A grande diversidade de análises computacionais e experimentais e o grande volume de dados gerados por aplicações de tecnologias *high-throughput* rotineiramente impossibilitam o estabelecimento manual das relações entre as variáveis envolvidas e conseqüentemente a correta interpretação do fenômeno estudado.

Os dados gerados, via de regra, passam por inúmeras análises descritivas e comparativas que idealmente são integradas via utilização de um banco de dados relacional empregando um modelo entidade-relacionamento capaz de responder complexas questões.

Se por um lado o fluxo de análises é freqüentemente interdependente, elas idealmente devem ser correlacionadas e isso é feito através da implementação de *pipelines* de análise integrados a banco de dados relacionais, por vezes bastante complexos, e que viabilizam a recuperação integrada dos dados para posterior análise.

Particularmente, um *pipeline* computacional de análise de seqüências biológicas tem como característica fundamental a rígida padronização na execução de suas etapas. Tal padronização garante que as análises sejam realizadas com os mesmos critérios estabelecidos durante sua concepção, em diferentes contextos e sem erro humano. Essa característica implica na reprodutibilidade das análises e representa uma importante premissa necessária em todo experimento científico.

Outro aspecto importante de um *pipeline* está relacionado à sua facilidade de execução para o usuário final onde idealmente todas as etapas são executadas de forma transparente e fácil.

Além dos pontos expostos acima relacionados às características desejáveis de um *pipeline* de análise de seqüências vale ressaltar que um *pipeline* robusto desenvolvido para manipulação de dados em larga escala deve analisar os resultados e gerar relatórios, utilizando técnicas estatísticas e/ou de mineração de dados e que conseqüente vão além dos relatórios de saída dos diferentes algoritmos de análise empregados na sua execução. Testes estatísticos simples, como o Qui-Quadrado, por exemplo, são capazes de elucidar a relação entre duas variáveis através da análise de uma tabela de contingência (David C. Howell: *Chi-Square Test - Analysis of Contingency Tables http://www.uvm.edu/~dhowell/methods8/Supplements/ChiSquareTests.pdf*) e integrar novas perspectivas às análises.

Dado o enorme volume de dados inicialmente analisados (5 genomas completos, 45.149 proteínas), essa dissertação de mestrado teve como objetivo central o desenvolvimento de um *pipeline* especializado e integrado a um banco de dados para identificação, caracterização e análise em larga escala de desordem estrutural (Fiers, Van Der Burgt *et al.*, 2008).

1.6 Organismo modelo

A família *Trypanosomatidae* pertence ao reino *Protista*, sub-reino *Protozoa*, filo *Sarcomastigophora*, sub-filo *Mastigophora*, classe *Zoomastigophorea* e a ordem *Kinetoplastida,* representa um grupo de parasitas unicelulares, flagelados (com flagelo único) que abrange importantes patógenos de humanos e animais e inclui organismos do gênero *Leptomonas*, *Leishmania*, *Phytomonas*, *Crithidia*, *Blastocrithidia*, *Herpetomonas* e *Trypanosoma*, sendo que dois desses gêneros possuem importância médica: o gênero *Leishmania* e o *Trypanosoma*.

Os tripanosomas africanos (*Trypanosoma brucei*, *T. congolense* e *T. vivax*) são endêmicos em áreas rurais do deserto da África sub-Sahariana, onde causam a doença do sono em humanos que é fatal se não tratada (Aslett, Aurrecoechea *et al.*,

2010) (Cox, 2004). Em 2004, 17.580 casos de infecções humanas foram reportados, sendo que juntamente com os casos não registrados de áreas pobres e rurais, a taxa de infecção real pode atingir 300.000 novos casos ao ano (WHO, 2009 - *The World Health Organization. Human African trypanosomiasis (sleeping sickness): epidemiological update*) (Aslett, Aurrecoechea *et al.*, 2010).

O *Trypanosoma cruzi* é endêmico nas Américas Central e do Sul, causando a doença de Chagas em aproximadamente 8-9 milhões de infecções e 14.000 mortes ao ano (WHO, 2002) (Hotez, Bottazzi *et al.*, 2008) (Aslett, Aurrecoechea *et al.*, 2010) com aproximadamente 50.000 novos casos por ano (Senior, 2007).

Os parasitos do gênero *Leishmania* provocam doenças infecto-parasitárias conhecidas como leishmanioses, as quais acometem o homem causando um amplo espectro de manifestações clínicas. Tais manifestações dependem da espécie infectante de *Leishmania* e da relação parasita-hospedeiro e variam de lesões cutâneas auto-cicatrizantes a distúrbios viscerais que podem levar à morte. A *Leishmania ssp* infecta aproximadamente 12 milhões de pessoas, com 2 milhões de novos casos registrados anualmente (WHO, 2009 / http://www.who.int/leishmaniasis/burden/magnitude/burden_magnitude/en/index.html) (Aslett, Aurrecoechea et al., 2010).

Até hoje, somente três vacinas estão licenciadas para uso: uma vacina atenuada para humanos no Uzbequistão, uma vacina de antígenos totais para a imunoterapia humana no Brasil e uma vacina de segunda geração para a profilaxia de cães no Brasil (Palatnik-De-Sousa, 2008). Além disso, o número de drogas disponíveis para tratamento da doença é limitado e o fenômeno de resistência a drogas tem sido observado com freqüência (El-On, 2009).

Além dos pontos levantados acima, os quais evidenciam a relevância do estudo desses parasitas no contexto da saúde pública, diferentes estudos na área de genômica, proteômica e bioinformática têm sido realizados na tentativa de melhor elucidar as características biológicas desses organismos. A pesquisa relacionada à organização estrutural e funcional peculiar desses protozoários pleomórficos tem ajudado no entendimento dos mecanismos envolvidos no controle da transcrição e nos determinantes moleculares que controlam a vida intracelular deste parasita, ainda não totalmente esclarecidos, e que podem estar relacionadas às diferentes formas de patogênese.

Devido à importância médica e para a saúde pública acima citadas dos tripanosomatídeos, neste projeto utilizaremos a informação do conteúdo genômico de *L. braziliensis*, *L. major*, *L. infantum*, *T. cruzi* e *T. brucei* em uma ampla análise comparativa *in silico* visando a identificação e caracterização de proteínas do tipo IUP.

1.6.1 Genomas dos organismos estudados

Com o objetivo de descrever a origem dos dados utilizados neste trabalho (proteomas preditos) inserimos uma breve descrição dos projetos genoma que geraram a informação aqui analisada.

O projeto genoma de *L. major* (que integrou laboratórios da Inglaterra, França, Estados Unidos, Canadá e Brasil) foi finalizado em janeiro de 2005 e é sem dúvida um marco para o nosso conhecimento desse parasito que causa leishmaniose cutânea. Esse projeto foi idealizado pela OMS com o intuito de integrar estudos realizados por diferentes grupos de pesquisa e acelerar a compreensão da biologia do parasito e suas interações com os diferentes hospedeiros e visava ao desenvolvimento de vacinas, drogas e terapias mais eficazes.

L. infantum foi a segunda espécie de *Leishmania* a ser seqüenciada. Esse parasito que é responsável pela leishmaniose visceral na América do Sul e região do Mediterrâneo foi finalizado em 2007 pelo *Sanger Institute* em colaboração com Debbie Smith da Universidade de York e Jeremy Mottram da Universidade de Glasgow.

O genoma de *L. braziliensis*, agente causador de leishmaniose muco-cutânea, foi seqüenciado e anotado em 2007 pelo *Sanger Institute* em colaboração com Ângela Cruz da Universidade de São Paulo (FMRP – USP) e Debbie Smith da Universidade de York. A seqüência consenso desse genoma foi obtida através de *whole shotgun genome* com cobertura de aproximadamente 5x.

Em julho de 2005, os genomas de *T. brucei, L. major* e *T. cruzi* foram depositados em bancos de dados de domínio público e nesse mesmo ano houve a publicação de um estudo comparativo de conteúdo gênico e arquitetura desses três parasitos (El-Sayed, Myler *et al.*, 2005).

A tabela 1 apresenta as principais características dos cinco genomas estudados.

Tabela 1: Informações dos genomas estudados. Dados obtidos do TriTrypDB em novembro de 2010 (http://tritrypdb.org/tritrypdb/).

Organismo	Tamanho do genoma	Número de cromossomos	Número de proteínas preditas
L. braziliensis	32 Mb	35	8310
L. major	32.8 Mb	36	8408
L. infantum	32 Mb	36	8216
T. cruzi	67 Mb	41	10320
T. brucei	26 Mb	11	9895

Através da tabela 1, podemos observar que neste projeto, foram analisados um total de 189.8 Mb, divididos em 159 cromossomos e 45149 proteínas preditas.

2 JUSTIFICATIVA

Proteínas do tipo *intrinsically unstructured* são uma classe de proteínas recentemente reconhecidas (Dyson e Wright, 2005) e atuantes em uma ampla gama de papéis biológicos importantes. Em *P. falciparum*, no único trabalho de caracterização desse tipo de proteínas publicado até o momento em protozoários parasitos (Feng, Zhang *et al.*, 2006), foi demonstrada a correlação dessas proteínas com os processos de adesão e invasão celular.

Além do envolvimento direto das proteínas do tipo IUP nos processos de adesão e invasão celular demonstrados por Feng (Feng, Zhang *et al.*, 2006), há ainda outros fatores importantes que merecem destaque. Russel e Cheng demonstraram que a presença de desordem estrutural em uma proteína representa um facilitador para o desenho racional de drogas devido à alta especificidade e baixa afinidade de interação apresentada pelas IUPs (Russell e Gibson, 2008) (Cheng, Legall *et al.*, 2006). Além disso, a identificação de IUPs em larga escala tem sido utilizada como método de triagem para a seleção de novos potenciais alvos de pequenas moléculas (Cheng, Legall *et al.*, 2006).

Os pontos levantados acima demonstram a importância do estudo integrado mais profundo dessa intrigante classe de proteínas, que pode levar a um melhor entendimento dos mecanismos de interação parasito/hospedeiro e, em última instância, a um novo modelo para controle das doenças parasitárias como a leishmaniose, a doença do sono e a doença de Chagas.

Decifrar a rede regulatória de eventos associados ao parasitismo é um grande desafio da parasitologia, biologia molecular e computacional que somente poderá ser alcançado pela utilização de abordagens integradas de estudo.

Nesse contexto, um dos enfoques principais do projeto é o estabelecimento de um *pipeline* computacional que viabilize a identificação e caracterização *in silico* de IUPs nos proteomas preditos de vários organismos, entre eles, *L. infantum*, *L. major*, *L. braziliensis*, *T. cruzi* e *T. brucei*.

Uma vez que não existe *pipeline* algum de caracterização de IUPs estabelecido até o momento, tão pouco existe uma metodologia integrativa para predição de

desordem estrutural protéica em genomas, propomos assim o desenvolvimento de um *pipeline* computacional capaz de realizar tais tarefas.

Vale ainda ressaltar a ausência completa da informação sobre proteínas com desordem estrutural na anotação dos genomas estudados e assim o projeto deverá também contribuir agregando mais essa camada de anotação a esses genomas.

3 OBJETIVOS

3.1 Objetivo geral

Desenvolvimento de um *pipeline* computacional de análise de desordem estrutural protéica.

3.2 Objetivos específicos

Estudar os principais algoritmos preditores de desordem estrutural protéica e avaliar o desempenho desses algoritmos frente ao proteoma predito de tripanosomatídeos;

Interagir analises computacionais em um banco de dados relacional;

Implementar programas capazes de realizar análises descritivas automatizadas dos dados contidos no banco de dados desenvolvido;

Utilizar o *pipeline* desenvolvido na analise de proteomas preditos de tripanosomatídeos

Validar experimentalmente por espectrometria de massas de potencias proteínas desordenadas.

Padronizar e definir os requisitos para instalação e execução do *pipeline* desenvolvido.

4 MATERIAIS E MÉTODOS

Todos os programas citados abaixo, exceto quando explicitado, deverão ser instalados em sistema operacional Linux http://www.linorg.usp.br (Unix para PC's).

4.1 Download das seqüências

A versão 2.3 dos proteomas de *Leishmania braziliensis*, *Leishmania major*, *Leishmania infantum*, *Trypanosoma brucei* e *Trypanosoma cruzi* foram obtidas através do site do TriTrypDB (http://tritrypdb.org/tritrypdb/).

4.2 Linguagem de programação

Para o desenvolvimento do *pipeline* de IUPs e da estrutura necessária a sua execução foram utilizados a linguagem de programação Perl (*Practical Extraction and Report Language*) disponível no site http://www.perl.org/ e o sistema de gerenciamento de banco de dados MySQL disponível em http://www.mysql.com que utiliza a liguagem SQL (*Structured Query Language*).

A linguagem Perl foi escolhida pela riqueza de códigos Perl existentes para a bioinformática, a capacidade de integração do código com sistemas baseados em Unix, além da existência de uma grande variedade de módulos disponíveis no CPAN (*Comprehensive Perl Archive Network* - http://www.cpan.org/) e no BioPerl que é uma comunidade que dedica-se à criação de bibliotecas de código aberto de módulos de pesquisa em bioinformática (http://www.bioperl.org/).

O sistema de gerenciamento de banco de dado MySQL foi escolhido por ser um dos sistemas de gerenciamento de banco de dados mais populares e de código aberto, está disponível para sistemas operacionais Unix, além da alta performance e de se integrar muito bem a programas Perl.

4.3 *Pipeline* de IUPs

4.3.1 Visão geral

Desenvolvemos um *pipeline* automatizado que utilizando diferentes metodologias de predição, proporciona a identificação e caracterização das proteínas do tipo IUP em diferentes organismos.

O *pipeline* de IUPs possui como dado de entrada somente o proteoma predito dos organismos em estudo e como resultado é gerado um banco de dados com informações sobre a caracterização das IUPs, dois arquivos de análise dos resultados (análise descritiva e de contingência) e uma estrutura de diretórios contendo os arquivos resultantes de cada predição (Figura 2).

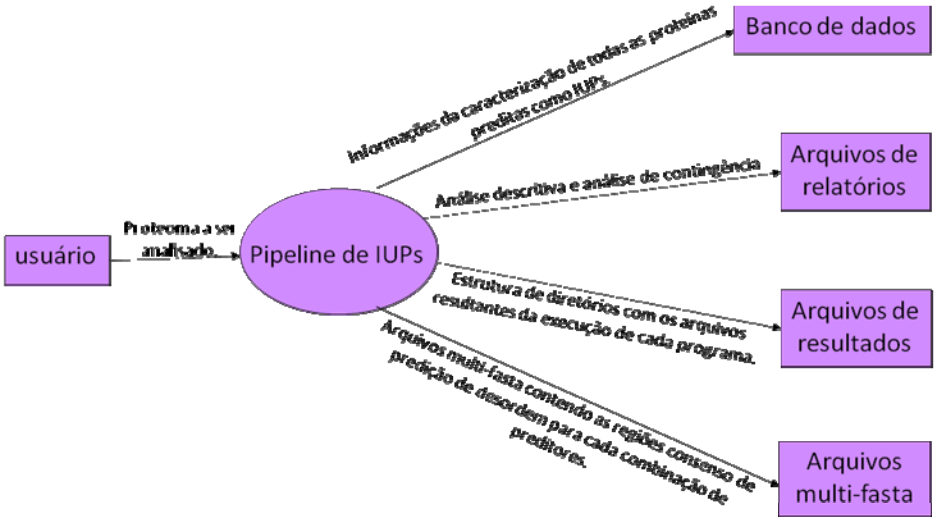


Figura 2: Diagrama de contexto do *pipeline* de IUPs

4.3.2 Funcionalidades

Para atingir os objetivos propostos, o *pipeline* de IUPs possui 10 funcionalidades principais:

1. Criar uma estrutura de diretórios que organizará a execução do *pipeline* de IUPs;

2. Criar um banco de dados para o armazenamento dos dados e consulta interna durante a execução do *pipeline* de IUPs;
3. Pré-processar as seqüências;
    2.1 Remover as seqüências com a presença de caracteres ilegais;
    2.2 Remover seqüências menores que 100 aminoácidos;
    2.3 Remover as seqüências que não possuem Metionina como aminoácido inicial;
4. Identificar as proteínas que possuem região desordenada maior que 40 aminoácidos;
5. Caracterizar as proteínas preditas como IUPs;
    5.1 Predizer regiões transmembranas e as coordenadas do peptídeo sinal quando presente;
    5.2 Predizer as características físico-químicas;
    5.3 Classificar as proteínas entre hipotética e com função predita.
    5.4 Predizer a localização celular;
    5.5 Predizer a anotação funcional;
6. Armazenar os resultados no banco de dados de IUPs;
7. Criar regiões consenso de predição para todas as combinações de preditores;
8. Calcular o número de resíduos desordenados de cada região desordenada, sua localização na seqüência e o número de regiões desordenadas para cada proteína predita como IUP;
9. Criar arquivos fasta contendo as regiões consenso formadas;
10. Criar arquivos de relatórios com a análise descritiva e análise de contingência dos resultados obtidos.

Todas essas funcionalidades são detalhadas nos itens abaixo de acordo com a ordem de execução dentro do *pipeline* de IUPs.

4.3.2.1 Criação da estrutura de diretórios

Foi necessária a criação de uma estrutura bem definida de diretórios para que todos os arquivos originais de resultados obtidos fossem armazenados de maneira organizada, levando-se em conta o organismo que estava sendo analisado, para que assim a extração das informações relevantes de cada resultado fosse feita de

maneira rápida e segura.

Os primeiros diretórios formados são: PREDICTIONS, PROTEOME, FASTA_FILES, CONTINGENCY e DESCRITIVE.

O diretório PREDICTIONS irá conter vários subdiretórios diferentes referenciando cada programa de predição utilizado no *pipeline* de iUPs com seus respectivos resultados.

O diretório PROTEOME irá conter o arquivo fasta com o proteoma após o pré-processamento das seqüências, ou seja, somente as seqüências das proteínas que passaram pelo pré-processamento. Além disso, nesse diretório existem dois subdiretórios onde um contém arquivos individuais que representam cada uma das seqüências fasta e o outro subdiretório contém arquivos no formato “flat” de seqüência para execução do programa VSL2B.

O diretório FASTA_FILES irá conter os arquivos com as seqüências das proteínas preditas como IUPs para cada diferente combinação de preditores, sendo que no cabeçalho das seqüências foram integradas as coordenadas das regiões ou região desordenada(s) predita(s) para determinada IUP.

O diretório CONTINGENCY que irá conter análise de contingência representando a associação de porcentagem de resíduos desordenados com outras características do banco.

Por último o diretório DESCRITIVE que irá conter o arquivo de análise descritiva e os gráficos referentes a essa análise.

4.3.2.2 Pré-processamento das seqüências

O pré-processamento do proteoma é uma filtragem das seqüências que deverão ser analisadas no *pipeline* de IUPs, sendo essa filtragem baseada nos seguintes pontos:

- Remover as seqüências com possíveis erros de anotação;
- Remover seqüências menores que um determinado tamanho;
- Remover as seqüências que não possuem Metionina como aminoácido inicial.

A etapa de pré-processamento está relacionada à remoção através do

emprego de expressões regulares de possíveis erros de anotação associados à identificação da Metionina inicial, códons de parada internos (TGA,TAG e TAA) à seqüências e caracteres ilegais (*, X, B, Z e U) que podem ter sido inseridos durante o processo de anotação automática.

Outro importante ponto do pré-processamento está relacionado ao tamanho das proteínas analisadas, pois a maioria dos preditores utilizados possui a tendência de identificar com mais precisão regiões desordenadas longas (aproximadamente maiores que 40 pb) (Han, Zhang *et al.*, 2009), portanto, o filtro do tamanho das seqüências que serão analisadas é importante na execução do *pipeline*.

4.3.2.3 Predição das IUPs

A próxima etapa do *pipeline* foi a predição *ab initio* das IUPs, para isso avaliamos os preditores existentes, e selecionamos aqueles que estavam disponíveis para download e execução local. Essa característica é imprescindível para a sua utilização em um *pipeline* local de análise. Selecionamos 4 preditores que implementam 6 metodologias diferentes de predição de IUPs: Disembl (implementa 3 metodologias), Globplot, IUPred e VSL2B. Todos os resultados de predição de IUPs foram armazenados no banco de dados de IUPs de maneira automática através da utilização de *scripts* Perl e sentenças SQL.

4.3.2.3.1 DisEMBL

A predição utilizando-se Disembl pode ser feita através de três critérios que consideram diferentes características para predição de uma IUP.

As três definições de desordem utilizadas pelo Disembl são: a) Loops/Coils segundo a definição do Dicionário de Estrutura Secundária de Proteína (DSSP – Dictionary of Protein Secondary Structure (Kabsch e Sander, 1983)) sendo sua predição promíscua e, portanto não considerado em nossas análises; b) Hot-loops que avalia um subgrupo do critério anterior considerando somente loops com alto grau de mobilidade e c) Remark465 que considera coordenadas perdidas em Raio-X

como regiões desordenadas baseando-se no banco de dados PDB (Protein Data Bank) (Linding, Jensen *et al.*, 2003).

O DisEMBL utiliza como dado de entrada um arquivo multi-fasta das seqüência protéicas analisadas, sendo que foi executado com os parâmetros sugeridos como padrão no programa.

4.3.2.3.2 IUPred

O algoritmo IUPred realiza predições avaliando a energia resultante das interações inter-resíduos, sendo que essa interação é maior em proteínas globulares. Sendo assim, o IUPred utiliza uma abordagem que estima o potencial dos polipeptídios de formarem contatos estabilizantes, através de um modelo de interação estatístico (Dosztányi, Csizmók *et al.*, 2005). Foi executado utilizando como dado de entrada um arquivo com uma única seqüência no formato fasta e utilizando o parâmetro para identificação de regiões longas de desordem, já que é o único preditor utilizado no *pipeline* de IUPs que possui a opção de identificar regiões curtas e longas de desordem.

O arquivo resultante do IUPred fornece a probabilidade de cada aminoácido ser desordenado. Como critério, consideramos um aminoácido como estando inserido em uma região desordenada quando seu valor de escore é maior ou igual a 0.5. O escore varia entre 0 e 1, valores maiores indicam maior probabilidade de desordem.

4.3.2.3.3 GlobPipe

No algoritmo GlobPipe 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) (Linding, Russell *et al.*, 2003).

O GlobPipe utiliza como dados de entrada um arquivo multi-fasta com as seqüências das proteínas analisadas, sendo que sua execução no *pipeline* de IUPs seguiu os parâmetros utilizados no servidor web.

4.3.2.3.4 VSL2B

O VSL2 é um preditor baseado em métodos de aprendizado de máquinas que possui três variantes: VSL2B, VSL2P e VSL2. A variante utilizada no *pipeline* foi a variante VSL2B que utiliza somente informações obtidas a partir da seqüência de aminoácidos como: flexibilidade da molécula, hidrofobicidade, carga elétrica, entropia, razão carga/hidropatia e a freqüência de cada resíduo (Peng, Radivojac *et al.*, 2006).

Para a sua execução, o VSL2B precisa de arquivos que contenham somente uma seqüência protéica e que estejam no formato “flat”, ou seja, que contenha somente a seqüência sem o cabeçalho fasta. O VSL2B foi executado pelo *pipeline* de IUPs com seus parâmetros padrão.

4.3.2.4 Predição das regiões transmembranas e peptídeo sinal das IUPs preditas

Nesta etapa utilizamos o algoritmo Phobius (Käll, Krogh *et al.*, 2004) que é um método combinado de predição de regiões transmembranas e peptídeo sinal, disponível para download em http://phobius.sbc.su.se/. O algoritmo Phobius é baseado em modelos ocultos de Markov (HMM) que modela a diferença entre regiões de peptídeo sinal e transmembranas.

O programa foi utilizado tendo como entrada o arquivo contendo o proteoma predito no formato fasta e a para maior clareza na disposição das informações do arquivo de resultado foi utilizada a opção “formato longo”.

4.3.2.5 Predição das características físico-químicas das IUPs preditas

O EMBOSS (*The European Molecular Biology Open Software Suite*) é um pacote de software de análise grátis e de código aberto especialmente desenvolvido para as necessidades da biologia molecular.

Um dos programas do EMBOSS é o *pepstats* que caracteriza propriedades físico-quimicas das proteínas analisadas incluindo: peso molecular, número de resíduos, carga, ponto isoelétrico entre outros. No *pipeline* de IUPs o *pepstats* foi executado com seus parâmetros padrão e tendo como dado de entrada o proteoma em estudo no formato fasta.

4.3.2.6 Predição da localização subcelular das IUPs preditas

Para a predição da localização subcelular utilizamos o programa WoLF PSORT (Horton, Park *et al.*, 2007) que é baseado na seqüência de aminoácidos das proteínas, faz suas predições baseadas em classificação de sinais de padrões conhecidos e algumas características das seqüências como o conteúdo de aminoácidos.

Esse programa foi escolhido por ser o que apresenta um maior detalhamento da localização predita e esse detalhamento vai além das predições clássicas (cloroplasto, mitocôndria e via de secreção) feitas por algoritmos similares como o TargetP (Emanuelsson, Nielsen *et al.*, 2000). A classificação do WoLF PSORT inclui os termos: cloroplasto, citosol, citoesqueleto, retículo endoplasmático, extracelular, Complexo de Golgi, lisossomo, mitocôndria, nuclear, peroxissomo, membrana plasmática e membrana vacuolar.

O WoLF PSORT utiliza como dado de entrada um arquivo multi-fasta das seqüências protéicas analisadas e foi executado com seus parâmetros padrão.

4.3.2.7 Predição da anotação funcional das IUPs preditas

4.3.2.7.1 *Gene Ontology*

A anotação funcional foi baseada no vocabulário de classificação funcional do *Gene Ontology* (GO). O projeto GO é a maior iniciativa em bioinformática com o objetivo de padronizar a representação dos genes e produtos gênicos. O projeto prevê um vocabulário controlado de termos para descrever as características dos produtos gênicos e seus dados de anotação através dos membros do Consórcio GO (http://www.geneontology.org/GO.consortiumlist.shtml) e de ferramentas para

acessar e processar esses dados. O projeto GO incentiva contribuições da comunidade tanto no conteúdo do GO quanto na anotação utilizando GO, e isso garante que o GO seja completo e acurado (http://www.geneontology.org/).

O Projeto GO desenvolveu três vocabulários estruturados e controlados (ontologias) que descrevem os produtos gênicos em termos de seus processos biológicos, componentes celulares e funções moleculares de maneira independente da espécie e em diferentes níveis, dependendo da profundidade de conhecimento requerida.

O GO não abrange algumas áreas e os termos relacionados a essas áreas não aparecerão nas ontologias, essas domínios não incluídos estão disponíveis em http://www.geneontology.org/GO.doc.shtml.

4.3.2.7.2 Blast2GO

O programa escolhido para realizar a anotação funcional foi o Blast2GO *Pipeline Version* (B2G4Pipe) (Conesa, Götz *et al.*, 2005). O B2G4Pipe executa o Blast2GO sem interface gráfica, sendo essa característica essencial para a inclusão no *pipeline* de IUPs.

O B2G4Pipe utiliza como entrada o resultado da comparação de similaridade de seqüências no formato XML, portanto, inicialmente é necessária a execução de um BLAST (*Basic Local Alignment Search Tool*) (Altschul, Gish *et al.*, 1990) das seqüências proteicas analisadas contra o banco de dados não redundante (nr) do NCBI disponível em ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/. Foram executados no total 5 blasts um para cada um dos parasitos estudados. Os seguintes parâmetros foram utilizados: (-p: programa utilizado) blastp que compara seqüências de proteína; (–d banco de dados) nr; (–i arquivos multifasta de entrada) seqüências das proteínas preditas de cada parasito; (–o nome do arquivo de saída) nome diferente para cada parasito; (–a número de processadores utilizados) 2; (–e limite de *e-value* considerado) 10^{-6} e (–m formato do arquivo de saída) XML.

Com o resultado blast em formato XML, é executado o programa b2g4pipe.

Para a completa identificação da anotação funcional das proteínas preditas como IUPs, uma informação adicional sobre a que domínio pertence determinado identificador GO (predito pelo Blast2GO4Pipe), é adquirida através de uma busca no banco de dados GO_terms explicado no item 4.3.4.

4.3.2.7.3 Análise de enriquecimento funcional

Para identificar os termos funcionais enriquecidos nas IUPs de cada organismo estudado, utilizamos a biblioteca GO::TermFinder (Boyle, Weng *et al.*, 2004) que permite identificar os termos funcionais enriquecidos, determinando o valor-p para a anotação associada em uma lista de genes. Para isso, é utilizada a distribuição hipergeométrica e um valor-p corrigido é calculado para corrigir o teste de múltiplas hipóteses.

O valor-p calculado na distribuição hipergeométrica utiliza a seguinte formula:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-1}}{\binom{N}{i}}$$

Sendo N o número total de genes na distribuição de fundo (todos os genes de um determinado organismo), M é o número de genes naquela distribuição que são anotados, n número de genes da lista de interesse (IUPs) e k número de genes da lista de interesse anotados. Utilizamos a correção de múltiplas hipóteses de bonferroni e valor-p corrigido menor que 0,05 como significante.

4.3.2.8 Armazenamento dos resultados no banco de dados formado

Para o armazenamento dos resultados das predições foram desenvolvidos *scripts* que extraem informações relevantes dos arquivos de resultados e armazenam no local correto do banco de dados de IUPs.

Os *scripts* desenvolvidos para os preditores de IUPs utilizam BioPerl para selecionar informações básicas sobre as seqüências analisadas e expressões regulares para identificar as informações relevantes sobre as regiões de desordem

presentes no resultado de cada proteína. Essas informações são então inseridas no banco de dados de IUPs através de sentenças SQL.

4.3.2.9 Classificação hipotética ou predita

Outra etapa da caracterização das proteínas preditas como IUPs é sua classificação entre hipotética ou com função predita. Essa classificação é feita baseando-se no resultado da comparação de similaridade de seqüência de cada proteína, através do algoritmo BLAST contra o banco de dados não redundante de proteínas do NCBI. O valor de corte de *e-value* utilizado foi 10^{-6}.

4.3.2.10 Criação das regiões consenso de predição e dos arquivos fasta contendo essas regiões formadas pelas diferentes combinações de preditores de IUPs

Para análise das regiões preditas como IUPs por diferentes combinações de preditores, tornou-se necessária a delimitação do consenso de predição de cada combinação, para isso, são considerados dois casos: quando as regiões tinham sobreposição de pelo menos um aminoácido e quando essas regiões estavam próximas.

No primeiro caso, é considerada a menor coordenada como o ponto inicial da região consenso e a maior coordenada o ponto final. Na segunda ocorrência, é verificado se uma margem de 10% além dos limites de cada região predita como desordenada se sobrepõem, caso isso aconteça, essas regiões são agrupadas em uma única região de consenso.

A combinação considerada na análise varia desde o preditor sozinho até todas as possíveis combinações com os demais preditores. Para que a geração das possíveis combinações fosse feita automaticamente, utilizamos a função Perl *powerset* presente na biblioteca Data::PowerSet.

As coordenadas finais de cada região desordenada consenso de cada diferente combinação foi inserida na tabela DISORDER_nr do banco de dados de IUPs para servir como base na formação dos arquivos multi-fasta com as

seqüências das proteínas que possuem regiões desordenadas. Esses arquivos contêm somente as seqüências no formato fasta das proteínas preditas como IUPs pela combinação analisada, sendo que no cabeçalho da seqüência estão presentes as coordenadas das regiões desordenadas encontradas para a proteína. Os arquivos multi-fasta estão presentes no diretório FASTA_FILES como citado no 4.3.2.1.

4.3.2.11 Calculo de estatísticas das regiões de IUPs preditas e armazenamento no banco de dados

Nesta etapa, são calculados: a localização das regiões desordenadas em relação à proteína analisada (N terminal, C terminal, intermediário e CN terminal), o número de resíduos desordenados e a porcentagem de resíduos desordenados para cada proteína por diferentes combinações de preditores. São consideradas como localizadas na região N ou C terminal, regiões presentes nos 15% inicial ou final do tamanho total da seqüência, respectivamente. A nomenclatura CN terminal indica que os resíduos desordenados integram os aminoácidos da região N e C terminal.

Todas essas informações são calculadas para cada combinação de preditores em cada proteína identificada como IUP pela combinação e armazenadas na tabela STATISTICS do banco de dados de IUPs.

4.3.2.12 Criação dos arquivos de relatório

A finalização do *pipeline* de IUPs gera arquivos de relatórios: um com a análise descritiva obtido através da execução do script create_analisys_V2.pl e arquivos com a análise de contingência obtidos através da execução do script create_categories.pl.

4.3.2.12.1 Arquivo de análise descritiva

Este arquivo resume algumas informações fundamentais encontradas com a execução do *pipeline* de IUPs. Os dados são obtidos através da execução de sentenças SQL feitas ao banco de dados formado. As informações incluem:

- Número total de proteínas do arquivo fasta de entrada;
- Número de proteínas maiores que o tamanho mínimo escolhido pelo usuário;
- Número de proteínas que começam com Metionina (somente se a opção de verificação de Metionina inicial foi escolhida pelo usuário);
- Número de proteínas que começam com Metionina e não contem erros de anotação (novamente se a opção foi escolhida pelo usuário);
- Número de IUPs preditas pela melhor combinação de preditores (descrito no item 4.3.7);
- Número de IUPs com função predita e hipotética;
- Número de resíduos desordenados;
- A freqüência de aminoácidos nas IUPs em relação a aminoácidos globulares (descrito no item 4.3.2.12.1.1);
- Número de IUPs com determinadas faixas (porcentagens) de resíduos desordenados;
- Número de IUPs com determinados tamanhos de resíduos desordenados;
- Número de IUPs em cada possível localização feita pelo PSORT;
- Número de IUPs com determinadas quantidades de regiões transmembranas;
- Os 20 termos GO mais enriquecidos para cada classe GO: função molecular, processo biológico e componente celular.

Para melhor visualização dessas informações, são gerados gráficos para cada um dos itens descritos acima, além de um gráfico com o comportamento apresentado pela carga e pelo ponto isoelétrico das IUPs. Esses gráficos são gerados com a utilização da biblioteca Gnuplot (http://www.gnuplot.info/) dentro de *scripts* Perl. Os gráficos resultantes e o arquivo de análise descritiva serão colocados dentro do diretório DESCRITIVE.

4.3.2.12.1.1 Calculo da freqüência de aminoácidos

Para analisar a freqüência de aminoácidos presente nas regiões desordenadas em comparação com regiões globulares, utilizamos a metodologia descrita por Romero e colaboradores (Romero, Obradovic *et al.*, 2001).

Para obter a freqüência de aminoácidos em regiões globulares, utilizamos a versão de Maio de 2010 do conjunto de dados PDB_S25 (Hobohm e Sander, 1994) disponível em http://bioinfo.tg.fh-giessen.de/pdbselect/ e que contem somente um representante de cada grupo de proteínas relacionadas (com mais de 25% de identidade) no PDB. Extraímos desses dados os aminoácidos com coordenadas perdidas no relatório do PDB, sendo essa característica um indicador de desordem estrutural.

4.3.2.12.2 Arquivo de análise de contingência

A análise de contingência avalia a associação entre duas variáveis analisando se a freqüência das variáveis relacionadas está acima ou abaixo de um valor esperado (Agresti, 2002) (http://cran.r-project.org/web/packages/vcdExtra/vignettes/vcd-tutorial.pdf). Para sua execução é necessária a montagem de uma tabela de contingência contendo uma linha para cada IUP e em cada coluna um atributo.

Essa análise é baseada em duas hipóteses, a hipótese nula que prevê que as variáveis não têm relação (ou estão aleatoriamente relacionadas) e a hipótese alternativa que corresponde à associação ou relacionamento entre as variáveis. Apesar disso, a estrutura dessa relação não é especificada, ou seja, ela demonstra que há associação, mas não aponta onde. Em todos os testes fixou-se em 0,05% o nível para a rejeição da hipótese de nulidade (nível de significância de 95%) de acordo com os padrões correntes em estudos biológicos.

Nosso interesse com essa análise é encontrar associação entre desordem estrutural com outras características, por isso, verificamos a relação das variáveis do banco de dados com a porcentagem resíduos desordenados. Cada analise feita tem a variável de porcentagem de resíduos desordenados fixada para comparação com as demais.

As outras variáveis do banco precisaram ser categorizadas para formar a tabela de categorias que será analisada, e isso é feito através de um script Perl que utiliza uma sentença SQL para selecionar os dados, depois é verificado o valor da variável e identificada sua classe de acordo com critérios estabelecidos previamente através da observação do banco de dados:

- Tamanho da proteína: dividimos em nove categorias de 100 em 100 aminoácidos, sendo a categoria 1 proteínas com tamanhos >= 100 e >200 aminoácidos até a categoria 9 com proteínas >1000 aminoácidos. Essa divisão foi feita para ter maior resolução onde os tamanhos das proteínas são mais comuns;
- Proteínas com função predita ou hipotética: simplesmente determinamos que o número 1 fosse a classe de hipotéticas e o número 2 a classe de preditas;
- Peso molecular (Da): dividimos aleatoriamente em quatro grupos: <=30.000, >30.000 e <=60.000, >60.000 e <=90.000 e >90.000;
- Localização: determinamos o número 1 como cloroplasto, 2 para citosol, 3 para citoesqueleto, 4 para retículo endoplasmático, 5 para extracelular, 6 para Complexo de Golgi, 7 para lisossomo, 8 para mitocôndria, 9 para nuclear, 10 para peroxissomo, 11 para membrana plasmática e 12 para membrana vacuolar;
- Carga: dividimos em 3 grupos: negativo (<0), neutro (=0) e positivo (>0);
- Ponto isoelétrico: dividimos em 4 grupos: muito ácidas (<3), ácidas (>=3 e <7), básicas (>=7 e <9) e muito básicas (>9);
- Número de regiões desordenadas: dividimos em 5 grupos (=1 e <=5, >5 e <=10, >10 e <=20, >20 e <=30 e >30), sendo que essa divisão foi feita para ter maior resolução onde o número de regiões desordenadas são mais comuns;
- Número de regiões transmembranas: dividimos em 4 grupos (=1, =2, =3 e =4 regiões transmembranas).

Como é um pré-requisito da análise de contingência, foram consideradas na análise somente as IUPs que continham informação para todos os atributos selecionados.

Para essa análise foi utilizada a linguagem R (http://www.r-project.org/) e o pacote VCD (*Visualizing Categorical Data* - http://cran.r-project.org/web/packages/vcd/vcd.pdf). Neste pacote são geradas três medidas de associação (coeficiente de contingência, coeficiente Phi e coeficiente Cramer's V) que quantificam a força da associação entre as duas variáveis analisadas, sendo o valor 0 a associação mais fraca e 1 a mais forte.

Considerando que a análise de contingência demonstra somente se há ou não associação entre duas variáveis, mas não demonstra onde, é essencial a produção de gráficos que facilitam a visualização de onde está a diferença encontrada. Portanto, para cada associação encontrada (p-valor<0.05), é feita a montagem do gráfico de associação que é sugerido por Cohen e Friendly (Cohen, 1980) (Friendly, 1992) e fornece um meio para melhor visualização dos resultados.

Os resultados da análise de contingência encontram-se dentro do diretório CONTINGENCY onde estarão os gráficos das associações encontradas e o arquivo .Rout com o resultado de execução para todas as variáveis analisadas.

4.3.3 Estrutura necessária e execução do *pipeline* de IUPs

Nesta seção será mostrada qual a estrutura necessária para a execução do *pipeline* de IUPs que requer a criação do banco de dados de IUPs e do banco de dados contendo todos os termos do GO. Todas as informações descritas nesse item estão colocadas no arquivo READ-ME (Anexo 1).

Todas as etapas descritas abaixo foram executadas para a criação do banco de dados de IUPs que contem todas as informações de identificação e caracterização das proteínas do tipo IUPs preditas para todos os tripanosomatídeos estudados.

4.3.3.1 Pré-requisitos necessários a instalação do *pipeline* de IUPs

O *pipeline* de IUPs foi desenvolvido para ser utilizado em sistemas Linux e todos os itens abaixo devem ser instalados pelo administrador do sistema que deve

acrescentar o caminho do diretório de cada executável na variável de ambiente PATH, editando o arquivo /etc/profile e acrescentando o caminho dos diretórios.

- Mysql;
- Perl – bibliotecas: Mysql, IO::File, Getopt::Long, Data::PowerSet, XML::SAX::ExpatXS;
- BioPerl – bibliotecas: Bio::SeqIO, Bio::SearchIO;
- Preditores de IUPs
  - DisEMBL
  - GlobPipe
  - IUPred
  - VSL2B
- Phobius;
- Wolf PSORT;
- EMBOSS – Pepstats;
- Blast2GO4Pipe;
- Blast – versão 2.2.21;
- Banco de dados não redundante do NCBI formatado – disponível em ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/.

4.3.4 Criação do banco de dados de IUPs

Utilizamos como sistema de gerenciamento de banco de dados o pacote MySQL de código fonte livre, disponível em http://www.mysql.com. O banco de dados IUPs foi desenvolvido e modelado utilizando o programa DBDesigner (Data Bank Designer) disponível em http://www.fabforce.net/dbdesigner4/index.php que é um programa de código livre, para o desenho e criação do código fonte do banco de dados para Linux/OS.

Utilizamos o modelo de entidade relacionamento para desenvolver o banco de dados de IUPs, que tem como finalidade descrever de maneira conceitual, os dados a serem utilizados em um sistema de informações, sendo representado graficamente pelo diagrama de entidade relacionamento (Figura 3).

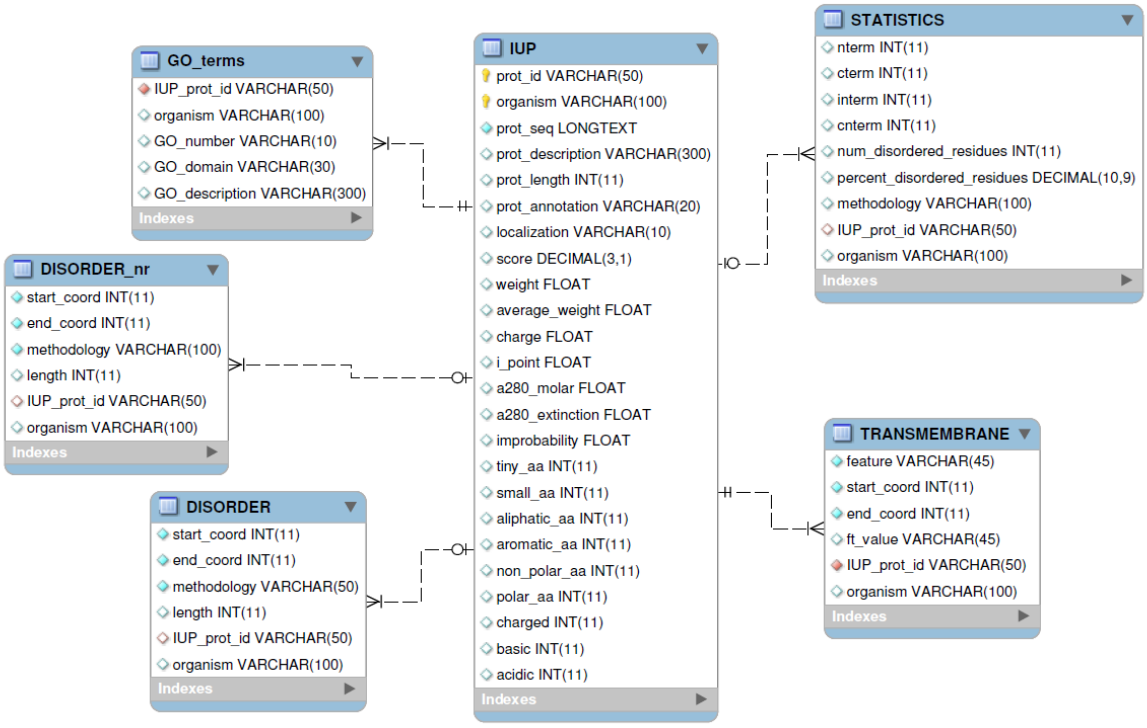


Figura 3: Diagrama entidade relacionamento do banco de dados de IUPs.

Como podemos observar no diagrama de entidade relacionamento, no total foram criadas 6 tabelas no banco de dados de IUPs: IUP, DISORDER, TRANSMEMBRANE, GO_terms, DISORDER_nr e STATISTICS.

A tabela IUP contem todas as proteínas dos organismos estudados que foram preditas como contendo regiões desordenadas além de algumas informações sobre a caracterização dessas regiões, sendo ao todo 30 características presentes nessa tabela (Tabela 2).

Tabela 2: Tabela IUP do banco de dados de IUPs

Tabela IUP	
Campo	**Descrição**
prot_id	Identificador da proteína
organism	Nome do organismo
prot_seq	Seqüência de aminoácidos da proteína
prot_description	Descrição da proteína, presente no arquivo fasta
prot_length	Tamanho da seqüência de aminoácidos da proteína
prot_annotation	Classificação da proteína: hipotética ou com função predita
localization	Predição da localização subcelular
score	Escore da predição de localização subcelular
weight	Peso molecular
average_weigth	Média do peso do resíduo
charge	Carga elétrica
i_point	Ponto isoelétrico
a280_molar	Coeficiente de extinção molar (A280)
a280_extinction	Coeficiente de extinção 1mg/ml (A280)
probability	Probabilidade de exclusão 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

A tabela DISORDER contem informações das predições de cada preditor considerado, sendo ao todo 6 características presentes na tabela (Tabela 3).

Tabela 3: Tabela DISORDER do banco de dados de IUPs

Tabela DISORDER	
Campo	**Descrição**
start_coord	Coordenada inicial da região de desordem predita
end_coord	Coordenada final da região de desordem predita
methodology	Metodologia responsável pela predição
length	Tamanho da região de desordem predita
IUP_prot_id	Identificador da proteína que possui a região de desordem predita
organism	Nome do organismo

A tabela TRANSMEMBRANE contem informações sobre possíveis regiões

transmembranas presentes nas IUPs preditas, além da coordenada do peptídeo sinal caso ele esteja presente, sendo ao todo 6 características descritas na tabela (Tabela 4).

Tabela 4: Tabela TRANSMEMBRANE do banco de dados de IUPs

Tabela TRANSMEMBRANE	
Campo	**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 que possui o domínio predito
organism	Nome do organismo

A tabela GO_terms apresenta os termos do GO encontrados para cada proteína, sua descrição e o domínio que pertence (função molecular, processo biológico ou componente celular) sendo ao todo 5 características presentes na tabela (Tabela 5).

Tabela 5: Tabela GO_terms do banco de dados de IUPs

Tabela GO_terms	
Campo	**Descrição**
GO_number	Identificador da funcionalidade predita
GO_domain	Domínio GO a qual pertence
GO_description	Descrição GO que possui
IUP_prot_id	Identificador da proteína
organism	Nome do organismo

A tabela DISORDER_nr contem as regiões de IUP resultantes tanto dos preditores individuais quanto de suas combinações de maneira não redundante, sendo ao todo 6 informações contidas na tabela (Tabela 6).

Tabela 6: Tabela DISORDER_nr do banco de dados de IUPs

Tabela DISORDER_nr	
Campo	**Descrição**
start_coord	Coordenada inicial da região de desordem predita - sem redundância
end_coord	Coordenada final da região de desordem predita - sem redundância
methodology	Metodologia responsável pela predição
length	Tamanho da região de desordem predita - sem redundância
IUP_prot_id	Identificador da proteína que possui a região de desordem predita
organism	Nome do organismo

A tabela STATISTICS contém as estatísticas gerais das informações contidas no banco de dados relacionadas às regiões desordenadas e complementarmente fornece a localização dessas regiões na proteína: N-terminal (amino-terminal), C-terminal (carboxi-terminal), tanto na região C-terminal como na N-terminal e intermediário, o número e a porcentagem de resíduos desordenados em cada IUP para cada metodologia, sendo que ao todo a tabela apresenta 9 campos de informação (Tabela 7).

Tabela 7: Tabela STATISTICS do banco de dados de IUPs

Tabela STATISTICS	
Campo	**Descrição**
nterm	Quantidade de regiões de desordem preditas na região N terminal
cterm	Quantidade de regiões de desordem preditas na região C terminal
interm	Quantidade de regiões de desordem preditas na região intermediária (entre C e N terminal)
cnterm	Quantidade de regiões de desordem preditas que abrangem as regiões C e N terminal
num_disordered_residues	Número de resíduos desordenados
percent_disordered_residues	Porcentagem de resíduos desordenados em relação ao total de aminoácidos
methodology	Metodologia responsável pela predição da região de desordem
IUP_prot_id	Identificador da proteína
organism	Nome do organismo

Inicialmente o administrador do sistema deve criar um usuário mysql específico para que tenha privilégios específicos para cada banco de dados. Para a criação automática do banco de dados de IUPs, o administrador do sistema deve executar o *script* create_db_iup-pipeline.pl e permitir ao usuário criado selecionar, inserir e atualizar informações no banco. Essas etapas tem como resultado final um banco de dados com toda a estrutura de tabelas descritas acima.

4.3.5 Criação do banco de dados GO_terms

Para que a anotação funcional das proteínas preditas como IUPs fossem feitas de maneira mais completa, complementamos o resultado obtido pelo programa Blast2GO4Pipe com o domínio (componente celular, processo biológico ou função molecular) e a descrição de cada identificador GO predito. Essa informação é extraída do arquivo go_20101016-termdb-tables.tar.gz presente no site do *Gene Ontology* (http://archive.geneontology.org/latest-lite/) que apresenta todos os identificadores GO com seus respectivos domínio e descrição.

Para que essa complementação dos dados fosse feita de maneira automática, rápida e segura, é necessária a criação de um banco de dados com todos os identificadores, descrição e domínio do GO que será consultado durante o processo de classificação funcional das IUPs preditas. Vale ainda ressaltar que a informação dessa consulta retorna ao banco de dados de IUPs inserindo o domínio e a descrição de um dado identificador GO.

Para a criação e inserção das informações acima, o administrador do sistema deve executar o script create-insert_db_GO-terms.pl que criará um banco de dados com uma tabela contendo as colunas: GO_id (identificador GO), domain (domínio ao qual pertence) e description (descrição do termo GO) e irá inserir as informações do arquivo go_20101016-termdb-tables.tar.gz no banco dados GO_terms formado. Além disso, o administrador do sistema deve dar ao usuário mysql criado o privilégio de selecionar informações do banco.

4.3.6 Parâmetros obrigatórios na execução do *pipeline* de IUPs

Para a execução do *pipeline* de IUPs, 13 parâmetros obrigatórios (Tabela 8) e devem ser fornecidos pelo usuário.

Tabela 8: Parâmetros do *pipeline* de IUPs.

Parâmetro	Descrição	Valores aceitos
-i	arquivo de entrada contendo as proteínas que serão analisadas em formato fasta	localização e nome do arquivo
-ls	tamanho mínimo das seqüências que serão consideradas na análise	número que representa o tamanho desejado
-met	verifica se as proteínas iniciam com Metionina	T para a opção ser verdadeira ou F para a opção ser falsa
-func	classifica as proteínas analisadas em preditas e hipotéticas	T para a opção ser verdadeira ou F para a opção ser falsa
-a	verifica a existência de erros de anotação	T para a opção ser verdadeira ou F para a opção ser falsa
-m	nome do organismo que está sendo analisado	nome do organismo sem conter espaço
-fa	executa a anotação funcional	F para a opção ser falsa ou a localização e o nome do arquivo de resultado blast no formato XML
-d	nome do banco de dados de IUPs	nome do banco sem conter espaço
-u	nome do usuário do banco de dados de IUPs	nome do usuário do banco de dados
-p	senha do usuário do banco de dados	senha
-o	base do nome dos arquivos resultantes	base do nome sem espaços
-db	caminho e nome do banco de dados não redundante do NCBI - nr localmente	localização e nome do arquivo
-ld	tamanho mínimo de desordem considerado	número que representa o tamanho desejado

Com a utilização dos parâmetros para execução do *pipeline*, o usuário pode escolher quais as etapas que deseja executar com seus dados, com isso, diferentes tipos de análise podem ser realizadas.

4.3.7 Melhor combinação de preditores de desordem estrutural

Na dissertação de mestrado de Raul Torrieri (finalizada em fevereiro de 2010), foi analisada a melhor combinação de preditores de regiões de desordem através da metodologia de análise de métodos de classificação chamada ROC (*Receiver Operating Characteristics*) (Green e Swets, 1966), sendo 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).

O conjunto de seqüências controle utilizado na análise de desempenho de predição de desordem estrutural foram obtidas do banco de dados DisProt (*Database of Protein Disorder*), versão 4.9 (http://www.disprot.org/data/version_4.9/).

Como resultado da análise do gráfico ROC gerado, a melhor combinação de preditores de desordem estrutural escolhida foi: REM465, GlobPipe, IUPred e VSL2B que apresentou menor valor de falsos positivos e o valor de verdadeiros positivos entre os 5 melhores valores encontrados, equilibrando-se entre os erros e acertos.

4.3.8 Arquivo READ-ME

Todos os requisitos e passos necessários para instalação e execução do *pipeline* foram agrupados e explicitados no arquivo READ-ME que está presente no mesmo pacote em que se encontram os códigos dos *scripts* (Anexo 1).

Esse arquivo irá direcionar e facilitar a instalação e execução de toda a estrutura do *pipeline* de IUPs pelos usuários do programa, tornando assim reprodutível e utilizável por outras pessoas.

4.4 Testes

Parte do processo que antecede a disponibilização e/ou liberação do código fonte de um software ao usuário final envolve a execução de vários testes que em essência avaliam a existência de falhas no processo de instalação e execução das diferentes etapas.

Até o momento foram realizados três testes distintos.

O teste 01 teve como objetivo verificar a capacidade instalação do *pipeline*. Utilizamos como SO (Sistema Operacional) padrão o Debian.

O teste 02 teve como objetivo verificar se a entrada é adequadamente aceita e a saída é corretamente produzida em cada passo do *pipeline* de IUPs e se a integridade dos bancos de dados é mantida durante a execução.

O teste 03 teve como objetivo a verificação de todos os possíveis caminhos de execução permitidos pelo *pipeline* de IUPs descritos no item 4.3.6.

Para a realização de todos os testes descritos acima foi utilizado um conjunto controlado de 10 seqüências de *Leishmania braziliensis*.

4.5 Limitações do *pipeline* de IUPs

O *pipeline* de IUPs cumpriu seu objetivo de identificar e caracterizar proteínas do tipo IUP de maneira automática, consistente e organizada além de gerar análises iniciais aos resultados obtidos.

Apesar disso, o *pipeline* de IUPs apresenta algumas limitações como: identificar regiões de desordem menores que 40 aminoácidos, sendo necessária a inserção de novos preditores de desordem (além do IUPred) que incluem essas regiões.

4.6 Análise Experimental

Com o objetivo de avaliar a acurácia da predição *in silico* de IUPs utilizamos uma pequena amostragem de proteínas extraídas dos organismos estudados (*L. braziliensis* e *L. major*).

Para tanto foram empregadas duas metodologias experimentais distintas que envolvem eletroforese 2D desenvolvidas especificamente para a identificação de IUPs (Csizmók, Szollosi *et al.*, 2006) e (Galea, Pagala *et al.*, 2006).

Na metodologia de Csizmók e colaboradores, um gel nativo, ou seja, sem desnaturante era utilizado na primeira dimensão, havendo uma separação das proteínas por carga/massa. A canaleta resultante é utilizada como *strip* para a segunda dimensão, onde é feita novamente a separação por caraga/massa em um gel desnaturante utilizando uréia 8M para não influenciar na carga das proteínas. Como resultado, as proteínas do tipo IUP encontram-se na diagonal do gel ou próxima a ela por terem percorrido a mesma distância nas duas dimensões do gel.

A metodologia de Galea e colaboradores tem como base o aquecimento prévio do extrato proteico à 100°C por uma hora e logo em seguida colocado no gelo por 15 minutos, para que assim as proteínas globulares se agreguem e precipitem. O sobrenadante estará então enriquecido de IUPs e é aplicado a uma *strip* para separação por focalização isoelétrica seguido pela segunda dimensão em uma separação por carga/massa.

Esses experimentos foram realizados na Faculdade de Medicina de Ribeirão Preto - USP em colaboração com a Dra Ângela Kaysel Cruz.

A metodologia 1 foi feita para o proteoma de *L. major* e a metodologia 2 para o proteoma de *L. major* e *L. braziliensis*.

5 RESULTADOS

A execução do *pipeline* de IUPs desencadeia três eventos principais: a criação de um banco de dados e a criação de arquivos de análise descritiva e análise de contingência, conforme item 3.1. O banco de dados contem todas as informações preditas durante o processo de execução e os arquivos de análise descritiva e de contingência sumarizam respectivamente os principais resultados descritivos e suas relações para a melhor combinação de preditores de desordem estrutural encontrada: REM465, GlobPipe, IUPred e VSL2B.

O *pipeline* de IUPs foi executado para todos os organismos estudados: *L. major*, *L. braziliensis*, *L. infantum*, *T. cruzi* e *T. brucei*, resultando em um banco de dados contendo as informações de todos os organismos e em um diretório geral para cada organismo com os arquivos de execução e de análise.

5.1 Banco de dados de IUPs

O banco de dados de IUPs possui seis tabelas contendo todas as informações geradas pelas etapas do *pipeline* de IUPs, que inclui o pré-processamento das seqüências, a predição de desordem estrutural e a caracterização das proteínas preditas como IUPs.

No total, para os cinco organismos estudados, o banco de dados de IUPs integra aproximadamente 540 Mb de informações contidas em 2.807.074 linhas que estão divididas nas seis tabelas: IUP, DISORDER, DISORDER_nr, GO_terms, STATISTICS e TRANSMEMBRANE.

5.2 Arquivos de análise descritiva

O arquivo de análise descritiva agrupa todas as informações de um determinado organismo, desde o pré-processamento das seqüências até a identificação e caracterização das proteínas preditas como IUPs pela melhor combinação de preditores de desordem estrutural (REM465, GlobPipe, IUPred e

VSL2B). Esses resultados incluem desde o número de proteínas que passaram pelo pré-processamento até a análise da freqüência de aminoácidos encontrados nas regiões de desordem estrutural, alguns desses resultados são apresentados graficamente para sua maior compreensão.

5.2.1 Pré-processamento

Os resultados presentes no arquivo de análise descritiva relacionados ao pré-processamento estão apresentados na tabela abaixo (Tabela 9).

Tabela 9: Resultado do pré-processamento das seqüências no *pipeline* de IUPs para os cinco organismos estudados.

Organismo	Proteínas maiores que 100 aminoácidos	Proteínas que começam com Metionina	Proteínas que começam com Metionina e não contem erros de anotação
L. braziliensis	98,2% (8165 proteínas)	97,2% (8084 proteínas)	95,8% (7965 proteínas)
L. major	98,1% (8253 proteínas)	97,7% (8222 proteínas)	97,1% (8171 proteínas)
L. infantum	98,2% (8072 proteínas)	97,7% (8034 proteínas)	95,3% (7835 proteínas)
T. cruzi	97,8% (10096 proteínas)	86,4% (8921 proteínas)	79,2% (8178 proteínas
T. brucei	97,0% (9599 proteínas)	96,3% (9537 proteínas)	95,7% (9472 proteínas)

Somente as proteínas que passaram pelo pré-processamento, ou seja, maiores que 100 aminoácidos, que iniciam com Metionina e não possuem erros de anotação continuaram sendo avaliadas pelo *pipeline* de IUPs para que assim as predições fossem feitas de maneira confiável.

5.2.2 Identificação das IUPs

Outro resultado presente no arquivo de análise descritiva é o número de proteínas preditas como IUPs, ou seja, proteínas que possuem pelo menos uma região de desordem maior que 40 aminoácidos e observamos que as espécies de *Leishmania* e de *Trypanosoma* estudadas possuem aproximadamente 70% e 55% de suas proteínas preditas como IUPs respectivamente. Essas proteínas preditas como IUP foram caracterizadas como tendo função predita ou hipotética adquirida através do primeiro *hit* BLAST encontrado em uma busca contra o banco de dados

não redundante do NCBI e podemos observar que para as cinco espécies estudadas a maioria das IUPs é hipotética (Tabela 10).

Tabela 10: Resultado do pré-processamento das seqüências no *pipeline* de IUPs para os cinco organismos estudados.

Organismo	Número de proteínas analisadas	Número de IUPs preditas	Porcentagem de IUPs com função predita	Porcentagem de IUPs hipotéticas
L. braziliensis	7965 (95,8%)	5725 (68,9%)	28,9%	70,9%
L. major	8171 (97,2%)	5959 (70,8%)	29,1%	70,6%
L. infantum	7835 (95,3%)	5797 (70,5%)	28,7%	71%
T. cruzi	8178 (79,2%)	5565 (53,9%).	38,5%	61,5%
T. brucei	9472 (95,7%)	5510 (55,6%)	32%	68%

Com relação ao número de resíduos desordenados, também presente no arquivo de análise descritiva, encontramos os resultados apresentados no gráfico 2.

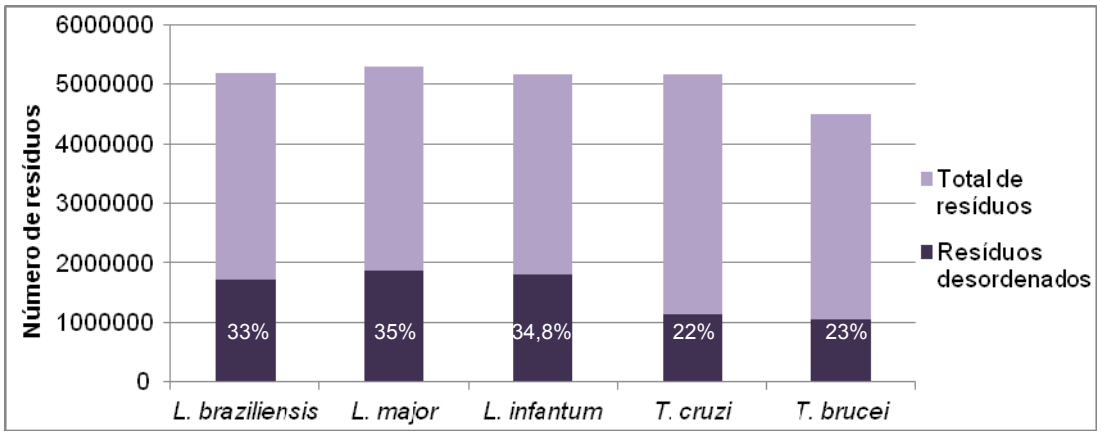


Gráfico 2: Número de resíduos desordenados.

5.2.3 Caracterização das IUPs

No arquivo de análise descritiva estão colocadas algumas análises que mostram a caracterização das proteínas preditas como IUPs, destacando alguns pontos principais das informações encontradas pela execução do *pipeline* de IUPs.

5.2.3.1 Análise da freqüência de aminoácidos

A freqüência de cada um dos aminoácidos desordenados está descrita no arquivo de análise descritiva, sendo que é gerado um gráfico que mostra a relação da freqüência dos aminoácidos desordenados com a de resíduos ordenados presentes no PDB_S25 de acordo com a metodologia de Romero e colaboradores, explicitado no item 4.3.2.12.1.1. Sendo assim, foram gerados cinco gráficos que mostram essa relação entre as freqüências, um para cada organismo estudado (Gráficos 3, 4, 5, 6 e 7).

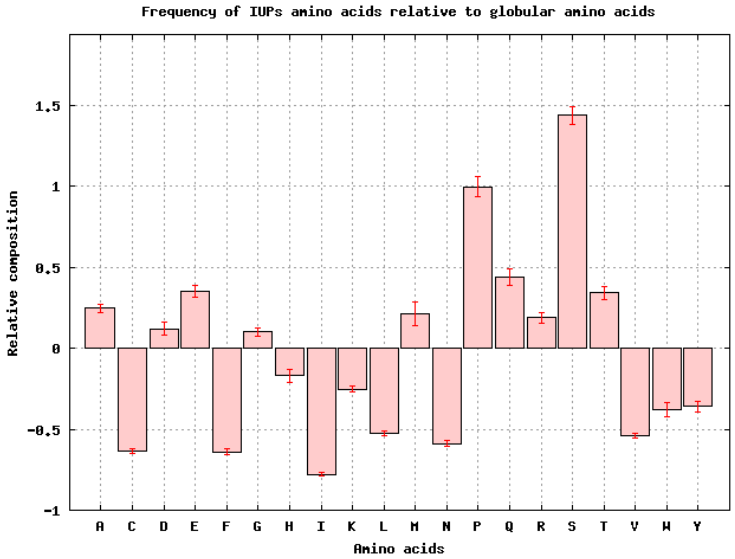


Gráfico 3: Perfil de composição ordem/desordem de *L. braziliensis*. Comparação da composição de aminoácidos de proteínas globulares com aminoácidos de IUPs. Valores negativos indicam o empobrecimento e valores positivos enriquecimento. No eixo Y: composição relativa. No eixo X: aminoácidos.

No gráfico 3 observamos que os aminoácidos enriquecidos em regiões desordenadas para *L. braziliensis* são principalmente: P, Q, S, E, A e T. Já os empobrecidos são: W, Y, F, V, I, L, N e C.

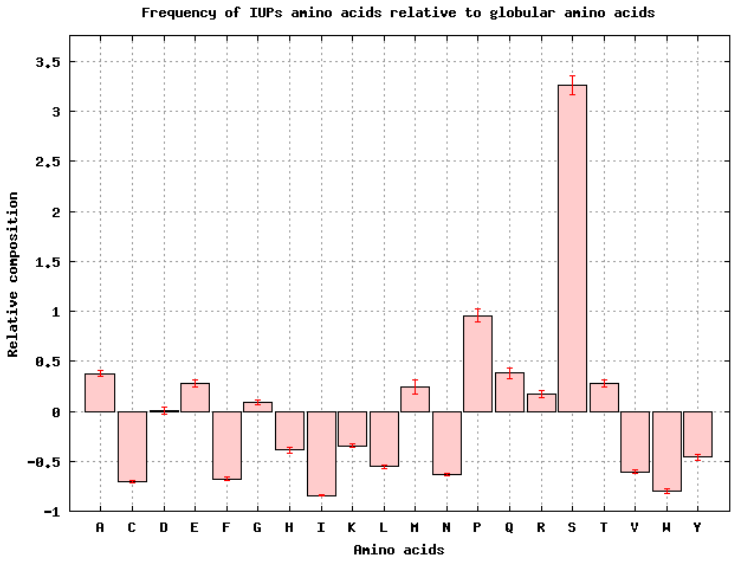


Gráfico 4: Perfil de composição ordem/desordem de *L. major*. Comparação da composição de aminoácidos de proteínas globulares com aminoácidos de IUPs. Valores negativos indicam o empobrecimento e valores positivos enriquecimento. No eixo Y: composição relativa. No eixo X: aminoácidos.

No gráfico 4 observamos que os aminoácidos enriquecidos em regiões desordenadas para *L. major* são principalmente: P, Q, E, S, A e T. Já os empobrecidos são: W, Y, F, V, I, L, N e C.

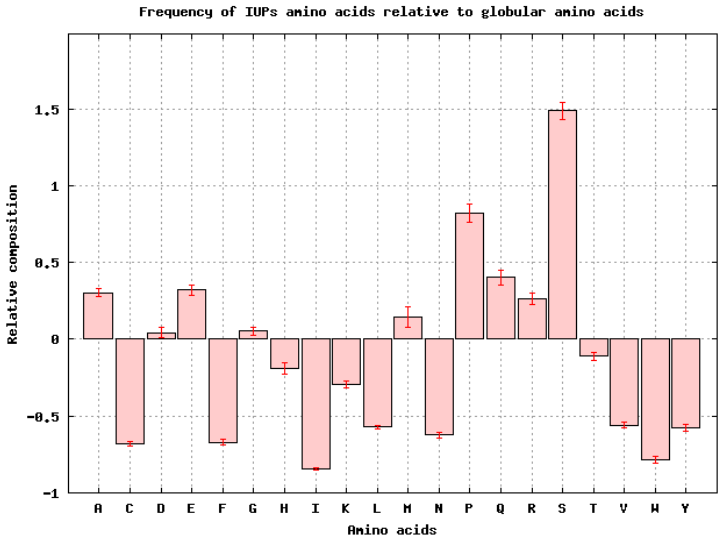


Gráfico 5: Perfil de composição ordem/desordem de *L. infantum*. Comparação da composição de aminoácidos de proteínas globulares com aminoácidos de IUPs. Valores negativos indicam o empobrecimento e valores positivos enriquecimento. No eixo Y: composição relativa. No eixo X: aminoácidos.

No gráfico 5 observamos que os aminoácidos enriquecidos em regiões desordenadas para *L. infantum* são principalmente: P, Q, E, S, A e R. Já os empobrecidos são: W, Y, F, V, I, L, N e C.

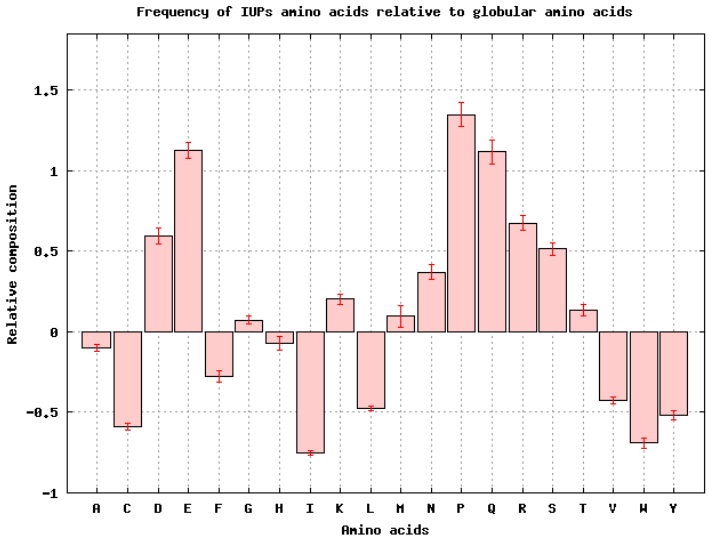


Gráfico 6: Perfil de composição ordem/desordem de *T. brucei*. Comparação da composição de aminoácidos de proteínas globulares com aminoácidos de IUPs. Valores negativos indicam o empobrecimento e valores positivos enriquecimento. No eixo Y: composição relativa. No eixo X: aminoácidos.

No gráfico 6 observamos que os aminoácidos enriquecidos em regiões desordenadas para *T. brucei* são principalmente: P, Q, E, S, D e R. Já os empobrecidos são: W, Y, F, V, I, L e C.

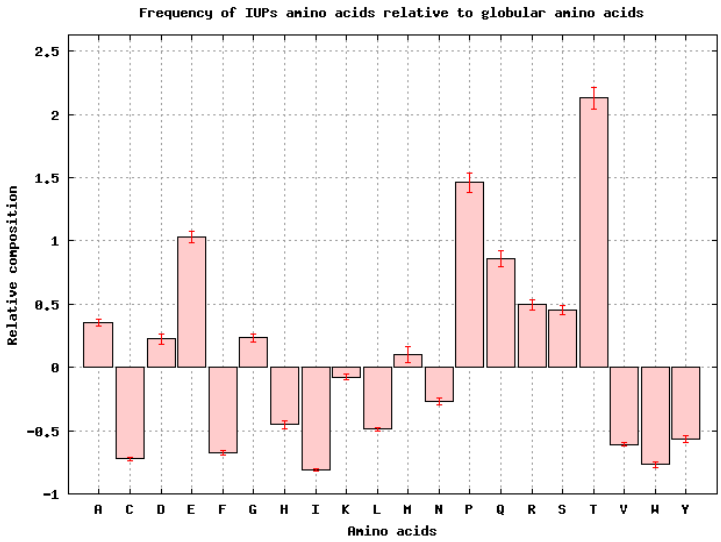


Gráfico 7: Perfil de composição ordem/desordem de *T. cruzi*. Comparação da composição de aminoácidos de proteínas globulares com aminoácidos de IUPs. Valores negativos indicam o empobrecimento e valores positivos enriquecimento. No eixo Y: composição relativa. No eixo X: aminoácidos.

No gráfico 7 observamos que os aminoácidos enriquecidos em regiões desordenadas para *T. cruzi* são principalmente: P, Q, E, S, T e R. Já os empobrecidos são: W, Y, F, V, I, L e C.

Com isso, podemos concluir que em tripanosomatídeos, os aminoácidos enriquecidos em regiões desordenadas são: P, Q, E, R e S e os empobrecidos são: W, Y, F, V, I, L e C.

5.2.3.2 Distribuição dos resíduos desordenados

Outro item analisado foi o número de proteínas que possuem determinadas faixas de porcentagem (de <=10% a 100%) de resíduos desordenados. Com esses dados também é gerado um gráfico mostrando a porcentagem de proteínas em cada faixa de porcentagem de resíduos desordenados.

O gráfico 8 correlaciona faixas de frações de resíduos desordenados com a porcentagem de proteínas preditas como IUPs em *L. braziliensis*, observamos que a grande maioria das IUPs (18%) nesse organismo está centrada na faixa de 10 a 20% de resíduos desordenados e que quase 70% de todas as IUPs têm menos que 50% dos resíduos desordenados.

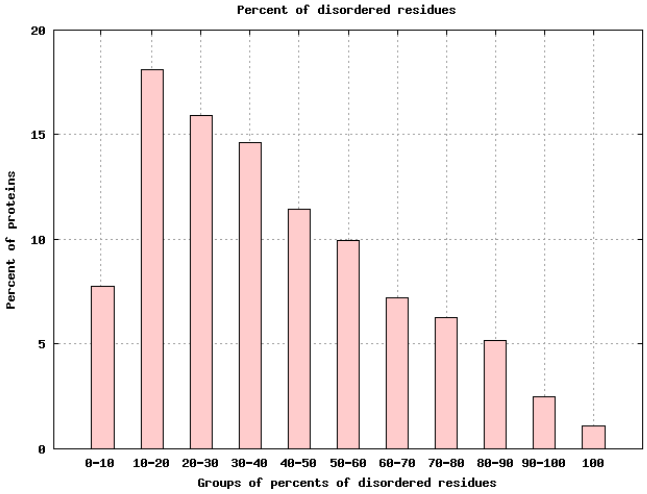


Gráfico 8: Distribuição de resíduos desordenados em *L. braziliensis*. No eixo X: faixas de porcentagem de resíduos desordenados. No eixo Y: porcentagens de proteínas.

Esse comportamento também é observado nos outros organismos estudados (Gráficos 9, 10, 11 e 12).

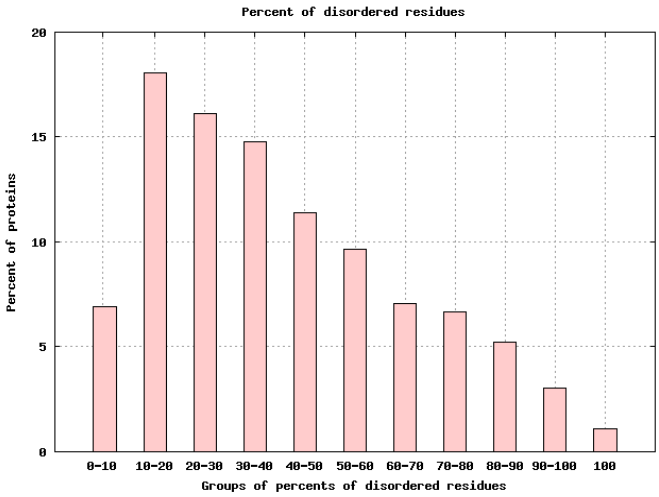


Gráfico 9: Distribuição de resíduos desordenados em *L. major*. No eixo X: faixas de porcentagem de resíduos desordenados. No eixo Y: porcentagens de proteínas.

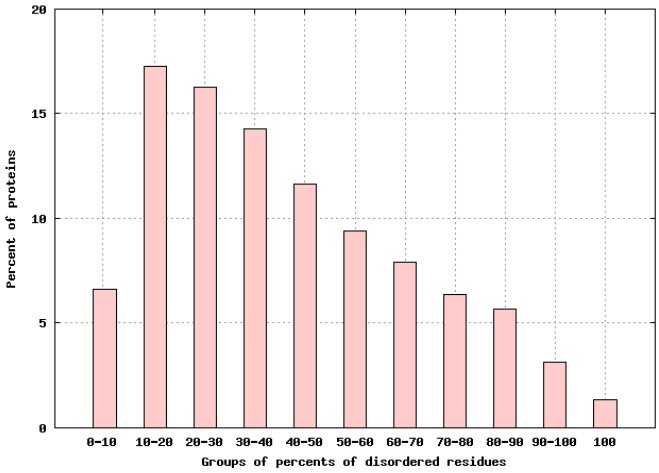


Gráfico 10: Distribuição de resíduos desordenados em *L. infantum*. No eixo X: faixas de porcentagem de resíduos desordenados. No eixo Y: porcentagens de proteínas.

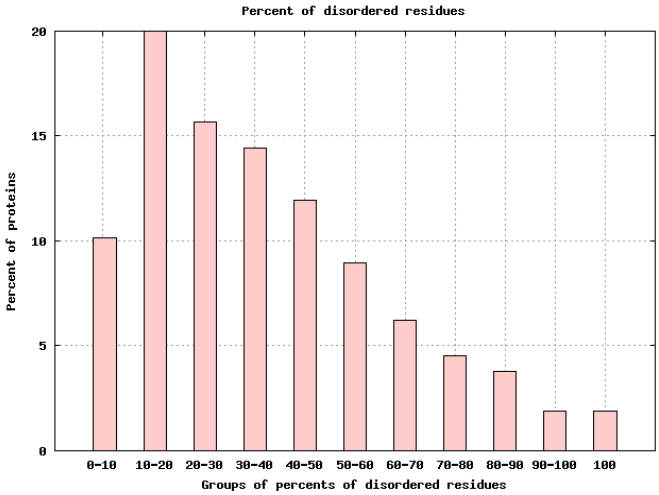


Gráfico 11: Distribuição de resíduos desordenados em *T. brucei*. No eixo X: faixas de porcentagem de resíduos desordenados. No eixo Y: porcentagens de proteínas.

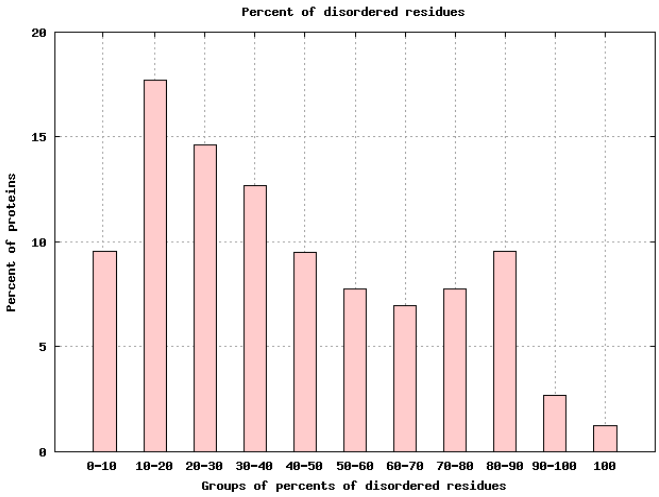


Gráfico 12: Distribuição de resíduos desordenados em *T. cruzi*. No eixo X: faixas de porcentagem de resíduos desordenados. No eixo Y: porcentagens de proteínas.

5.2.3.3 Tamanho das regiões desordenadas

O tamanho da região desordenada presente nas IUPs também é mostrado no arquivo de análise descritiva. Os resultados encontrados foram similares para os cinco organismos estudados, como está exemplificado no gráfico 13 que demonstra o tamanho das regiões desordenadas em *L. braziliensis*.

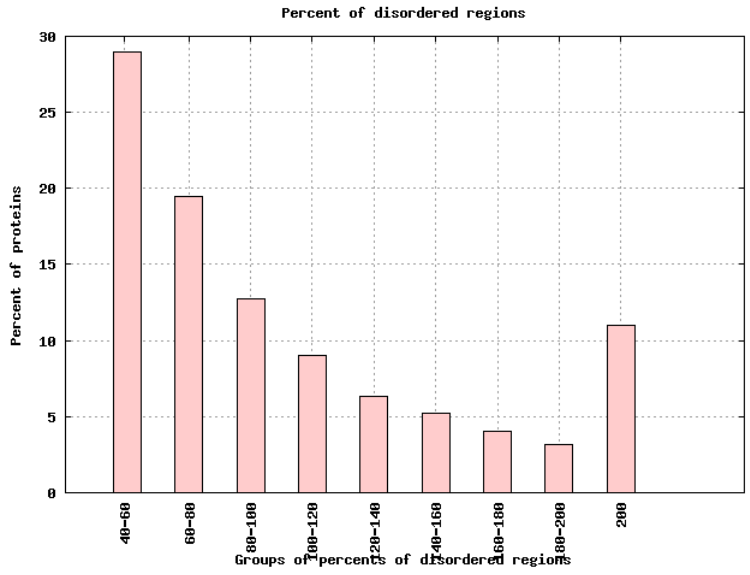


Gráfico 13: Distribuição de regiões desordenados em *L. braziliensis*. No eixo X: faixas de tamanho das regiões desordenadas em AA. No eixo Y: porcentagens de proteínas.

No gráfico observamos que quase 30% das regiões desordenadas possuem de 40 a 60 aminoácidos, sendo que quase metade das regiões desordenada tem até 80 aminoácidos.

5.2.3.4 Predição da localização das IUPs

Outro ponto presente no arquivo de análise descritiva é a localização subcelular das proteínas preditas como IUPs. No arquivo estão presentes o número de proteínas em cada possível classificação disponível no programa WolfPSort, além da geração de um gráfico com a porcentagem de IUPs existente em cada localização (Gráfico 14).

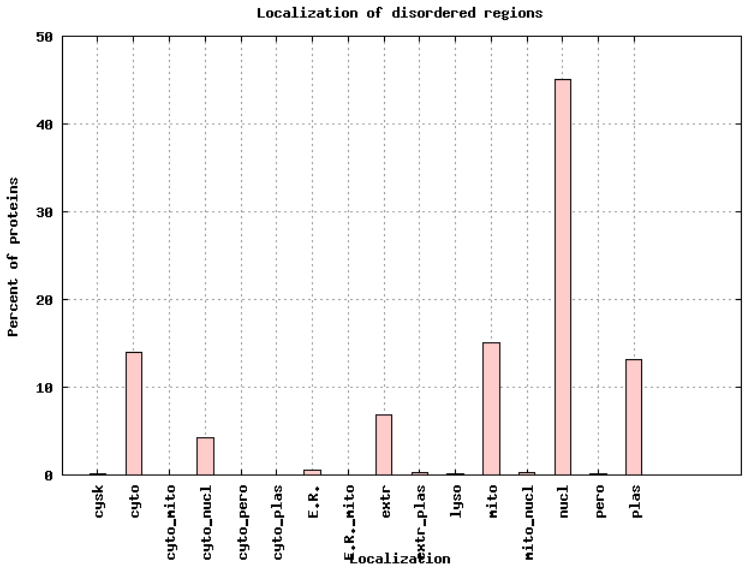


Gráfico 14: Localização subcelular das IUPs em *L. braziliensis*. A predição foi feita pelo algoritmo Wolf Psort com os termos: cysk (citoesqueleto), cyto (citosol), cyto_mito (não definiu a predição entre citosol e mitocôndria), cyto_nucl (não definiu a predição entre citosol e nuclear), cyto_pero (não definiu a predição entre citosol e peroxisomo), cyto_plas (não definiu a predição entre citosol e membrana plasmática), E.R. (Retículo Endoplasmático), E.R._mito (não definiu a predição entre Retículo Endoplasmático e mitocôndria), extr (extracelular), extr_plas (não definiu a predição entre extracelular e membrana plasmática), lyso (lisossomo), mito (mitocôndria), mito_nucl (não definiu a predição entre mitocôndria e nuclear), nucl (nuclear), pero (peroxisomo) e plas (membrana plasmática).

De acordo com o gráfico, as IUPs estão localizadas em 45% dos casos no núcleo, seguido pela mitocôndria, citoplasma, membrana plasmática e extracelular. Os cinco organismos estudados apresentaram resultados similares com relação à porcentagem de IUPs em cada localização subcelular.

5.2.3.5 Número de domínios transmembranas das IUPs

No arquivo de análise descritiva há também a verificação do número de regiões transmembranas presentes em cada IUP. As classes foram divididas desde

uma região transmembrana até mais que 10 regiões e um gráfico foi feito demonstrando a porcentagem de IUPs dentro de cada classe (Gráfico 15).

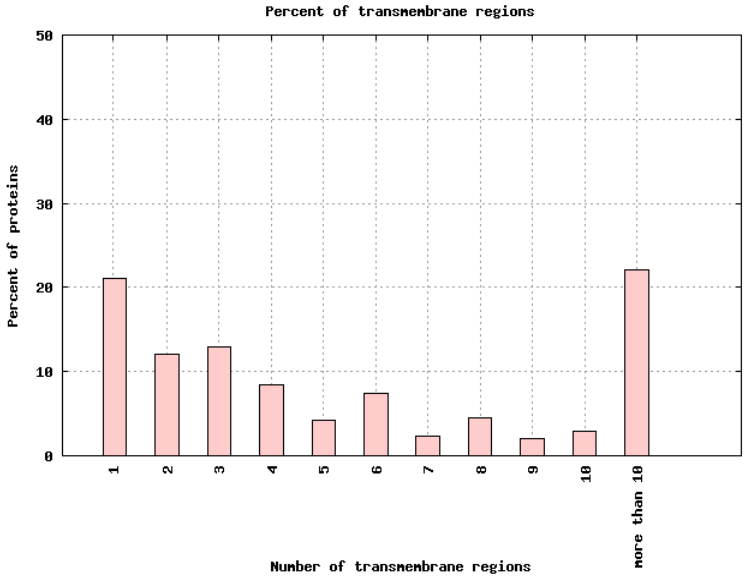


Gráfico 15: Número de regiões transmembranas nas IUPs de *L. braziliensis*. No eixo X: número de regiões transmembrana. No eixo Y: porcentagem de proteínas.

Através do gráfico podemos observar que aproximadamente 50% das IUPs em *L. braziliensis* possui até 3 regiões transmembranas, sendo que o mesmo comportamento acontece com as outras duas espécies de *Leishmania*. Já para as espécies de *Trypanosoma* (Gráfico 16) a um maior número de IUPs com uma região transmembrana, sendo que 65% das IUPs possuem até 3 regiões transmembranas.

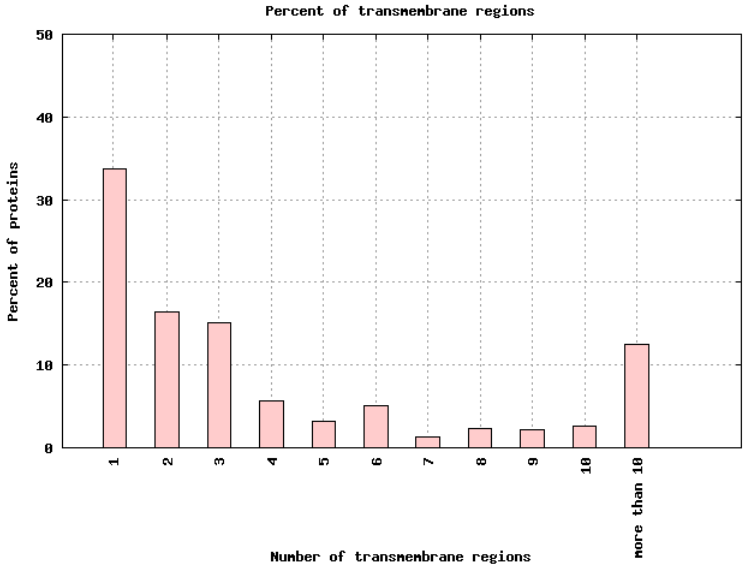


Gráfico 16: Número de regiões transmembranas nas IUPs de *T. brucei*. No eixo X: número de regiões transmembrana. No eixo Y: porcentagem de proteínas.

5.2.3.6 Ancoramento das regiões desordenadas nas proteínas

Outra informação presente no arquivo de análise descritiva é o número de regiões desordenadas em cada região das proteínas: C-terminal, N-terminal, intermediária e CN-terminal. Colocamos o resultado dessa análise para os cinco organismos estudados em um único gráfico (Gráfico 17).

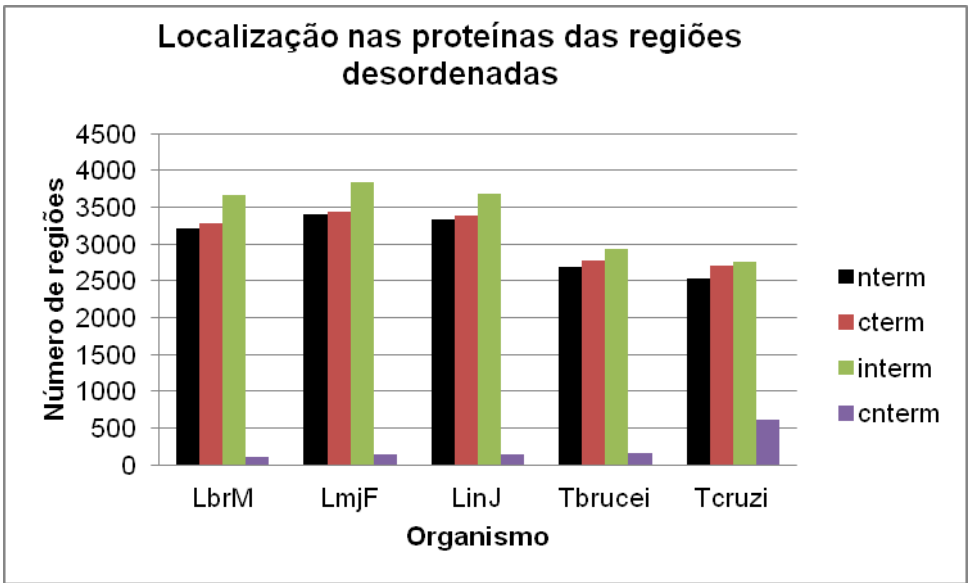


Gráfico 17: Localização das regiões desordenadas nas proteínas. No eixo X: nome do organismo. No eixo Y: número de regiões desordenadas.

Observamos no gráfico que para os organismos estudados há uma tendência nas regiões desordenadas em estarem primeiramente na região intermediária que é a maior região da proteína, em segundo lugar, está a posição C-terminal. Notamos também que o *T. cruzi* tem um número bem maior de regiões desordenadas CN-terminal quando comparado com os outros quatro organismos.

5.2.3.7 Predição funcional das IUPs

No arquivo de análise descritiva estão apresentados os 20 primeiros termos mais enriquecidos entre as proteínas preditas como IUPs, para cada categoria do GO: função molecular, processo biológico e componente celular. Também é gerado um gráfico com a porcentagem daquele termo GO entre todos os termos encontrados para cada uma dessas três categorias acima citadas (Gráficos 18, 19 e

20). Como o perfil observado é muito similar entre os genomas estudados, apresentamos somente os resultados obtidos para *L. braziliensis*.

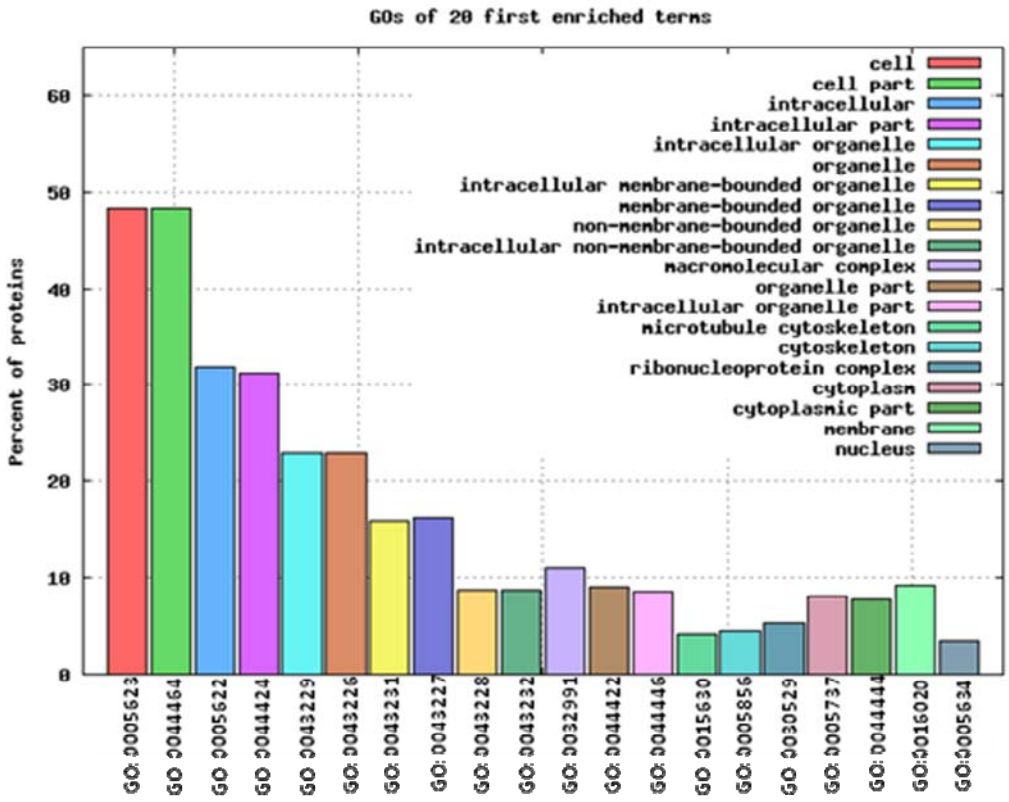


Gráfico 18: 20 primeiros termos GO enriquecidos da categoria componente celular das IUPs de *L. braziliensis*. No eixo X: número GO e na legenda a descrição de cada um. No eixo Y: porcentagem de proteínas.

Analisando o gráfico, observamos que os termos parte da célula e intracelular são os mais comuns entre os 20 termos mais enriquecidos para a categoria componente celular. Segundo a definição do GO, parte celular é qualquer parte constituinte de uma célula, a estrutura básica e unidade funcional de todos os organismos. Já a definição de intracelular são os conteúdos vivos de uma célula; o que está contido (mas não incluído) na membrana plasmática, geralmente excluídos os grandes vacúolos e massas de material de secreção ou ingerido. Em eucariotos inclui o núcleo e o citoplasma.

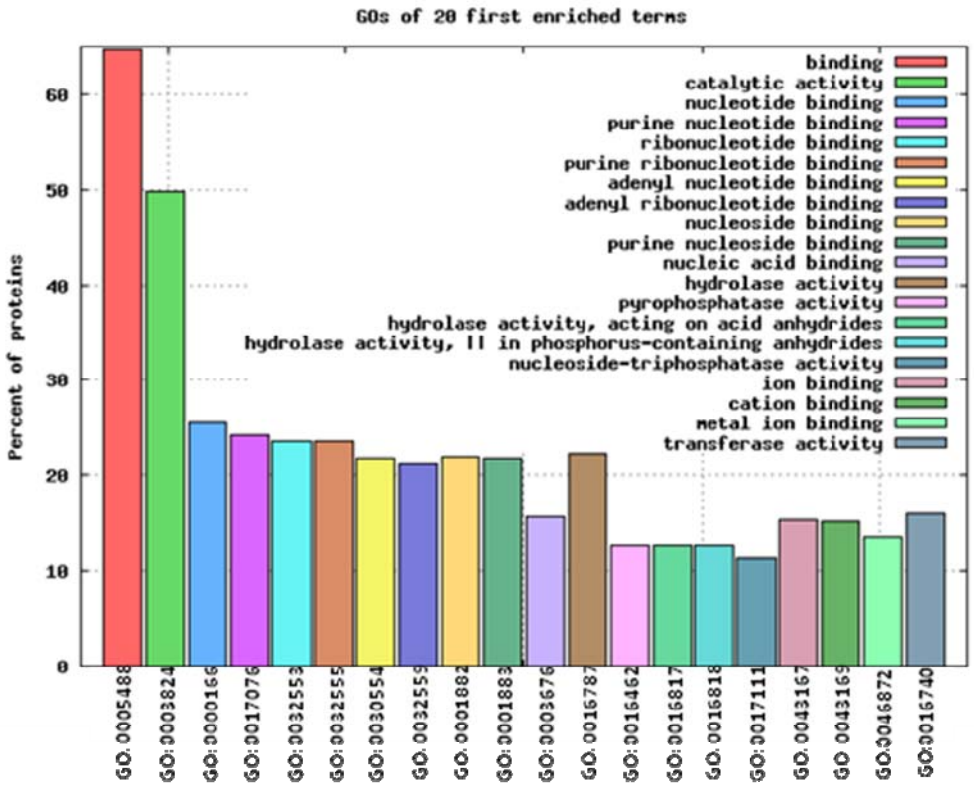


Gráfico 19: 20 primeiros termos GO enriquecidos da categoria função molecular das IUPs de *L. braziliensis*. No eixo X: número GO e na legenda a descrição de cada um. No eixo Y: porcentagem de proteínas.

Através da análise dos 20 termos GO mais enriquecidos na categoria função molecular, observamos que o termo *binding* e atividade catalítica são os mais comuns. Segundo definição do GO, *binding* significa interação seletiva, não covalente, muitas vezes estequiométrica, de uma molécula com um ou mais sítios específicos de outra molécula; e atividade catalítica é a catálise de uma reação bioquímica em temperaturas fisiológicas, nas reações catalisadas biologicamente, os reagentes são conhecidos como substratos e os catalisadores são as enzimas.

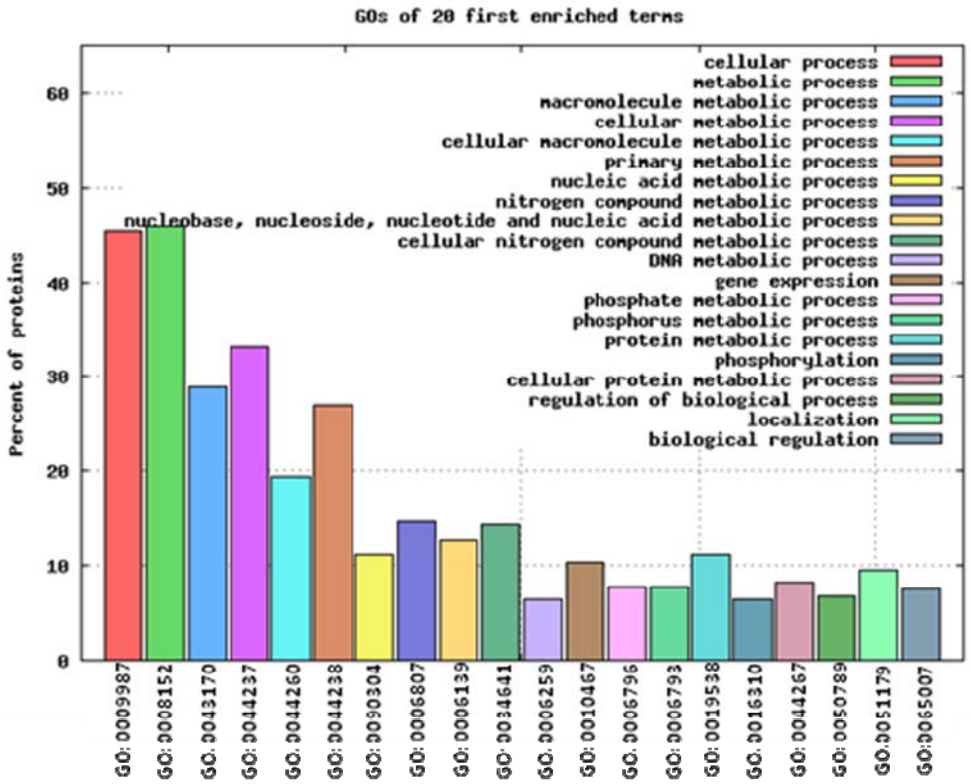


Gráfico 20: 20 primeiros termos GO enriquecidos da categoria processo biológico das IUPs de *L. braziliensis*. No eixo X: número GO e na legenda a descrição de cada um. No eixo Y: porcentagem de proteínas.

Na figura dos 20 primeiros termos GOs enriquecidos na categoria processo biológico, observamos que se destacam os termos processos metabólicos e processos celulares.

Segundo a definição do GO, processos metabólicos são reações químicas e vias, incluindo o anabolismo (parte do metabolismo que se refere à síntese de substâncias em um organismo, ou seja, a partir de moléculas mais simples, são criadas moléculas mais complexas, só ocorre em alta energética, caso esteja em baixa energética, acontece o catabolismo) e o catabolismo (parte do metabolismo que se refere à assimilação ou processamento da matéria orgânica adquirida pelos seres vivos para fins de obtenção de energia), pelo qual os organismos vivos transformam substâncias químicas.

Processos celulares são definidos segundo o GO como qualquer processo que é realizado a nível celular, mas não necessariamente restrita a uma única célula. Por exemplo, a comunicação celular ocorre entre mais de uma célula, mas ocorre a nível celular.

5.2.3.8 Predição de características físico-químicas das IUPs

Gráficos de duas características físico-químicas, a carga e o ponto isoelétrico, são gerados com o objetivo de mostrar o comportamento geral das IUPs nessas duas características (Gráficos 21 e 22 respectivamente).

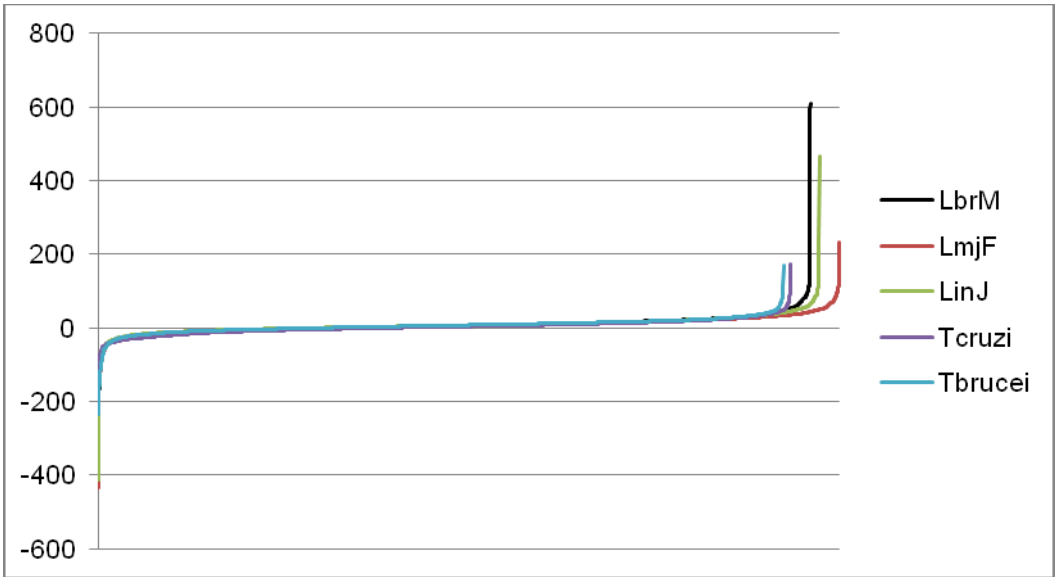


Gráfico 21: Carga das IUPs em *L. braziliensis*. No eixo X: IUPs identificadas; No eixo Y: carga em coulomb (C).

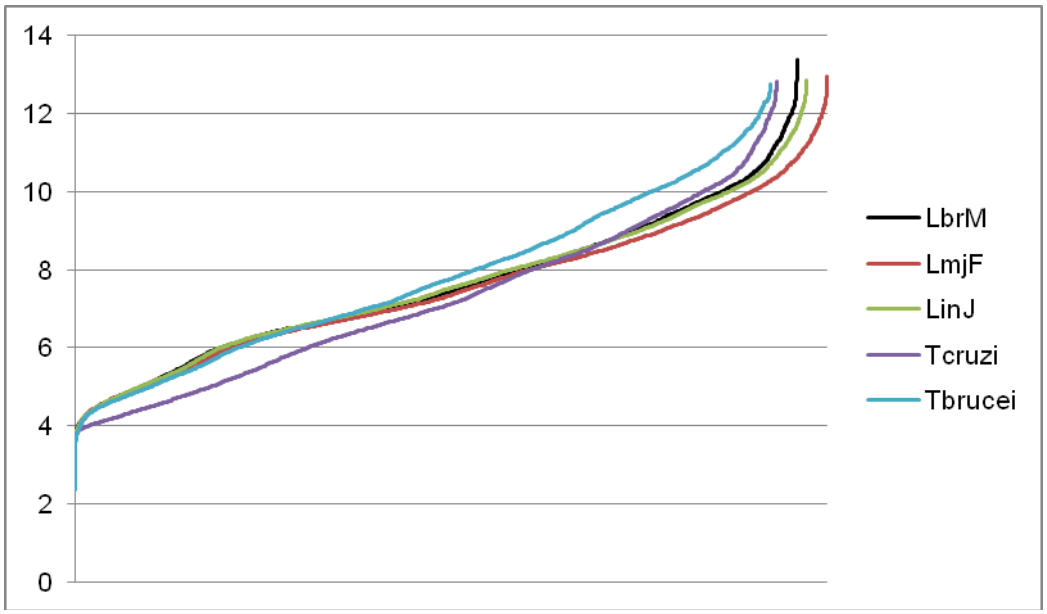


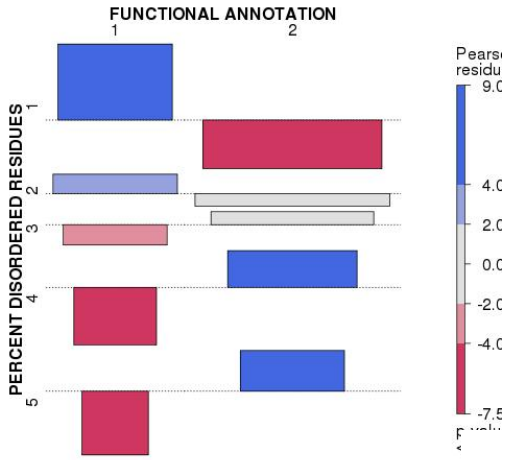
Gráfico 22: Ponto isoelétrico das IUPs em *L. brasiliensis*. No eixo X: IUPs identificadas; No eixo Y: ponto isoelétrico.

Para *L. braziliensis* observamos que para a carga, não há uma tendência das IUPs por cargas positivas ou negativas e para o ponto isoelétrico, notamos que IUPs com pI acima de 7, ou seja, com caráter básico ocorrem em maior quantidade. Os

outros organismos estudados apresentaram o mesmo tipo de comportamento para essas características.

5.3 Arquivos de análise de contingência

Como resultados da análise de contingência, obtivemos gráficos que mostram a associação ($p<0.05$) entre duas variáveis e arquivos com as tabelas de contingência de todas as análises feitas. Os gráficos de associação que apresentaram significado biológico com relação a porcentagem de resíduos desordenados foram: a) predito/hipotético (Gráfico 23); b) ponto isoelétrico (Gráfico 24); c) localização (Gráfico 25) e d) regiões transmembranas (Gráfico 26).



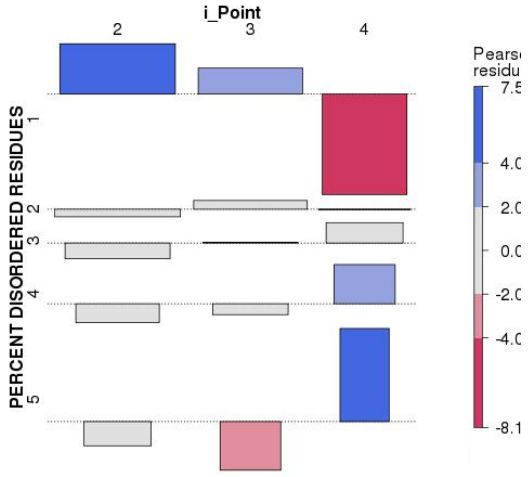
FUNCTIONAL ANNOTATION	
predicted	1
hypothetical	2

PERCENT OF DISORDERED RESIDUES	
> 0 AND <= 0.2	1
> 0.2 AND <= 0.4	2
> 0.4 AND <= 0.6	3
> 0.6 AND <= 0.8	4
> 0.8	5

Gráfico 23: Associação entre porcentagem de resíduos desordenados e a anotação funcional da proteína (predita ou hipotética) nas IUPs de *L. braziliensis*. As cores representam a freqüência maior (azul) ou menor (rosa) que o esperado. Os números representam as categorias dos atributos analisados.

Observamos no gráfico que a primeira categoria de porcentagem de resíduos desordenados apresenta uma freqüência maior que o esperado de proteínas com função predita e uma freqüência menor que o esperado de proteínas hipotéticas. Essa situação se inverte nas categorias seguintes onde a porcentagem de resíduos desordenados aumenta, havendo uma freqüência maior que o esperado de proteínas hipotéticas e uma freqüência menor de proteínas com função predita.

Portanto, conforme a porcentagem de resíduos desordenados aumenta, as IUPs tendem a ser hipotéticas. Perfil semelhante é observado em todos os genomas estudados (gráficos não apresentados).

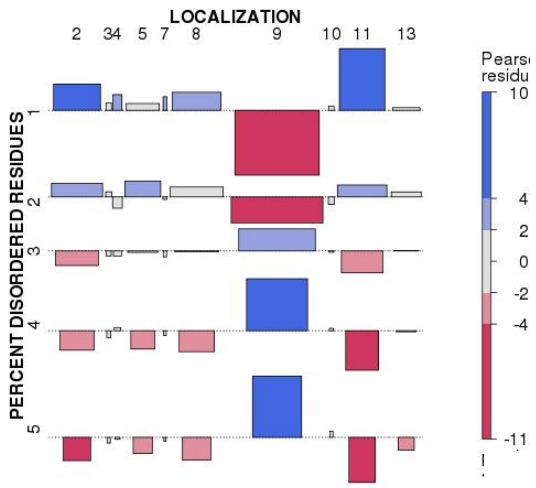


I_POINT	
<= 3	1
> 3 AND <= 7	2
> 7 AND <= 9	3
> 9	4
PERCENT OF DISORDERED RESIDUES	
> 0 AND <= 0.2	1
> 0.2 AND <= 0.4	2
> 0.4 AND <= 0.6	3
> 0.6 AND <= 0.8	4
> 0.8	5

Gráfico 24: Associação entre porcentagem de resíduos desordenados e o ponto isoelétrico nas IUPs de *L. braziliensis*. As cores representam a freqüência maior (azul) ou menor (rosa) que o esperado. Os números representam as categorias dos atributos analisados.

Observamos no gráfico que a primeira categoria que seria o ponto isoelétrico <=3 não está presente, significando que nenhuma IUP se classificou na categoria muito ácida. Notamos também que as categorias 2 e 3 tem uma freqüência maior que o esperado e a categoria 4 tem uma freqüência menor que o esperado, apenas nas proteínas que possuem até 20% de resíduos desordenados.

Este comportamento se inverte conforme aumenta a porcentagem de resíduos desordenados. Assim, observamos que de maneira geral, as proteínas tendem a se tornar mais básicas conforme a porcentagem de resíduos desordenados aumenta. Perfil semelhante é observado em todos os genomas estudados (gráficos não apresentados).



LOCALIZATION	
chloroplast	1
cytosol	2
cytoskeleton	3
endoplasmic reticulum	4
extracellular	5
Golgi apparatus	6
lysosome	7
mitochondria	8
nuclear	9
peroxisome	10
plasma membrane	11
vacuolar membrane	12
without prediction	13

PERCENT OF DISORDERED RESIDUES	
> 0 AND <= 0.2	1
> 0.2 AND <= 0.4	2
> 0.4 AND <= 0.6	3
> 0.6 AND <= 0.8	4
> 0.8	5

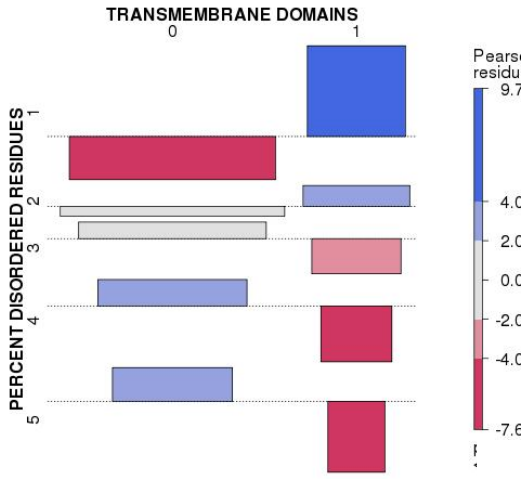
Gráfico 25: Associação entre porcentagem de resíduos desordenados e a localização predita para as IUPs de *L. braziliensis*. As cores representam o quanto a freqüência maior (azul) ou menor (rosa) que o esperado. Os números representam as categorias dos atributos analisados.

Através do gráfico, observamos que a categoria 11 de membrana plasmática tem uma freqüência maior que o esperado nas proteínas com até 20% de resíduos desordenados, sendo que esse comportamento se inverte nas proteínas com mais de 40% de resíduos desordenados.

Considerando a localização nuclear, há uma freqüência menor que o esperado nas proteínas com até 20% de resíduos desordenados sendo que à partir de 40% de resíduos desordenados, essa situação se inverte e há uma freqüência maior que o esperado da localização nuclear para as IUPs.

Para a localização extracelular, observamos uma freqüência maior que o esperado para proteínas com até 40% de resíduos desordenados, comportamento que se inverte para proteínas com mais de 40% de resíduos.

Para a localização citosol, notamos uma freqüência maior que o esperado para proteínas com até 40% de resíduos desordenados, situação que se inverte as demais. Perfil geral semelhante é observado em todos os genomas estudados (gráficos não apresentados).



TRANSMEMBRANE DOMAINS	
without transmembrane regions	0
with transmembrane regions	1

PERCENT OF DISORDERED RESIDUES	
> 0 AND <= 0.2	1
> 0.2 AND <= 0.4	2
> 0.4 AND <= 0.6	3
> 0.6 AND <= 0.8	4
> 0.8	5

Gráfico 26: Associação entre porcentagem de resíduos desordenados e domínios transmembrana para as IUPs de *L. braziliensis*. As cores representam a freqüência maior (azul) ou menor (rosa) que o esperado. Os números representam as categorias dos atributos analisados.

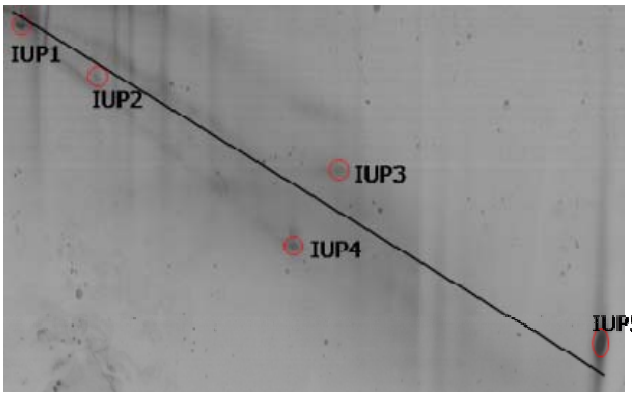
Para a análise relacionada a regiões transmembranas, notamos que a uma freqüência maior que o esperado de IUPs com regiões transmembranas com até 40% de resíduos desordenados. Essa situação se inverte, ou seja, há uma freqüência maior que o esperado de IUPs sem regiões transmembranas com mais de 40% de resíduos desordenados. Perfil geral semelhante é observado em todos os genomas estudados (gráficos não apresentados).

5.4 Análise experimental

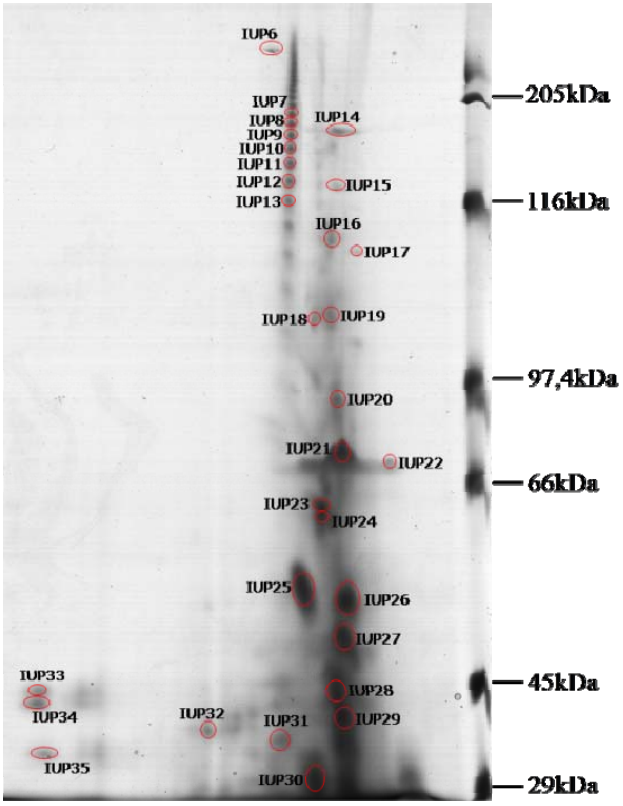
Com o intuito de estabelecer um protocolo inicial de validação experimental, foram realizadas duas metodologias de análise experimental, a desenvolvida por Csizmók e colaboradores (metodologia 1) e a metodologia de Galea e colaboradores (metodologia 2), descritas no item 4.6, onde somente as proteínas do tipo IUP são identificadas.

A metodologia 1 foi realizada somente com o proteoma de *L. major* (Figura 4 - A) pois era necessário um volume muito grande de proteínas para realização do experimento e a metodologia 2 foi utilizada para o proteoma de *L. major* e *L. braziliensis* (Figura 4 – B e C), não utilizamos o proteoma de *L. infantum* para a validação experimental devido a proximidade com *L. major*.

A


IUP1
IUP2
IUP3
IUP4
IUP5

B


IUP6
IUP7
IUP8
IUP9
IUP10
IUP11
IUP12
IUP13
IUP14
IUP15
IUP16
IUP17
IUP18
IUP19
IUP20
IUP21
IUP22
IUP23
IUP24
IUP25
IUP26
IUP27
IUP28
IUP29
IUP30
IUP31
IUP32
IUP33
IUP34
IUP35
205kDa
116kDa
97,4kDa
66kDa
45kDa
29kDa

C

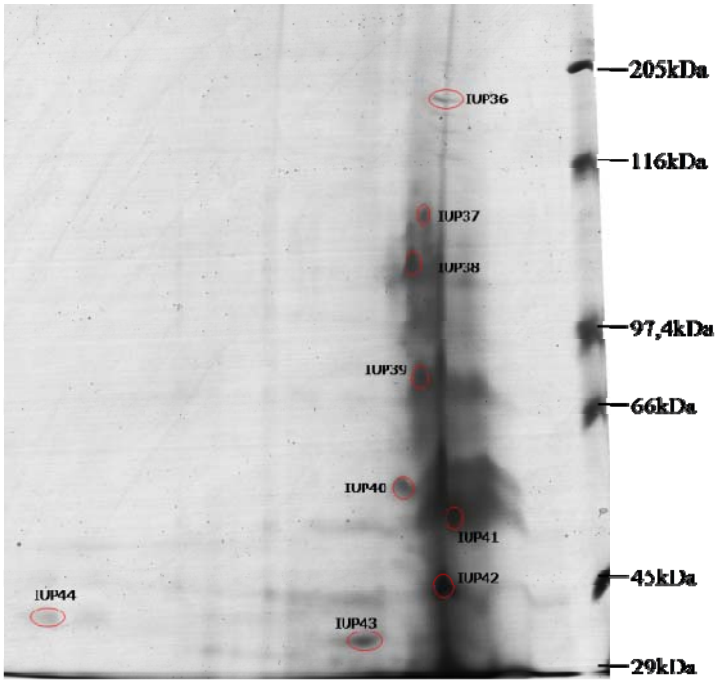


Figura 4: Análise do gel 2D de IUPs em *Leishmania ssp*. (A) O extrato protéico de *L. major* correu em um gel nativo 7,5% na primeira dimensão (com aquecimento por 10 min. em 100ºC) e em um gel 5-20% contendo 8M uréia na segunda dimensão no formato grande (18x16 cm). A segunda dimensão foi corado com colloidal Coomassie blue G-250 por 30 horas. As IUPs identificadas ficaram próximas a linha diagonal. (B) Extrato protéico de *L. major* enriquecido com IUPs por aquecimento a 100ºC por uma hora que foi carregada em uma fita de pH 3-5.6 (13 cm, gradiente não linear). SDS-PAGE foi realizado com um gel 12,5% e corado com colloidal Coomassie blue G-250 por 30 horas. (C) Extrato protéico de *L. braziliensis* enriquecido com IUPs por aquecimento a 100ºC por uma hora que foi carregada em uma fita de pH 4-7 (13 cm, gradiente não linear). SDS-PAGE foi realizado com um gel 12,5% e corado com colloidal Coomassie blue G-250 por 30 horas.

A validação experimental com espectrometria de massa foi feita para 17 spots que foram selecionados dos géis de eletroforese 2D descritos anteriormente. Os resultados estão na tabela 11.

Tabela 11: Validação experimental e *in silico* das IUPs preditas.

GEL #	IUP (SPOT)	ORGANISMO	PROTEINA IDENTIFICADA	FUNÇÃO	ESCORE	ALGORITMO DE PREDIÇÃO DE IUP MELHOR COMBINAÇÃO
						REM465_GLOBPIPE_IUPRED_VSL2B
1	IUP1	*L. major*	LmjF27.0240	kinetoplast-associated protein-like protein	285	X
1	IUP4	*L. major*	LmjF32.2520	hypothetical protein, unknown function	31	X
1	IUP5	*L. major*	LmjF09.0910	calmodulin, putative	505	X
2	IUP6	*L. major*	LmjF27.0240	kinetoplast-associated protein-like protein	435	X
2	IUP8	*L. major*	LmjF05.0380	microtubule-associated protein, putative	188	X
2	IUP10	*L. major*	LmjF05.0380	microtubule-associated protein, putative	449	X
2	IUP11	*L. major*	LmjF05.0380	microtubule-associated protein, putative	228	X
2	IUP25	*L. major*	LmjF23.1020	hypothetical protein, unknown function	1973	X
2	IUP26a	*L. major*	LmjF04.0770	nascent polypeptide associated complex subunit- like protein	1093	X
2	IUP26b	*L. major*	LmjF19.1160	hypothetical protein, conserved	1054	X
2	IUP27	*L. major*	LmjF04.0770	nascent polypeptide associated complex subunit- like protein	1854	X
2	IUP32	*L. major*	LmjF28.2770	heat-shock protein hsp70, putative	608	X
2	IUP34	*L. major*	LmjF30.2460	heat shock 70-related protein 1, mitochondrial precursor, putative	1591	X
3	IUP37	*L. braziliensis*	no hits	-	-	
3	IUP40	*L. braziliensis*	no hits	-	-	
3	IUP42	*L. braziliensis*	LbrM25_V2.0580	eukaryotic initiation factor 5a, putative	2160	X
3	IUP43	*L. braziliensis*	LbrM28_V2.1300	glucose-regulated protein 78, putative	363	X

Os resultados experimentais indicaram que todos os 15 spots de proteína seqüenciados e identificados foram preditos como IUPs também *in silico*.

6 DISCUSSÃO

6.1 IUPs nos Tripanosomatídeos

Proteínas desestruturadas existem sem uma estrutura tridimensional estável e são caracterizadas por conjuntos conformacionais altamente flexíveis. A disponibilidade de seqüências genômicas tem mostrado que a ocorrência de regiões desestruturadas de tamanho significante (>40 resíduos) é surpreendentemente comum em proteínas funcionais (Dunker, Brown *et al.*, 2002) (Uversky, 2002). Contribuindo para esse contexto, a existência de proteínas desestruturadas funcionais como hormônios (Boesch, Bundi *et al.*, 1978) e a observação experimental desse tipo de proteínas em células intactas (Daniels, Williams *et al.*, 1978) evidenciam o papel crucial dessa classe de proteínas.

Proteínas desordenadas têm papel funcional vital em vários processos biológicos como a regulação da transcrição, tradução e transdução de sinais (Wright e Dyson, 1999) (Iakoucheva, Brown *et al.*, 2002) (Iakoucheva, Radivojac *et al.*, 2004) (Tompa, 2002) e sua existência representa um forte argumento motivador para uma releitura do paradigma estrutura-função (Dunker e Obradovic, 2001). Complementarmente, a viabilidade do estudo em larga escala de IUPs só é possível frente às inovações tecnológicas e computacionais que tem gerado nas últimas décadas uma vasta quantidade de dados de seqüência e genomas completos.

Dentro da família *Trypanosomatidae*, parasitos do gênero *Leishmania* e *Trypanosoma* têm importância médica relevante em diferentes países (Donelson, Gardner *et al.*, 1999). Além da inexistência de vacinas efetivas para a prevenção das enfermidades de que são causadores, as drogas usualmente utilizadas no tratamento são altamente tóxicas.

Até o momento não existem estudos de identificação e caracterização em larga escala de desordem estrutural protéica em genomas de tripanosomatídeos fato que já representa um enorme motivador para esse estudo. Considerando ainda a existência de trabalhos que descrevem as propriedades de interações proteína-proteína das IUPs como características importantes e muitas vezes cruciais para sua utilização como promissores alvos terapêuticos, o presente estudo tem apelo cientifico ainda maior.

Nossos resultados indicam uma grande fração desses genomas contendo proteínas desestruturadas (Tabela 10). Considerando que ~70% das proteínas preditas como IUPs não tem função predita, essa constatação pode estar diretamente ligada as matrizes e abordagens utilizadas durante o processo de atribuição computacional de função durante a etapa de anotação genômica automática. Assim, o presente estudo contribui também para anotação desses genomas através da adição dessa nova camada de informação.

O *pipeline* de IUPs desenvolvido, além de permitir uma identificação e caracterização sistemática das IUPs, integra informações, viabilizando, em um contexto mais amplo de análise contribuir para uma escolha mais racional e direcionada de potenciais alvos para drogas.

6.2 *Pipeline* de IUPs

Existem pelo menos duas grandes contribuições do presente estudo: o desenvolvimento de um *pipeline* automático de predição, análise e caracterização de IUPs; e o entendimento do papel da desordem estrutural em tripanosomatídeos.

Com relação ao desenvolvimento do *pipeline* estivemos centrados na integração das análises e na execução de testes específicos definidos por alguns padrões de engenharia de software relacionados à implementabilidade, robustez e confiabilidade na construção e execução de um programa. Nesse contexto, avaliamos todos os possíveis caminhos de execução disponíveis ao usuário final. Com o resultado final desses testes básicos comprovamos que pontos centrais relacionados à integridade dos dados e banco de dados assim como a implementabilidade foram validados (Tabela 12).

Tabela 12: Teste de validação do *pipeline* de IUPs.

	Aprovado	Não aprovado
Instalação de componentes	X	
Validação de caminhos de execução	X	
Teste de entrada e saída de dados	X	
Teste de integridade de dados	X	
Validação de resultados e relatórios	X	

O teste de instalação (implementabilidade do pacote) seguiu exatamente as etapas e os requisitos descritos no arquivo READ-ME (Anexo 1). Assim sendo, esse tutorial permite a instalação do pacote por qualquer usuário que conheça o ambiente Linux. Além disso, o *pipeline* é capaz de lidar com um grande volume de dados de maneira automática e robusta, de acordo com seu propósito inicial. Atualmente o *pipeline* pode ser utilizado por usuários externos ao nosso grupo via solicitação no link: http://iup.cpqrr.fiocruz.br/iup-pipeline.

Outra etapa importante do *pipeline* está relacionada à escolha da extensão das regiões de desordem a ser considerada. Na execução do *pipeline* somente regiões desordenadas acima de 40 aminoácidos foram consideradas e esse critério foi adotado por várias razões.

Inicialmente precisamos considerar que mesmo os arquivos relacionados à estrutura de proteínas do PDB podem conter trechos de coordenadas ausentes. Esses trechos variam de 1 a 30 aminoácidos e ocorrem freqüentemente nas regiões terminais ou de pequenos *loops* da proteína e não representam regiões desestruturadas verdadeiras (Uversky, 2010) (He, Wang *et al.*, 2009).

Por outro lado, no Disprot que é um banco de referência em desordem estrutural, estão presentes somente regiões maiores que 30 aminoácidos.

Por último é necessário que se diga que os dados contidos no CASP (*Critical Assessment of Structure Prediction*) são restritos a uma única metodologia experimental (difração de raios X) e que, além disso, existem poucas proteínas de tripanosomatídeos cristalizadas (He, Wang *et al.*, 2009).

Com relação ao pré-processamento das seqüências, no *pipeline* adotamos como critério de corte a remoção de proteínas com erros de anotação (descritos em materiais e métodos). A principal razão que nos levou ao estabelecimento desse corte esta associado à anotação automática que é muito utilizada atualmente. Devido ao grande volume de dados gerados nos diferentes projetos genoma, poucas as proteínas são curadas manualmente fato que induz a ocorrência e perpetuação de erros de anotação. Como exemplo dessa constatação, temos o *pipeline* de anotação automática MAKER, que tem especificidade de 91% e sensibilidade de 89% na anotação automática de genomas e uma especificidade de 11% e uma sensibilidade de 34% na anotação de códon de início e fim (Cantarel, Korf *et al.*,

2008).

6.3 Freqüência de aminoácidos

Para investigar a composição de aminoácidos presente nas IUPs dos tripanosomatídeos, fizemos uma análise comparativa com base na freqüência de aminoácidos presentes em regiões globulares. Para tanto utilizamos os dados filtrados contidos no PDB_S25 (vide metodologia item 4.3.2.12.1.1). Nossos resultados indicam que nos tripanossomatídeos, os aminoácidos enriquecidos em regiões desordenadas são: P, Q, E, R e S e os empobrecidos são: W, Y, F, V, I, L e C.

Observamos nesses resultados que regiões desordenadas são empobrecidas em aminoácidos hidrofóbicos (I, L e V) e aminoácidos aromáticos (W, Y e F), que normalmente formam o núcleo hidrofóbico das proteínas globulares (Uversky, 2010). A diminuição do conteúdo de Cisteína (C) em regiões desordenadas vai de encontro com o propósito da formação de núcleos hidrofóbicos de proteínas globulares, pois, esse aminoácido freqüentemente ocorre em sítios de ativação ou estabilizando pontes disulfeto que têm um papel importante na manutenção da estrutura de algumas proteínas, o que não é necessário nas IUPs (Tompa, 2002). Portanto esses aminoácidos (I, L, V, W, Y, F e C) são promotores de ordem.

Por outro lado, identificamos aminoácidos promotores de desordem nas regiões desordenadas dos tripanosomatídeos: P, Q, E, R e S.

O enriquecimento de Prolina (P), por exemplo, está relacionado à falta de estrutura, esse aminoácido é conhecido como desfavorecedor de formação de estrutura secundária rígida, está envolvido ativamente nas regiões de interação proteína-proteina (Williamson, 1994) e conseqüentemente tem uma forte preferência por motivos conformacionais abertos, o que mostra a dimensão funcional para a prevalência deste aminoácido nas IUPs, que dependem fortemente do reconhecimento de alvos (Tompa, 2002).

Proteínas desordenadas possuem: baixa concentração de aminoácidos aromáticos (F, W e Y) e hidrofóbicos (A, I, L, M, F, P, W e V) e alta concentração de

aminoácidos polares (S, T, N, Q, Y, C, K, R, H, D e E) e carregados (K, R, H, D e E) (Uversky, 2010).

Contudo existem variações nesse comportamento como o aminoácido Prolina que é enriquecido nas IUPs e é hidrofóbico, mas interrompe estruturas secundárias (Dunker, Lawson *et al.*, 2001) (Williams, Obradovi *et al.*, 2001) (Romero, Obradovic *et al.*, 2001).

Pequenas variações do comportamento geral descrito anteriormente também podem depender: a) do método experimental utilizado para identificar a região (Ressonância Magnética Nuclear, cristalografia de Raio-X e dicroísmo circular) (Williams, Obradovi *et al.*, 2001); b) do comprimento da região de desordem (Radivojac, Obradovic *et al.*, 2004); c) do preditor de desordem estrutural (Uversky, 2010) e d) da localização da região desordenada na seqüência (N-terminal, C-terminal e intermediária) (Li, Romero *et al.*, 1999).

6.4 IUPs e suas funções nos tripanosomatídeos

Dados experimentais têm revelado que para muitas proteínas a função desempenhada depende do grau de desestruturação ao invés do grau de estruturação da proteína (Sugase, Dyson *et al.*, 2007) (Galea, Nourse *et al.*, 2008).

Assim sendo, dada a importância e envolvimento das IUPs em processos biológicos essenciais, investigamos o enriquecimento de termos anotadores de função (termos GO) no conjunto de proteínas preditas como IUPs. Dentre os inúmeros termos GO que encontramos significantemente enriquecidos ($p<0.05$) nas IUPs, escolhemos alguns que descrevemos em maior detalhe a seguir.

Transcrição e regulação da transcrição: Interações proteína-DNA e proteína-proteína são processos centrais durante o processo de transcrição. Vários exemplos de IUPs envolvidas na regulação da transcrição têm sido relatados (Dyson, H. J. e Wright, P. E., 2002) (Dyson e Wright, 2005). Por exemplo, o domínio de ativação C-terminal da proto-oncoproteína (bZIP) é desestruturado e altamente flexível e ainda assim suprime efetivamente a transcrição *in vitro* (Campbell, Terrell *et al.*, 2000).

Em tripanosomatídeos o controle de transcrição é predominantemente pós-transcricional. De fato, imagina-se que os genes são transcritos e processados continuamente e sua expressão é então regulada seja pelo transporte seletivo para o citoplasma, seja pela estabilidade do mRNA ou então pela seleção das seqüências de mRNA que serão traduzidas, através de um mecanismo de mobilização polisomal diferencial. No caso da estabilidade, que é um fenômeno largamente estudado nos tripanosomatídeos, a maior parte dos estudos se concentram em demonstrar o papel da região 3'-não codificante (3'-UTR). Obviamente a seqüência UTR em si não tem o papel regulador, mas sim as proteínas a elas associadas. Estas proteínas poderiam então modular a expressão gênica, participando da seleção das seqüências de RNA mensageiro a serem traduzidas já que várias evidências apontam para um mecanismo de mobilização seletiva de seqüências de mRNA para os polisomos.

Ainda há um vasto campo a ser elucidado e pesquisado relacionado à regulação da expressão gênica em tripanosomatídeos e dentro desse contexto podemos sugerir que as IUPs devam ter um papel importante.

Processamento de RNA e *Splicing*: Em tripanosomatídeos, o processo de *trans-splicing* é responsável pelo processamento dos pré-mRNAs policistrônicos, resultando na individualização de cada gene, de modo que cada mRNA contenha a seqüência *spliced leader* (SL) na extremidade 5´ e uma cauda poly A na extremidade 3'.

O processamento dos pré-mRNAs em tripanosomatídeos seja por *cis* e/ou *trans-splicing* é catalisado pelo *spliceosomo*, uma maquinaria de alto peso molecular composta por ribonucleoproteínas (RNPs) e outras proteínas. As ribonucleoproteínas (U1, U2, U4/U6, U5 e SL) se apresentam na forma de complexos de pequenos RNAs (snRNAs) e proteínas, capazes de catálise de RNA.

O enriquecimento significativo dos termos ribonucleoproteínas, processo metabólico de RNA (processamento e *splicing*), *protein binding* e *ribonucleotide binding* entre outras. Tais resultados estão em completa sintonia com os mecanismos celulares descritos acima.

Aspecto interessante que também corrobora nossos resultados está relacionado ao fato de proteínas envolvidas na "montagem" (pela facilitação do recrutamento de componentes celulares do *spliceosomo*) serem proteínas ricas em

Serina e Arginina. Nossas análises evidenciam o enriquecimento desses aminoácidos (vide gráficos 3, 4, 5, 6 e 7).

Citoesqueleto: um conjunto de estruturas protéicas, microtúbulos e filamentos determinam a estrutura e forma da célula e contribuem para o citoesqueleto. IUPs desempenham papéis cruciais na montagem e função do citoesqueleto. O grau de desordem de proteínas do citoesqueleto é comparável àquelas observadas nas proteínas envolvidas em regulação e sinalização celular (Iakoucheva, Brown *et al.*, 2002).

Flagelo: *Leishmania* e *Trypanosoma* são protozoários flagelados cujo ciclo de vida envolve a infecção seqüencial do inseto vetor e do hospedeiro mamífero. Durante esse processo modificações morfológicas e bioquímicas complexas acontecem sucessivamente e são particularmente importantes para *Leishmania* que possui uma forma extracelular flagelada (no inseto vetor) e uma forma intracelular sem flagelo (no hospedeiro mamífero).

Em bactérias o componente principal é um filamento que varia de 12 a 25nm de diâmetro conhecido como “flagellin”. Tem sido descrito que a desordem estrutural tem um papel crucial na montagem do flagelo bacteriano (Namba, 2001). Durante a formação do flagelo regiões terminais desordenadas (N e C-terminais) adotam uma conformação para formar uma estrutura concêntrica tubular que é em sua maioria constituída por alfa-hélices alinhada paralelamente ao eixo do filamento (Namba, 2001) (Mimori-Kiyosue, Vonderviszt *et al.*, 1997).

Com base nessas observações podemos sugerir que a desordem terminal de proteínas associadas ao flagelo em tripanosomatídeos tem papel importante e similar ao “flagelin”.

6.5 Análise de contingência

No intuito de avaliar possíveis associações entre as inúmeras variáveis associadas à caracterização das IUPs, implementamos no nosso *pipeline* uma abordagem conhecida como análise de contingência.

Nessa estratégia de estudo avaliamos associações com significado biológico tendo como base uma variável relacionada à porcentagem de desordem estrutural (porcentagem de resíduos desordenados).

Observamos um padrão extremamente interessante em todas as variáveis com associações significantes ($p<0.05$). Esse padrão, que surge quando as proteínas alcançam aproximadamente 40% de resíduos desordenados, revela dois grupos que apresentam comportamentos de relação inversa.

Uma das associações significantes foi a que relaciona a porcentagem de resíduos desordenados à classificação funcional associada aos termos hipotético e predito. Dentro dos conjuntos de dados analisados observamos uma inversão na freqüência observada. Os primeiros intervalos de porcentagem de resíduos desordenados apresentam a função predita com freqüência maior que o esperado. Essa freqüência diminui gradativamente com o aumento da porcentagem de resíduos desordenados e se inverte quando se atinge o valor de ~40% de resíduos desordenados.

Comportamento inverso é observado para IUPs com porcentagem de resíduos desordenados maior que 40% onde a freqüência maior que o esperado ocorre IUPs anotadas como hipotéticas.

Aproximadamente 60% dos genomas de tripanosomatídeos é constituído por proteínas hipotéticas (sem similaridade significante com bancos de dados de domínio publico). Considerando o cenário de anotação automática de genomas, podemos sugerir que o aumento da porcentagem de desordem estrutural esteja associado a esse fato, uma vez que as matrizes utilizadas são construídas pela utilização de conjuntos de dados que não contemplam proteínas desordenadas e, além disso, filtros de baixa complexidade são regularmente empregados nas buscas por similaridade de seqüências e eles excluem tais proteínas e/ou regiões das comparações.

Outra associação significante encontrada foi a que relaciona a porcentagem de resíduos desordenados ao ponto isoelétrico. Observamos uma freqüência maior que o esperado de IUPs muito básicas ($pH>9$) com mais de 40% de resíduos desordenados.

A localização subcelular também mostrou uma associação significante com a porcentagem de resíduos desordenados.

Dentro do conjunto de dados analisados observamos uma freqüência maior do que o esperado das classes: membrana plasmática, citosol e mitocôndria e uma freqüência menor que o esperado da classe nuclear. Esse padrão ocorre para IUPs com até 40% de resíduos desordenados. Para IUPs com mais de 40% de resíduos desordenados observa-se uma inversão do padrão acima descrito.

A definição da localização subcelular é importante para o entendimento da função da proteína e é uma etapa crítica no processo de anotação dos genomas.

O conhecimento da localização subcelular de uma proteína pode melhorar significantemente a identificação de alvos durante o processo de descobrimento de drogas, como por exemplo, proteínas secretadas e proteínas de membrana plasmática são facilmente acessíveis por moléculas de drogas devido a sua localização no espaço celular ou na superfície celular e são de interesse pelo seu potencial uso como candidato a vacina ou como alvos para diagnostico.

Na análise da associação da porcentagem de resíduos desordenados e da presença de regiões transmembranas, IUPs com até 40% de resíduos desordenados tem uma freqüência maior que o esperado de regiões transmembrana. Esse resultado vai de encontro ao discutido anteriormente para a variável localização subcelular, onde IUPs localizadas na membrana plasmática apresentam uma freqüência maior que o esperado de regiões transmembrana. Perfil inverso acontece para IUPs com mais de 40% de resíduos desordenados.

6.6 Validação experimental das predições *in silico*

Com o intuito de validar as predições *in silico* realizadas pelo *pipeline* desenvolvido, em colaboração com Dra. Ângela Kaysel Cruz da FMRP – USP, empregamos duas metodologias experimentais de eletroforese bidimensional que identificam IUPs (Csizmók, Szollosi *et al.*, 2006) (Galea, Pagala *et al.*, 2006).

Apesar de não termos seqüenciado todos os *spots* identificados nos géis obtidos, os dados resultantes da espectrometria de massas de 17 *spots* são muito promissores.

Os resultados apresentados na tabela 11 (Validação experimental e *in silico* das IUPs preditas) sumarizam a inter-relação dos dados experimentais e predições *in silico*. Dos *spots* identificados por similaridade de seqüência, 100% deles (15 *spots*) foram preditos como IUPs pela melhor combinação de algoritmos preditores de desordem estrutural (REM465, GlobPipe, IUPRED, VSL2B). Esse resultado fornece indícios de que o *pipeline* desenvolvido assim como a combinação de algoritmos utilizada é extremamente eficaz.

6.7 IUPs como alvos para o desenvolvimento de drogas

Recentemente as IUPs vêm sendo descritas como potenciais alvos de interação de drogas. A baixa energia de interação entre as IUPs e seus potenciais ligantes estruturados facilita a intervenção na interação por outra molécula (droga) (Cheng, LeGall *et al.*, 2006).

Levando-se em consideração que a interação de uma IUP com seu ligante tem caráter fraco e que essa característica é decorrente do fato de parte da energia de ligação ser gasta na organização estrutural da IUP, a interação do ligante estruturado com uma molécula de droga vai ser facilitada quando comparada com o parceiro natural desordenado.

Dados preliminares obtidos através da mineração do banco de dados de IUPs de tripanosomatídeos revelaram a associação de IUPs com proteínas envolvidas na interação parasita/hospedeiro (Romao, Castro *et al.*, 2009) (Marín-Villa, Vargas-Inchaustegui *et al.*, 2008). Tais interações são fundamentais no processo de resposta imune do hospedeiro e têm sido considerado potenciais candidatos a alvos para droga.

Dentro desse contexto fica claro o uso potencial da informação relacionada a desordem estrutural para auxiliar na identificação de potencias candidatos ao desenvolvimento de droga. Alguns exemplos de proteínas identificadas como IUPs

pelo *Pipeline* de IUPs e que relacionam desordem estrutural com alvos descritos na literatura para o desenvolvimento de drogas são:

1-) Ornitina descarboxilase: Essencial para a sobrevivência do parasita. Alvo do DFMO (DiFluoroMethylOrnithine) (Boitz, Yates *et al.*, 2009) (Heby, Persson *et al.*, 2007).

2-) Triparedoxina: Essencial para a sobrevivência do parasita. Em *L. infantum*, envolvido em processos como síntese de DNA ou interação parasita/hospedeiro (Romao, Castro *et al.*, 2009).

3-) Phosphatidyl serina: Papel central na disseminação da infecção; atua como sinal para imersão de macrófagos (Getti, Cheke *et al.*, 2008).

4-) Tripanotiona sintetase: Envolvido na defesa contra estresse químico e oxidativo. Descrito como potencial alvo para droga (Heby, Persson *et al.*, 2007).

5-) Cisteína peptidase: Pode estar envolvido na infecção da célula hospedeira nos estágios iniciais da infecção em *L. pifanoi* e *L. amazonensis* (Marín-Villa, Vargas-Inchaustegui *et al.*, 2008).

7 CONCLUSÕES

As principais conclusões obtidas com o desenvolvimento deste trabalho foram:

1. O desenvolvimento do *pipeline* de identificação, caracterização e análise de proteínas preditas como IUPs cumpre seu objetivo de maneira integrada, segura, automática e organizada;
2. O banco de dados integrado viabilizou o armazenamento de todas as predições realizadas e a conseqüente extração de resultados;
3. As espécies de *Leishmania* estudadas apresentam aproximadamente 70% do proteoma predito como IUPs e as espécies de *Trypanosoma* possuem aproximadamente 55% do proteoma predito como IUPs;
4. As IUPs de tripanosomatídeos são enriquecidas com os aminoácidos P, Q, E, R e S e empobrecidas nos aminoácidos W, Y, F, V, I, L e C;
5. Análises de contingência revelaram a correlação da desordem estrutural com importantes características desses organismos incluindo: localização sub-celular, número de regiões transmembrana, ponto isoelétrico e função predita;
6. Análises funcionais relacionando as ontologias definidas no GO revelaram um enriquecimento significativo dos seguintes termos: ribonucleoproteínas, processo metabólico de RNA (processamento e *splicing*), *protein binding* e *ribonucleotide binding*, entre outras.

8 ANEXOS

8.1 Anexo 1 - README

Pre-requisites for IUP pipeline implementation :
--- Install:
Mysql
Perl -> libraries: Mysql, IO:: File Getopt:: Long, Date: PowerSet, XML:: SAX:: ExpatXS
BioPerl -> libraries: Bio:: SeqIO, Bio:: SearchIO
Predictors of IUPS -> DisEMBL, GlobPipe, IUPred, VSL2B
Predictor of transmembrane region -> Phobius
Prediction of cellular localization -> Wolf pSort
Predictor of physical-chemical characteristics -> EMBOSS -> Pepstats
Predictor of functional annotation -> Blast2GO4Pipe
Blast
Non redundant database of NCBI formatted

--- Creation of the IUPS database
Made by the administrator of the machine!
Run the script create_db_iup-pipeline.pl to create the database with the name you want and then create a user to the bank with the necessary privileges to run the pipeline:

perl create_db_iup-pipeline.pl -d bank_name -u root -r root_password

GRANT SELECT, INSERT, UPDATE ON .* TO bank_name user@localhost IDENTIFIED BY 'password' WITH GRANT OPTION;

--- Creating and loading the GO database
Made by the administrator of the machine!
Run the program create-insert_db_GO-terms.pl for create the database and loads the information and then give privileges to the appropriate user

perl create_db_GO-terms.pl -i GO_file -d GO_terms -u root -r root_password

GRANT SELECT ON .* TO GO_terms user@localhost IDENTIFIED BY 'password' WITH GRANT OPTION;

--- Pipeline Execution
To see the parameters just execute the line:
perl iup_pipeline_V6.pl

9 REFERÊNCIAS BIBLIOGRÁFICAS

Abercrombie BD, Kneale GG, Crane-Robinson C, Bradbury EM, Goodwin GH, Walker JM, et al. Studies on the conformational properties of the high-mobility-group chromosomal protein HMG 17 and its interaction with DNA. Eur J Biochem. 1978 Mar;84(1):173-7.

Ackerman MS, Shortle D. Robustness of the long-range structure in denatured staphylococcal nuclease to changes in amino acid sequence. Biochemistry. 2002 Nov;41(46):13791-7.

Agresti A, Categorical Data Analysis. 2nd ed. Hoboken(NJ): JohnWiley & Sons; 2002.

Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol. 1990 Oct;215(3):403-10.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Res. 2010 Jan;38(Database issue):D457-62.

Bai Y, Chung J, Dyson HJ, Wright PE. Structural and dynamic characterization of an unfolded state of poplar apo-plastocyanin formed under nondenaturing conditions. Protein Sci. 2001 May;10(5):1056-66.

Barlow PN, Vidal JC, Lister MD, Hancock AJ, Sigler PB. Synthesis and some properties of constrained short-chain phosphatidylcholine analogues: (+)- and (-)-(1,3/2)-1-O-(phosphocholine)2,3-O- dihexanoylcyclopentane-1,2,3-triol. Chem Phys Lipids. 1988 Mar;46(3):157-64.

Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, et al. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol. 2000 Nov;7 Suppl:957-9.

Blain SW, Massagué J. Breast cancer banishes p27 from nucleus. Nat Med. 2002 Oct;8(10):1076-8.

Boesch C, Bundi A, Oppliger M, Wüthrich K. 1H nuclear-magnetic-resonance studies of the molecular conformation of monomeric glucagon in aqueous solution. Eur J Biochem. 1978 Nov;91(1):209-14.

Boitz JM, Yates PA, Kline C, Gaur U, Wilson ME, Ullman B, et al. Leishmania donovani ornithine decarboxylase is indispensable for parasite survival in the mammalian host. Infect Immun. 2009 Feb;77(2):756-63.

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics. 2004 Dec;20(18):3710-5.

Bracken C, Iakoucheva LM, Romero PR, Dunker AK. Combining prediction, computation and experiment for the characterization of protein disorder. Curr Opin Struct Biol. 2004 Oct;14(5):570-6.

Brown C, Takayama S, Campen A, Vise P, Marshall T, Oldfield C, et al. Evolutionary rate heterogeneity in proteins with long disordered regions. J Mol Evol. 2002 Jul;55(1):104-10.

Campbell KM, Terrell AR, Laybourn PJ, Lumb KJ. Intrinsic structural disorder of the C-terminal activation domain from the bZIP transcription factor Fos. Biochemistry. 2000 Mar;39(10):2708-13.

Cantarel BL, Korf I, Robb SM, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan;18(1):188-96.

Cary PD, Moss T, Bradbury EM. High-resolution proton-magnetic-resonance studies of chromatin core particles. Eur J Biochem. 1978 Sep;89(2):475-82.

Cheng Y, LeGall T, Oldfield CJ, Mueller JP, Van YY, Romero P, et al. Rational drug design via intrinsically disordered protein. Trends Biotechnol. 2006 Oct;24(10):435-42.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. 2005 Sep;21(18):3674-6.

Cox FE. History of sleeping sickness (African trypanosomiasis). Infect Dis Clin North Am. 2004 Jun;18(2):231-45.

Csizmók V, Szollosi E, Friedrich P, Tompa P. A novel two-dimensional electrophoresis technique for the identification of intrinsically unstructured proteins. Mol Cell Proteomics. 2006 Feb;5(2):265-73.

Daniels AJ, Williams RJ, Wright PE. The character of the stored molecules in chromaffin granules of the adrenal medulla: a nuclear magnetic resonance study. Neuroscience. 1978;3(6):573-85.

Daughdrill GW, Pielak GJ, Uversky VN, Cortese MS, Dunker K. Natively Disordered Proteins. In: Protein Folding Handbook. Darmstadt(Germany): Wiley-VCH; 2005. P.1-128. Disponível em: < http://www.disprot.org/data/detection/NatDisPro.pdf >. Acesso em 27 abr. 2011.

Donelson JE, Gardner MJ, El-Sayed NM. More surprises from Kinetoplastida. Proc Natl Acad Sci U S A. 1999 Mar;96(6):2579-81.

Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol. 2005 Apr;347(4):827-39.

Dosztányi Z, Mészáros B, Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform. 2010 Mar;11(2):225-43.

Dunker A, Brown C, Lawson J, Iakoucheva L, Obradović Z. Intrinsic disorder and protein function. Biochemistry. 2002 May;41(21):6573-82.

Dunker A, Lawson J, Brown C, Williams R, Romero P, Oh J, et al. Intrinsically disordered protein. J Mol Graph Model. 2001;19(1):26-59.

Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 2005 Oct;272(20):5129-48.

Dunker AK, Obradovic Z. The protein trinity--linking function and disorder. Nat Biotechnol. 2001 Sep;19(9):805-6.

Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ. Intrinsic protein disorder in complete genomes. Genome Inform Ser Workshop Genome Inform. 2000;11:161-71.

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

Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005 Mar;6(3):197-208.

El-On J. Current status and perspectives of the immunotherapy of leishmaniasis. Isr Med Assoc J. 2009 Oct;11(10):623-8.

El-Sayed NM, Myler PJ, Blandin G, Berriman M, Crabtree J, Aggarwal G, et al. Comparative genomics of trypanosomatid parasitic protozoa. Science. 2005 Jul;309(5733):404-9.

Emanuelsson O, Nielsen H, Brunak S, von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol. 2000 Jul;300(4):1005-16.

Fawcett T, ROC Graphs: Notes and Practical Considerations for Researchers. Palo Alto(CA): HP Laboratories; 2004

Feng Z, Zhang X, Han P, Arora N, Anders R, Norton R. Abundance of intrinsically unstructured proteins in P. falciparum and other apicomplexan parasite proteomes. Mol Biochem Parasitol. 2006 Dec;150(2):256-67.

Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins. 2006 Oct;65(1):1-14.

Fiers MW, van der Burgt A, Datema E, de Groot JC, van Ham RC. High-throughput bioinformatics with the Cyrille2 pipeline system. BMC Bioinformatics. 2008;9:96.

Fischer E. Einfluss der Configuration auf die Wirkung der Enzyme. Berichte der deutschen chemischen Gesellschaft. 1894;Vol. 27(3):2985-93.

Friendly M. Graphical methods for categorical data. Proceedings of the 17th Annual SAS User Group International Conference; 1992 April 12-15; Honolulu, Hawaii; p.190-200.

Galea C, Pagala V, Obenauer J, Park C, Slaughter C, Kriwacki R. Proteomic studies of the intrinsically unstructured mammalian proteome. J Proteome Res. 2006 Oct;5(10):2839-48.

Galea CA, Nourse A, Wang Y, Sivakolundu SG, Heller WT, Kriwacki RW. Role of intrinsic flexibility in signal transduction mediated by the cell cycle regulator, p27 Kip1. J Mol Biol. 2008 Feb;376(3):827-38.

Galea CA, Wang Y, Sivakolundu SG, Kriwacki RW. Regulation of cell division by intrinsically unstructured proteins: intrinsic flexibility, modularity, and signaling conduits. Biochemistry. 2008 Jul;47(29):7598-609.

Getti GT, Cheke RA, Humber DP. Induction of apoptosis in host cells: a survival mechanism for Leishmania parasites? Parasitology. 2008 Oct;135(12):1391-9.

Green DM, Swets JM. Signal detection theory and psychophysics. New York: John Wiley and Sons Inc.; 1966.

Han P, Zhang X, Feng ZP. Predicting disordered regions in proteins using the profiles of amino acid indices. BMC Bioinformatics. 2009;10 Suppl 1:S42.

Haynes C, Iakoucheva LM. Serine/arginine-rich splicing factors belong to a class of intrinsically disordered proteins. Nucleic Acids Res. 2006;34(1):305-12.

He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK. Predicting intrinsic disorder in proteins: an overview. Cell Res. 2009 Aug;19(8):929-49.

Heby O, Persson L, Rentala M. Targeting the polyamine biosynthetic enzymes: a promising approach to therapy of African sleeping sickness, Chagas' disease, and leishmaniasis. Amino Acids. 2007 Aug;33(2):359-66.

Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci. 1994 Mar;3(3):522-4.

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W585-7.

Hotez PJ, Bottazzi ME, Franco-Paredes C, Ault SK, Periago MR. The neglected tropical diseases of Latin America and the Caribbean: a review of disease burden and distribution and a roadmap for control and elimination. PLoS Negl Trop Dis. 2008;2(9):e300.

Iakoucheva L, Brown C, Lawson J, Obradović Z, Dunker A. Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol. 2002 Oct;323(3):573-84.

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, et al. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004;32(3):1037-49.

Ishida N, Hara T, Kamura T, Yoshida M, Nakayama K, Nakayama KI. Phosphorylation of p27Kip1 on serine 10 is required for its binding to CRM1 and nuclear export. J Biol Chem. 2002 Apr;277(17):14355-8.

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

Klein-Seetharaman J, Oikawa M, Grimshaw SB, Wirmer J, Duchardt E, Ueda T, et al. Long-range interactions within a nonnative protein. Science. 2002 Mar;295(5560):1719-22.

Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE. Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. Proc Natl Acad Sci U S A. 1996 Oct;93(21):11504-9.

Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol. 2004 May;338(5):1027-36.

Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, Weiss S, et al. p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. Nat Struct Mol Biol. 2004 Apr;11(4):358-64.

Li X, Romero P, Rani M, Dunker AK, Obradovic Z. Predicting Protein Disorder for N-, C-, and Internal Regions. Genome Inform Ser Workshop Genome Inform. 1999;10:30-40.

Linding R, Jensen L, Diella F, Bork P, Gibson T, Russell R. Protein disorder prediction: implications for structural proteomics. Structure. 2003 Nov;11(11):1453-9.

Linding R, Russell R, Neduva V, Gibson T. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res. 2003 Jul;31(13):3701-8.

Lynch WP, Riseman VM, Bretscher A. Smooth muscle caldesmon is an extended flexible monomeric protein in solution that can readily undergo reversible intra- and intermolecular sulfhydryl cross-linking. A mechanism for caldesmon's F-actin bundling activity. J Biol Chem. 1987 May;262(15):7429-37.

Marín-Villa M, Vargas-Inchaustegui DA, Chaves SP, Tempone AJ, Dutra JM, Soares MJ, et al. The C-terminal extension of Leishmania pifanoi amastigote-specific cysteine proteinase Lpcys2: a putative function in macrophage infection. Mol Biochem Parasitol. 2008 Nov;162(1):52-9.

Mimori-Kiyosue Y, Vonderviszt F, Namba K. Locations of terminal segments of flagellin in the filament structure and their roles in polymerization and polymorphism. J Mol Biol. 1997 Jul;270(2):222-37.

Namba K. Roles of partly unfolded conformations in macromolecular self-assembly. Genes Cells. 2001 Jan;6(1):1-12.

Obradovic Z, Peng K, Vucetic S, Radivojac P, Brown CJ, Dunker AK. Predicting intrinsic disorder from amino acid sequence. Proteins. 2003;53 Suppl 6:566-72.

Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK. Comparing and combining predictors of mostly disordered proteins. Biochemistry. 2005 Feb;44(6):1989-2000.

Oldfield CJ, Meng J, Yang JY, Yang MQ, Uversky VN, Dunker AK. Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics. 2008;9 Suppl 1:S1.

Palatnik-de-Sousa CB. Vaccines for leishmaniasis in the fore coming 25 years. Vaccine. 2008 Mar;26(14):1709-24.

Peng K, Radivojac P, Vucetic S, Dunker A, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC Bioinformatics. 2006;7:208.

Penkett CJ, Redfield C, Dodd I, Hubbard J, McBay DL, Mossakowska DE, et al. NMR analysis of main-chain conformational preferences in an unfolded fibronectin-binding protein. J Mol Biol. 1997 Nov;274(2):152-9.

Pullen RA, Jenkins JA, Tickle IJ, Wood SP, Blundell TL. The relation of polypeptide hormone structure and flexibility to receptor binding: the relevance of X-ray studies on insulins, glucagon and human placental lactogen. Mol Cell Biochem. 1975 Jul;8(1):5-20.

Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, et al. Protein flexibility and intrinsic disorder. Protein Sci. 2004 Jan;13(1):71-80.

Romao S, Castro H, Sousa C, Carvalho S, Tomás AM. The cytosolic tryparedoxin of Leishmania infantum is essential for parasite survival. Int J Parasitol. 2009 May;39(6):703-11.

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK. Sequence complexity of disordered protein. Proteins. 2001 Jan;42(1):38-48.

Russell RB, Gibson TJ. A careful disorderliness in the proteome: sites for interaction and targets for future therapies. FEBS Lett. 2008 Apr;582(8):1271-5.

Senior K. Chagas disease: moving towards global elimination. Lancet Infect Dis. 2007 Sep;7(9):572.

Shojania S, O'Neil JD. HIV-1 Tat is a natively unfolded protein: the solution conformation and dynamics of reduced HIV-1 Tat-(1-72) by NMR spectroscopy. J Biol Chem. 2006 Mar;281(13):8347-56.

Shortle D, Ackerman MS. Persistence of native-like topology in a denatured protein in 8 M urea. Science. 2001 Jul;293(5529):487-9.

Smyth E, Syme CD, Blanch EW, Hecht L, Vasák M, Barron LD. Solution structure of native proteins with irregular folds from Raman optical activity. Biopolymers. 2001 Feb;58(2):138-51.

Sugase K, Dyson HJ, Wright PE. Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature. 2007 Jun;447(7147):1021-5.

Tompa P. Intrinsically unstructured proteins. Trends Biochem Sci. 2002 Oct;27(10):527-33.

Tompa P, Csermely P. The role of structural disorder in the function of RNA and protein chaperones. FASEB J. 2004 Aug;18(11):1169-75.

Torrieri, R. Identificação e caracterização computacional de proteínas do tipo IUP no proteoma predito de Schistosoma mansoni. Orientação: Jeronimo Conceição Ruiz. Belo Horizonte: [s.n.], 2010. 116 p. Capa dura, 30 cm., il. [Dissertação]. Ministério da Saude. Fundação Oswaldo Cruz. Centro de Pesquisas René Rachou.Programa de Pós-graduação em Ciências da Saúde. Área de concentração: Biologia Celular e Molecular. Disponível em: < http://www.cpqrr.fiocruz.br/texto-completo/D_49.pdf >. Acesso em: 26 abr. 2011.

Tsvetkov LM, Yeh KH, Lee SJ, Sun H, Zhang H. p27(Kip1) ubiquitination and degradation is regulated by the SCF(Skp2) complex through phosphorylated Thr187 in p27. Curr Biol. 1999 Jun;9(12):661-4.

Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci. 2002 Apr;11(4):739-56.

Uversky VN. The mysterious unfoldome: structureless, underappreciated, yet vital part of any given proteome. J Biomed Biotechnol. 2010;2010:568068.

Uversky VN, Gillespie JR, Fink AL. Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins. 2000 Nov;41(3):415-27.

Uversky VN, Permyakov SE, Zagranichny VE, Rodionov IL, Fink AL, Cherskaya AM, et al. Effect of zinc and temperature on the conformation of the gamma subunit of retinal phosphodiesterase: a natively unfolded protein. J Proteome Res. 2002 2002 Mar-Apr;1(2):149-59.

Vucetic S, Brown C, Dunker A, Obradovic Z. Flavors of protein disorder. Proteins. 2003 Sep;52(4):573-84.

Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT. NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry. 1996 Oct;35(43):13709-15.

Williams RM, Obradovi Z, Mathura V, Braun W, Garner EC, Young J, et al. The protein non-folding problem: amino acid determinants of intrinsic order and disorder. Pac Symp Biocomput. 2001:89-100.

Williamson MP. The structure and function of proline-rich regions in proteins. Biochem J. 1994 Jan;297 (Pt 2):249-60.

World Health Organization. Expert Committee on the Control of Chagas Disease. Control of Chagas disease: second report of the WHO expert committee. Genebra: WHO, 2002. 115 p. (WHO technical report series ; 905). ISBN-13 9789241209052. ISBN-10 9241209054.

Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999 Oct;293(2):321-31.

Yuan Z, Bailey TL, Teasdale RD. Prediction of protein B-factor profiles. Proteins. 2005 Mar;58(4):905-12.