BÁRBARA NOBREGA RODRIGUES

AVALIAÇÃO QUANTITATIVA DE SISTEMAS PREDITORES DE FUNÇÃO DE PROTEÍNAS

Dissertação apresentada como requisito parcial para obtenção do grau de Mestre pelo Programa de Pós-Graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná, área de concentração Bioinformática.

Orientador: Professora Doutora Jeroniza Nunes Marchaukoski

Co-orientador: Professora Doutora Maria Berenice Reynaud Steffens

CURITIBA
2013

Rodrigues, Bárbara Nobrega
R696 Avaliação quantitativade sistemas preditores de função de proteínas / Bárbara Nobrega Rodrigues. - Curitiba, 2013
108 f.: il., tabs, grafs.

Orientadora: Profa. Dra. Jeroniza Nunes Marchaukoski
Co-orientadora: Profa. Dra Maria Berenice Reynaud Steffens
Dissertação (Mestrado) – Universidade Federal do Paraná, Setor de Educação Profissional e Tecnológica, Curso de Pós-Graduação em Bioinformática.
Inclui Bibliografia.

1. Síntese protéica. 2. Seqüencia de nucleotidios. 3. Sistemas de predição de função protéica. 4. Bioinformática. I. Marchaukoski, Jeroniza Nunes. II. Steffens, Maria Berenice Reynaud. III.Título. IV. Universidade Federal do Paraná.

CDD 574.19296

TERMO DE APROVAÇÃO

BÁRBARA NOBREGA RODRIGUES

Avaliação Quantitativa de Sistemas Preditores de Função de Proteínas

Dissertação aprovada como requisito parcial para obtenção do grau de Mestre em Bioinformática, pelo Programa de Pós-graduação em Bioinformática, Setor de Educação Profissional e Tecnológica, da Universidade Federal do Paraná, pela seguinte banca examinadora:

Orientadora: Profª Drª Jeroniza Nunes Marchaukoski

Coorientadora: Profª Drª Maria Berenice R.Steffens

Dr. Marco Antonio Seiki Kadowaki
Instituto de Física de São Carlos- USP

Dr. Paulo Afonso Bracarense Costa
Departamento de Estatística- UFPR

Curitiba, 15 de fevereiro de 2013

DEDICATÓRIA

AO MEU COMPANHEIRO DE JORNADA,

AOS MEUS PAIS

E ÀS OUTRAS PESSOAS QUE TAMBÉM ME AJUDARAM A TORNAR POSSÍVEL ESTE TRABALHO.

AGRADECIMENTOS

Agradeço ao programa de pós-graduação pela oportunidade, à professora doutora Jeroniza Nunes Marchaukoski e à professora doutora Maria Berenice Reynaud Steffens pela orientação e auxílio. Ao professor Wagner Hugo Bonat pela indispensável ajuda com a parte estatística.

À CAPES, MEC e ao CNPq pelo apoio a pesquisa, fornecendo auxílio financeiro e auxílio com equipamentos. À secretária Léa pela disponibilidade em ajudar. À secretária Susana pela amizade e por ajudar em tudo o que estava ao seu alcance. Aos colegas de laboratório pela convivência e conversas.

À Kátia Paiva Lopes, pela amizade e pelo seu exemplo. Aos meus pais pelo apoio, pois chegaram até a ler partes da minha dissertação. Aos amigos Larissa Martins Lopes, Thiago Abreu e Rafael Pereira pelo apoio, torcida e incentivos.

À amiga Tatiana Machado de Souza por sempre se disponibilizar a discutir detalhes do meu trabalho. Ao Rafael Antonio Covre por me ajudar durante todo o mestrado, em todos os aspectos do desenvolvimento do trabalho e mais um pouco.

“Nunca ande pelo caminho traçado, pois ele conduz somente até onde os outros já foram.”

Alexandre Graham Bell

RESUMO

Este trabalho propõe aprofundar as avaliações dos sistemas preditores de função de proteína que utilizam como dado de entrada a sequência de uma proteína, podendo na sua conclusão determinar qual a melhor estratégia de predição. Foram selecionados os seguintes sistemas preditores Blast2GO, InterProScan, Panther Score, Pfam scan e ScanProsite. Os critérios utilizados para a seleção dos sistemas foram: 1- compor o conjunto de sistemas relatados em revisões literárias sobre predição de função de proteína, 2- possuir número de citações superior a 500, 3- ser sistema com a última atualização em menos de 3 anos e 4- análises automatizadas sem interação humana. Os 12 conjuntos de sequências selecionados foram: 690 sequências de proteínas enzimáticas, 487 sequências não enzimáticas, 85 sequências proteicas bifuncionais, 358 sequências proteicas de Aminergic GPCR, 412 sequências proteicas de NHR e 153 sequências proteicas de “Secretin-like”, 927 sequências proteicas de enolase, 262 sequências proteicas de crotonase, 389 sequências proteicas de haloacid dehalogenase, 145 sequências proteicas de vicinal oxygen chelate, 145 sequências proteicas de radical SAM and 863 sequências proteicas de padrão-ouro. Os resultados obtidos mostram divergências entre os resultados dos sistemas avaliados. São observadas diferenças entre nível de descrição da função, grafias distintas para uma mesma anotação e divergência de classificação. Os programas que apresentaram maior semelhança foram o Panther Score e o ScanProsite, embora a semelhança média entre os diferentes conjuntos testados seja inferior a 40%. O com maior acurácia foi o Blast2GO, mas com a acurácia máxima de 34%. Não houve nenhuma sequência em a classificação foi unanime para os 5 programas testados.

Palavras-chave: Sistemas de predição de função proteica. Comparação. Recursos de predição de função proteica. Caracterização de sequências.

ABSTRACT

This work assessed predict function systems that utilize a protein sequence as query data and return one function name, being possible to determine which is the best prediction strategy. Selected resources are Blast2GO, InterProScan, Panther Score, Pfam scan and ScanProsite. Selection criteria of these systems were: 1- mention in protein function prediction reviews, 2- more than 500 citations, 3- at least one update on tree years, and 4- to be completely automated system without human intervention. The 12 selected protein sequences sets were: 690 enzymes, 487 non-enzymes, 85 bifunctional, 358 aminergic GPCR, 412 of NHR, 153 secretin-like, 927 enolase, 262 crotonase, 389 haloacid dehalogenase, 145 vicinal oxygen chelate, 145 radical SAM and 863 gold standard. The results demonstrate divergence between the tested systems for function prediction and between the systems and the standard. The divergences are found in level of function description, spelling of the same terms and distinct classification. The Panther Score and the ScanProsite produced the most similar classifications, but this mean similarity is less than 40%. Blast2GO exhibit the highest accuracy, with the maximum value 34%. There are not sequences with the same output from the 5 programs.

Key-words: Predict protein function systems. Comparison. Resources for predict protein function. Sequence characterization.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

AIC - *Akaike Information Criterion* (critério de informação de Akaike).
ANOVA - Análise de variância
DAG - *Directed Acyclic Graph*
DNA - *Deoxyribonucleic Acid* (ácido desoxirribonucleico)
EC - *Enzyme Commission*
FASTA - Padrão de formato de arquivo de sequência
FunCat - *Functional Cataloge*
GI - *GenInfo Identifier*
GO - *Gene Ontology*
HMM - *Hidden Markov Model* (Modelo Oculto de Markov)
HMMER - Sistema de busca por sequências similares baseado em perfis de HMM
JRE - Java *Runtime Environment*
KEGG - *Kyoto Encyclopedia of Genes and Genomes*
MLG - Modelo Linear Generalizado
mRNA - *Messenger RNA* (ácido ribonucleico mensageiro)
NCBI - *National Center for Biotechnology Information*
NR - *Non-Redundant* (base de dados do NCBI)
PDB - *Protein Data Bank*
PDBID - PDB *identifier* (identificador único composto por 4 caracteres)
PFP - *Protein Function Prediction*
RNA - *Ribonucleic Acid* (ácido ribonucleico)
ROC - *Receiver Operanting Characteric* (análise de desempenho)
SFLD - *Structure-Function Linkage Database*

SUMÁRIO

1 INTRODUÇÃO

A catálise biológica foi descoberta em 1794 e em 1876 ela foi vinculada a ação de certas proteínas, as enzimas (LAIDLER, 1986; TIPTON; BOYCE, 2000), Paralelamente, o conceito de gene se desenvolveu a partir 1865 e em 1909 o termo gene aparece pela primeira vez descrito por Wilhelm Johannsen. Em 1941, é proposto o conceito “um gene, uma enzima” que em seguida se torna “um gene, um polipeptídio”, ligando o conceito de gene ao de proteína (GERSTEIN *et al.*, 2007).

Em 1955, o gene é descrito como uma estrutura física e correspondente a uma região definida do genoma. Em 1958, foi formulado dogma central da biologia molecular, que propõe que a informação contida no gene é transcrita em uma molécula de ácido ribonucleico (RNA mensageiro - mRNA) que é então traduzida em uma proteína, ou é transcrita em um RNA funcional. Em 1986, a maioria dos genes que codificam proteínas pode ser identificada nas sequências genômicas dos diferentes organismos por apresentar um padrão de fase de leitura aberta (*open reading frame* - ORF) (GERSTEIN *et al.*, 2007). A partir da década de 1990, as sequências gênicas são definidas como entidades anotadas no genoma, que podem ser identificadas por ferramentas computacionais. Tanto suas sequências quanto a de seus produtos são armazenadas em bases de dados, com ênfase maior nos genes codificantes de proteínas (GERSTEIN *et al.*, 2007). Nessas bases de dados podem estar presentes informações na forma de uma sequência de caracteres, espécie de origem, o genoma de origem, nomenclatura, o mRNA e a sequência do produto funcional relacionado. Dentre essas informações, a nomenclatura é a que reflete o vínculo do gene com seu produto, pois geralmente ela descreve a função que o gene apresenta no organismo em que é expresso, representando sua contribuição na determinação do fenótipo (BLABY-HAAS; CRÉCY-LAGARD, 2011).

As análises de expressão gênica, como análises de microarranjos e RNA-seq, realizam a caracterização experimental da função de produtos da expressão gênica, permitindo a definição da função para os genes estudados (BLABY-HAAS; CRÉCY-LAGARD, 2011). Porém, esses estudos contemplam a definição da função de uma pequena parcela dos genes pertencentes à crescente quantidade de genomas sequenciados (BLABY-HAAS; CRÉCY-LAGARD, 2011; CHITALE; KIHARA, 2011; GERLT *et al.*, 2012; NIKOLOSKI *et al.*, 2011; SCHNOES *et al.*,

2009; VALENCIA, 2005). Para as sequências gênicas sem dados experimentais, a definição de uma nomenclatura ou anotação geralmente envolve análises computacionais, utilizando as informações disponíveis de sequências conhecidas.

A análise computacional mais difundida é a por similaridade entre sequências, em que sequências similares teriam função conservada por apresentarem uma sequência ancestral comum, segundo o princípio de homologia. Essa análise compara uma sequência de resíduos com outras sequências presentes em alguma base de dados, identificando quais sequências conhecidas são mais similares (CHITALE; KIHARA, 2011; GERSTEIN *et al.*, 2007; NIKOLOSKI *et al.*, 2011; SCHNOES *et al.*, 2009).

As bases de dados utilizadas nos diferentes tipos de análises computacionais podem ser curadas ou não. Ao se utilizar uma base não curada para, por exemplo, transferir a nomenclatura de uma sequência conhecida para uma recém descoberta, assume-se o erro da acurácia do banco de dados e o erro do experimento que deu origem aquela nomenclatura (BLABY-HAAS; CRÉCY-LAGARD, 2011; FRIEDBERG, 2006; SCHNOES *et al.*, 2009). Há uma atribuição de erro em uma taxa conhecida e crescente no depósito de sequências em bases de dados, como por exemplo, a base de dados NR passou de 20% de sequências com anotação incorreta em 2000 para 40% em 2005 (SCHNOES *et al.*, 2009). Mesmo em métodos que não utilizam o princípio de homologia ou utilizam homologia agregada com outras metodologias, a predição da função está relacionada à qualidade da base de dados a qual a metodologia está vinculada (FONTANA *et al.*, 2009).

Outras metodologias elencam várias possibilidades, deixando a decisão final ao pesquisador (ALTSCHUL *et al.*, 1997; HAWKINS *et al.*, 2006), mas dificultam análises automatizadas, que permitem a caracterização de novas sequências no ritmo de sua descoberta, viabilizando análises posteriores ao sequenciamento. Tanto para métodos automáticos quanto manuais a definição de uma nomenclatura correta evita a propagação de erro e permite outras análises (FRIEDBERG, 2006).

A ambiguidade que a nomenclatura em si pode apresentar dificulta análises posteriores. Diferente da sequência, a nomenclatura de um determinado gene ou proteína pode variar dependendo de diversos aspectos, como por exemplo, do organismo fonte ou se o nome descreve sua função bioquímica ou sua ação química (FRIEDBERG, 2006; NATURE PUBLISHING GROUP, 2003; TAMAMES;

VALENCIA, 2006; TSURUOKA; MCNAUGHT; ANANIADOU, 2008). Há ainda sequências que apresentam mais de uma função, em que a atribuição de um nome identificando uma de suas funções gera uma representação limitada (FRIEDBERG, 2006; JEFFERY, 1999).

Existem vários estudos que vão desde aqueles que buscam minimizar os efeitos da variação da nomenclatura em análises até estudos determinando a taxa de erro em bases de dados, todos abordam diferentes problemas relacionados à função de genes e proteínas (SCHNOES *et al.*, 2009; TSURUOKA; MCNAUGHT; ANANIADOU, 2008). E dentre esses estudos, os métodos de predição de função buscam agilidade e eficiência desenvolvendo meios para reduzir o efeito desses problemas. As técnicas desenvolvidas por esses diferentes métodos de predição podem ser encontradas em diferentes revisões (BLABY-HAAS; CRÉCY-LAGARD, 2011; HENRY et al., 2011; PANDEY; KUMAR; STEINBACH, 2006; RENTZSCH; ORENGO, 2009). Estes estudos descrevem as metodologias, fundamentações teóricas e outras características geralmente extraídas dos artigos de cada metodologia. Porém não há uma análise comparando os resultados ou o desempenho de diferentes programas sem utilizar dados de validação descritos em artigos desses programas. Propostas similares a essa são encontradas em trabalhos comparando bases de dados (SCHNOES *et al.*, 2009), ferramentas de análise de ortologia (CHEN *et al.*, 2007) e de agrupamento de sequências (SIKIC; CARUGO, 2010).

2 OBJETIVOS

Realizar uma análise comparativa e quantitativa dos programas Blast2GO, InterProScan, Panther Score, Pfam scan e ScanProsite utilizando conjuntos de sequências bem caracterizadas e com nomenclatura funcional definida. Pela avaliação da quantidade de sequências preditas corretamente, considerando o tempo de execução, as divergências de resultados e as características de cada sistema, determinar qual dos sistemas é o mais indicado para realizar a predição de função de proteínas.

3 JUSTIFICATIVA

Uma comparação quantitativa permite observar as diferenças práticas de se optar por um dos programas de predição de função analisados, agregando um conhecimento distinto dos comparativos existentes, que tem um enfoque nas vantagens e desvantagens de cada programa destacando aspectos da metodologia de cada programa, por uma análise qualitativa.

4 REVISÃO DA LITERATURA

4.1 NOMENCLATURA DE GENES E SEUS PRODUTOS

A nomenclatura dos genes geralmente é herdada de seus produtos, que tem seu nome associado com a função que desempenham (TIPTON; BOYCE, 2000). A FIGURA 1 ilustra, como exemplo, a sequência de eventos que envolvem a síntese de enzimas e culminam com a definição da função. No contexto de um genoma, a nomenclatura representa uma classificação sistemática que segue padrões e regras de, por exemplo, um comitê de nomenclatura. Existem comitês de nomenclatura voltados principalmente para a nomenclatura em espécies modelo, como por exemplo, Human Genome Organisation Nomenclature Committee (HGNC), Mouse Nomenclature Committee (MGNC), Saccharomyces Genome database (SGD), The Arabidopsis Information Resource (TAIR) e FlyBase (Database of *Drosophila* Genes & Genomes) (NATURE PUBLISHING GROUP, 2003). Esta nomenclatura geralmente reflete o organismo do qual a sequência originou-se e seu papel nesse organismo (Quadro 1).

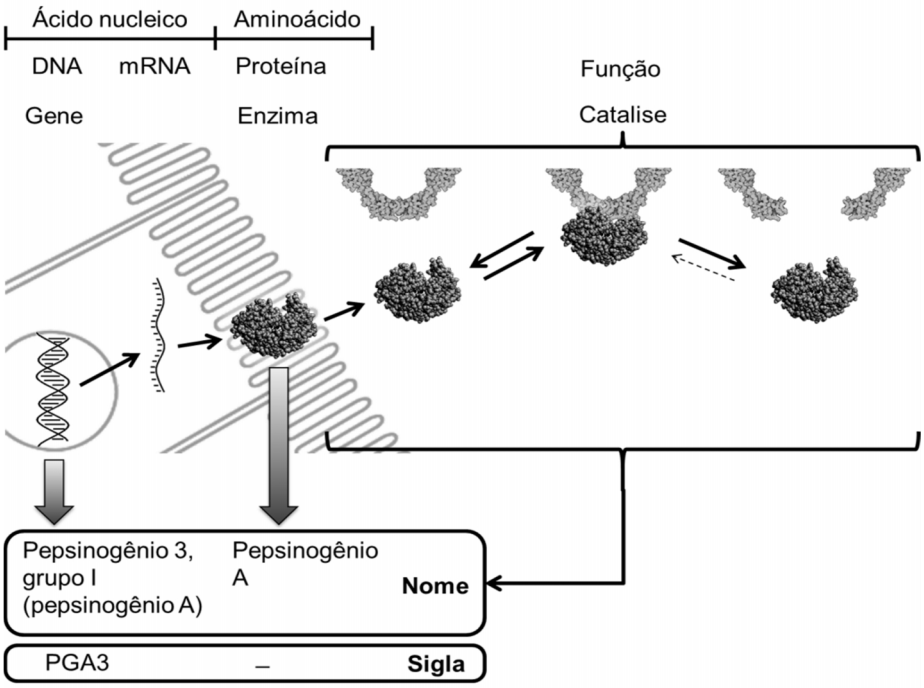


FIGURA 1 - TRANSCRIÇÃO E TRADUÇÃO DE UM GENE. A informação contida no gene é transmitida para seu produto funcional através do mRNA. No caso, o produto funcional trata-se de uma enzima cuja função influência no nome do gene. FONTE: adaptado KEGG (2012) e NCBI (2012).

FOCO	CITAÇÃO	SITE
Homo sapiens	Wain*et al.* (2002)	http://www.genenames.org/guidelines.html
Arabidopsis thaliana	Meinke e Koomneef (1997)	http://www.arabidopsis.org/portals/nomenclature/guidelines.jsp#symbol
Mus musculus	Maltais*et al.* (1997)	http://www.informatics.jax.org/mgihome/nomen/
Saccharomyces cerevisiae	Wood (1998)	http://www.yeastgenome.org/help/community/gene-registry
Bactérias	Demerec*et al.* (1968)	-
Drosophila sp.	Lindxley e Zimm (1992)	http://flybase.org/static_pages/docs/nomenclature/nomenclature3.html

QUADRO 1 - EXEMPLOS DE RECURSOS QUE FORNECEM REGRAS DE NOMENCLATURA PARA ALGUMAS ESPÉCIES DE INTERESSE. FONTE: TAIR web site e Reddien *et al.* (2008).

Existe também a nomenclatura funcional que pode ser atribuída a uma sequência e que geralmente corresponde à função molecular (RENTZSCH; ORENGO, 2009). A padronização dessa nomenclatura é abordada por sistemas de vocabulário controlado. Esses sistemas podem ser destinados a todo tipo de produto gênico, como o *Gene Ontology* (GO) e o *Functional Cataloge* (FunCat), ou para um grupo específico, como o *Enzyme Commission* (EC) voltado para a classificação de enzimas (TIPTON; BOYCE, 2000).

O EC provê uma classificação hierárquica (FIGURA 2a) representando a reação que a enzima catalisa (ANEXO 1). O FunCat também é uma classificação hierárquica, mas com estrutura em árvore para representar as funções de produtos da expressão gênica (ANEXO 2). O GO é uma classificação hierárquica em DAG (*Directed Acyclic Graph*) (FIGURA 2b) separada em três categorias principais: função molecular, processo biológico e componente celular (ANEXO 3) (TIPTON; BOYCE, 2000). Com essas três categorias o GO visa caracterizar produtos gênicos através de atributos (ANEXO 4).

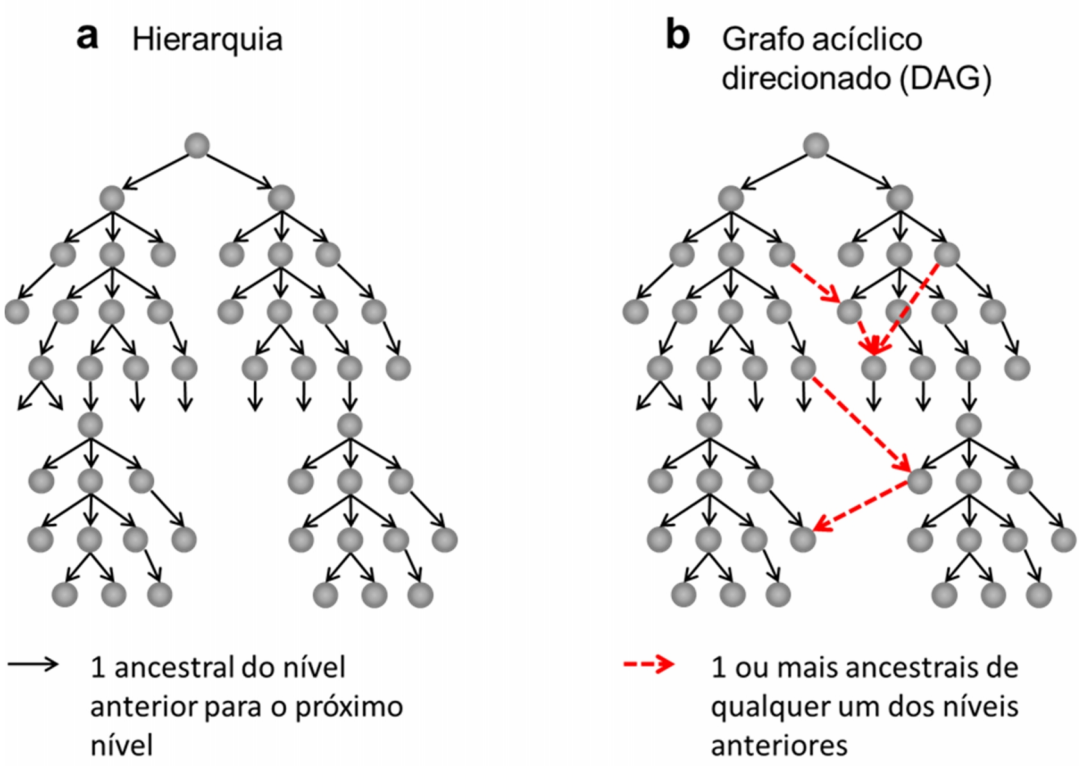


FIGURA 2 - DIFERENÇAS ENTRE ESTRUTURA DE ONTOLOGIA EM HIERARQUIA E DAG. Estrutura (a) hierárquica e em (b) grafo acíclico direcionado (directed acyclic graph - DAG). FONTE: adaptado de Bard (2003).

Esses sistemas buscam reduzir a ambiguidade, reduzir nomes redundantes e possibilitam a interpretação automática desses nomes (CHITALE; KIHARA, 2011; FRIEDBERG, 2006). Porém representam a função na forma de "linguagem humana", que pode ser ambígua para seres humanos e principalmente para máquinas (FRIEDBERG, 2006). Essa ambiguidade, presente nas diferentes formas de se classificar uma sequência, pode ser representada desde diferentes grafias para um mesmo nome até nomes iguais representando sequências distintas (Quadro 2) (FRIEDBERG, 2006; NATURE PUBLISHING GROUP, 2003; TAMAMES; VALENCIA, 2006). Para organismos eucariotos, foi observada uma preferência dos autores em utilizar sinônimos ao invés de nomes oficiais o que dificulta a utilização do conhecimento existente e a correta identificação de novas sequências. Como exemplo, em genes de camundongo foi encontrado 85,1% de ambiguidade com outros genes (CHEN; LIU; FRIEDMAN, 2005).

Existem diversos esforços para obter o significado dos nomes de genes e proteínas através de métodos computacionais supervisionados ou automáticos (LIU *et al.*, 2006; PILLET *et al.*, 2005). Abordagens de normalização dos nomes permitem igualar diferentes formas de redigir um mesmo nome e abordagens de entidades

nominais para harmonizar anotação de genes e proteínas, exemplificam estudos abordando esse tema (CAMPOS *et al.*, 2012; TSURUOKA; MCNAUGHT; ANANIADOU, 2008).

Tipo	Sigla Oficial	Descrição	Termo também encontrado
Sinônimos	PVR	Receptor poliovírus	CD155 (CD de *cell-surface protein*)
	TFF	Fator intestinal trefoil	ITF
	CASP1	Caspase 1	ICE (protease interleukin-1β-converting enzyme)
Acrônimos iguais	PAP	Refere-se a 5 genes humanos diferentes	-
Homônimos	Mad	Enzima mioadenilato deaminase	-
	Mad	Fator de transcrição *mothers against decapentaplegic*	-
Diferenças de grafia	IL2	Interleucina 2	IL-2
	IL3	Interleucina 3	IL-3
Diferença de enfoque	-	Quinase	Phosphorylation of a hydroxyl group

QUADRO 2 - EXEMPLOS DE AMBIGUIDADE QUE PODEM SER ENCONTRADOS EM NOMENCLATURA DE SEQUÊNCIAS. FONTE: Adaptado de Tamames e Valencia (2006), Nature Publishing Group (2003), Tsuruoka *et al.* (2008) e Friedberg (2006).

4.2 PROTEÍNA E FUNÇÃO

Proteínas são moléculas constituídas por resíduos de aminoácidos ligados por ligações peptídicas em que a identidade e posição de cada resíduo é determinada por uma trinca de nucleotídeos no gene e no mRNA, respectivamente. Seu papel dentro do organismo está relacionado a uma série de características: sequência de aminoácidos, estrutura, superfície, sítio ativo, regiões de interação, localização e via metabólica a que pertence. A função proteica normalmente se refere à função molecular representada por reações, substratos e atividades que uma proteína pode apresentar, podendo ser uma enzima, uma proteína estrutural, de transporte ou transmembrana (RENTZSCH; ORENGO, 2009).

A definição da função de uma sequência proteica pode apresentar algumas dificuldades, como falta de dados experimentais, existência de proteínas com mais de uma função e proteínas derivadas da fusão de genes (BLABY-HAAS; CRÉCY-

LAGARD, 2011; JEFFERY, 1999; MARCOTTE *et al.*, 1999). Nestas últimas a função pode não ser a mesma dos genes a qual ela é homóloga. Por exemplo, a topoisomerase II de levedura que apresenta parte da sequência homologa a girase B de *Escherichia coli* e parte homóloga a girase A de *E. coli* (MARCOTTE *et al.*, 1999) (FIGURA 3). No entanto, a sequência refere-se a uma topoisomerase, não a uma girase. A girase B apresenta uma atividade de ATPase. E a girase A e B de *E. coli* estão relacionadas quebra e religação das cadeias de DNA e são subunidades da topoisomerase II de *E. coli*. A topoisomerase II de levedura além de estar envolvida na quebra e religação das cadeias de DNA, para amenizar o superenovelamento, também é responsável pela localização do núcleo axial na meiose (NCBI, 2012).

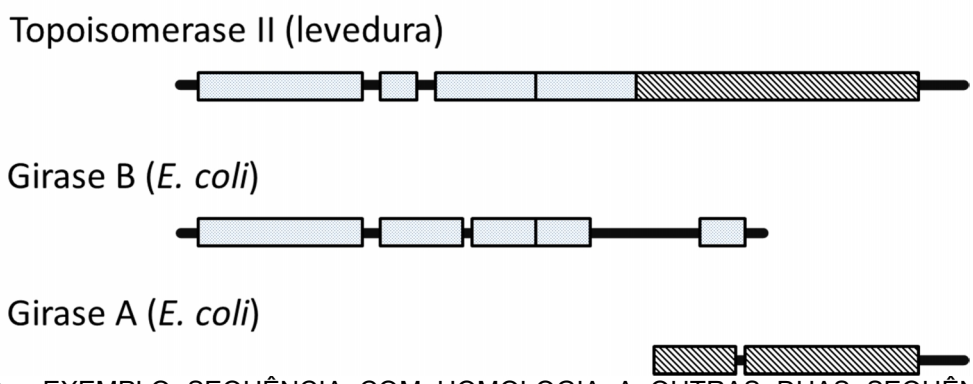


FIGURA 3 - EXEMPLO SEQUÊNCIA COM HOMOLOGIA A OUTRAS DUAS SEQUÊNCIAS. A topoisomerase II de levedura apresenta parte de sua sequência homologa a girase b de *E. coli* e parte homologa a girase a de *E. coli*. FONTE: Marcotte *et al.* (1999).

Proteínas multifuncionais que não resultam da fusão de genes, de variantes de s*plice*, família de proteínas homologas ou com atividades enzimáticas promiscua (sem especificidade) são as proteínas conhecidas como proteínas *moonlighting*. Nessas proteínas a função de uma mesma cadeia polipeptídica pode ser alterada em decorrência de mudanças na localização celular, tipo celular, estado oligomerizado, concentração de um ligante, substrato, cofator ou produto (JEFFERY, 1999; JEFFERY, 2009) (FIGURA 4).

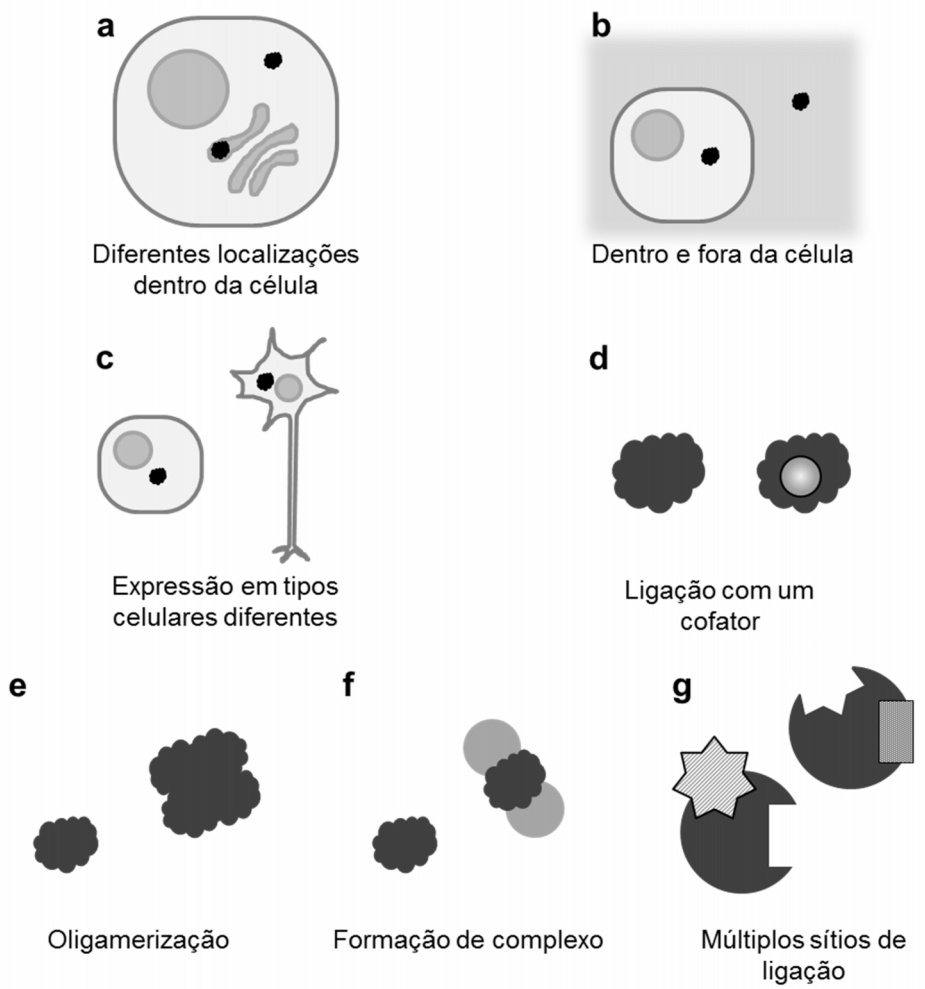


FIGURA 4 - EXEMPLOS DE MECANISMOS QUE ALTERNAM AS FUNÇÕES NAS PROTEÍNAS *MOONLIGHTING*. Uma proteína pode alternar entre diferentes funções: (a) em diferentes localizações celulares, (b) por estar no citoplasma ou estar secretada, (c) em diferentes tipos celulares, (d) por estar ou não ligada a um substrato, cofator ou produto, (e) estar em monômero ou em multímero, (f) por interação com outros polipeptídeos e (g) por apresentar diferentes sítios de ligação para diferentes substratos. FONTE: Adaptado de Jeffery (1999).

A determinação da função proteica de maior confiabilidade é a realizada com uso de dados experimentais (BLABY-HAAS; CRÉCY-LAGARD, 2011). São exemplos de experimentos que fornecem características: os microarranjos de DNA, RNA-seq, DNA-chip e camundongo *knockout*. Mas as estratégias de caracterização de produtos gênicos utilizando dados de técnicas experimentais de *high-thoughput* ainda não são capazes de preencher a lacuna entre uma sequência e sua função (BLABY-HAAS; CRÉCY-LAGARD, 2011). A função de 60% dos genes de 50% dos genomas ainda não é conhecida (genes anotados com termos como *hypothetical*, *putative*, *unknown*, *uncharacterized*, *predicted*, *no hits*, *codon recognized, expressed protein* e *conserved protein*).

A anotação das proteínas deriva de menos de 5% das proteínas que tem função comprovada experimentalmente (CHITALE; KIHARA, 2011; VALENCIA,

2005). Aproximadamente 40% das proteínas são marcadas como não apresentando função conhecida e são registradas como *Unknown function* (BLABY-HAAS; CRÉCY-LAGARD, 2011; FRIEDBERG; JAMBON; GODZIK, 2006).

4.3 PREDIÇÃO DE FUNÇÃO

A predição automática de função de proteínas, ganhou destaque com o ainda consagrado conceito de homologia baseado na similaridade entre sequências, com destaque para a ferramenta Blast do NCBI (CHITALE; KIHARA, 2011; MOUNT, 2004). Atualmente estudos vêm indicando que a utilização dessa metodologia pode não ser a melhor opção, mesmo em casos de sequências com alta similaridade (BLABY-HAAS; CRÉCY-LAGARD, 2011; CHITALE; KIHARA, 2011; FRIEDBERG, 2006). Experimentos que geram dados de expressão em larga escala propõem-se a preencher, futuramente, a lacuna entre produtos gênicos e sua função, em que aproximadamente 40% das proteínas anotadas como não apresentando função conhecida (BLABY-HAAS; CRÉCY-LAGARD, 2011). As abordagens computacionais de predição de função permitem a caracterização de sequências que não contam com dados laboratoriais específicos.

As ferramentas computacionais existentes podem ser embasadas em aprendizagem automática, interações proteicas, filogenia, comparação de estrutura e comparação de sequências. As abordagens adotadas podem ser por agrupamento, busca por padrões e por similaridade de sequência (Quadro 3) (HENRY et al., 2011; LOEWENSTEIN et al., 2009; RENTZSCH; ORENGO, 2009). Ferramentas de predição de função são foco de interesse de vários grupos de pesquisa e cerca de 100 sistemas estão registrados em diferentes revisões da área de predição (HENRY et al., 2011; JANGA; DÍAZ-MEJÍA; MORENO-HAGELSIEB, 2011; PANDEY; KUMAR; STEINBACH, 2006; RENTZSCH; ORENGO, 2009) (APÊNDICE 1). Elas podem estar associadas à curadoria manual ou automática, receber diferentes entradas para realizar a predição de função (arquivo com sequências de caracteres, arquivos contendo a informação estrutural ou um conjunto de interações, por exemplo), realizar as buscas por comparação com sequências completas ou com HMMs, focar na busca de domínios ou motivos, e podem estar

disponíveis para execução por servidor web, para execução local por linha de comando e/ou com interface gráfica (HENRY et al., 2011; RENTZSCH; ORENGO, 2009). Esses sistemas são foco de grupos, encontros e avaliações, como o Critical Assessment of Techniques for Protein Struture Predictions (CASP) (ANEXO 5), Systems in Molecular Biology (ISMB) e Automatic Function Prediction Special Interest Group (AFP-SIG) (CHITALE; KIHARA, 2011; FRIEDBERG; JAMBON; GODZIK, 2006) (Quadro 4).

MÉTO-DO	NOME	CITAÇÃO (1ª PUBLICAÇÃO)	ÚLTIMA ATUALIZA-ÇÃO	NÚMERO DE CITA-ÇÕES[1]	BASE DE DADOS	PLA-TA-FOR-MA
Similaridade	Gotcha[2]	Martin *et al.* (2004)	-	134	NR e GO	W
	PFP	Hawkins *et al.* (2006)	2009	70	NR e GO	W
	Goblet[2]	Groth et al (2004)	2010	82	NR e GO	W
	Gosling[2]	Jones *et al.* (2008)	-	6	NR e GO	
	BLAST2GO	Conesa *et al.* (2005)	2011	736	NR e GO	I,P
	OntoBlast[2]	Zehetner (2003)	-	83	NR e GO	
Filogenia	RIO[2]	Zmasek e Eddy (2002)	2006	146	Pfam, UniProtKB, “the tree of life” do NCBI	W
	FIGENIX*	Gouret (2005)	-	57	NR, “the tree of life” do NCBI	W
	AFAWE	Jocker (2008)	-	4	UniProt, SwissProt, RefSed, InterPro, GO	W
Busca por padrão	InterPro	Apweiler *et al.* (2001)	2011	861	Gene3D, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY e TIGRFAMs	P, W
	PROSITE[3]	Sigrist *et al.* (2002)	2012	510	Assinaturas (1308 padrões e 1034 perfis)	P, W
	Pfam[3]	Sonnhammer *et al.* (1997)	2011	730	Manualmente curada com 13000 famílias (Pfam-A)	P, W
	SUPERFAMILY[3]	Gough *et al.* (2001)	2010	660	própria a partir 900 genomas	W

Continua.

MÉTO-DO	NOME	CITAÇÃO (1ª PUBLICAÇÃO)	ÚLTIMA ATUALIZA-ÇÃO	NÚMERO DE CITA-ÇÕES[1]	BASE DE DADOS	PLA-TA-FOR-MA
Busca por padrão	SMART[3]	Schultz *et al.* (1998)	2011	2057	Manualmente curada com 1009 domínios proteicos	W
	PANTHER[3]	Thomas *et al.* (2003)	2012	680	Manualmente curada com 6594 famílias e 62972 subfamílias	P, W
	PIRSF[3]	Wu *et al.* (2004)	2006	122	PIR	W
	PRIAM	Claudel-Renard *et al.* (2003)	2011	102	ENZYME	P, W
	PropSearch	Hobohm e Sander (1995)		171	SWISS-PROT (vetores com as 144 características de cada sequência no ano de 1995)	W
Grupo de homólogos	ProtoNet	Sasson *et al.* (2003)	2011	65	UniProt (Swiss-Prot)	W
	eggNOG[3]	Jensen *et al.* (2008)	2011	80	721801 grupos ortólogos em 1133 espécies (4396591 proteínas)	W
	InParanoid[3]	Remm *et al.* (2001)	2009	728	1687023 sequências e 100 organismos (eucariotos e a Escherichia coli) construída com o programa inParanoid	P*,W
	OrthoMCL[3]	Li *et al.* (2003)	2011	640	1398546 sequências de proteínas de 150 genomas	P, W
Aprendizagem automática	ProtFun	Jensen *et al.* (2002)	2007	217	InterPro, UniProt (Swiss-Prot), GO	W
	SVM-Prot	Cai *et al.* (2003)	-	244	Pfam, BRENDA, GPCRDB, NuclearRDB, NCBI, TCDB, SWISS-PROT	W
	FFPred	Lobley *et al.* (2008)	-	12	uniref90,GO	W
	EzyPred	Shen e Chou (2007)	-	151	ENZYME	W
	EVEREST[2]	Portugaly *et al.* (2006)	2007	18	UniProtKB, PDB, SCOP, CATH, Pfam	P, W

Continua.

MÉTO-DO	NOME	CITAÇÃO (1ª PUBLICAÇÃO)	ÚLTIMA ATUALIZA-ÇÃO	NÚMERO DE CITA-ÇÕES[1]	BASE DE DADOS	PLA-TA-FOR-MA
Aprendizagem automática	EFICAz	Tian *et al.* (2004)	2009	49	UniProt (Swiss-Prot)	P, W
	CombFunc	Wass e Sternberg (2008)	2011	36	InterPro, Pfam, MINT, IntAct, COXPRESdb e 3DLigandSite, GO	W
Rede de interação funcional	STRING[3]	Snel *et al.* (2000)	2011	192	5214234 proteínas de 1133 organismos	W

[1] baseado em dados do Google do primeiro semestre de 2012

[2] apresentaram algum problema, como não funcionamento adequado do site ou do sistema de busca.

[3] sistemas com base de dados própria associada

* permite baixar os dados, mas não apresenta um sistema de busca próprio.

- dados não disponíveis

NR: base de dados não redundante do NCBI

GO: Gene Ontology

P: execução em linha de comando (*prompt*)

W: execução por *web server*

QUADRO 3 - EXEMPLOS DE PROGRAMAS QUE PERMITEM REALIZAR A PREDIÇÃO DE FUNÇÃO. FONTE: Pandey *et al.* (2006), Rentzsch e Orengo (2009), Henry *et al.* (2011) e Janga *et al.* (2011).

Nome	**Sigla**	**Foco**	**Descrição**
Automatic Function Prediction Special Interest Group	AFP-SIG	Vinculado ao CAFA	Encontro para promover discussões a respeito da predição de função de genes e produtos gênicos.
Critical Assessment of protein Function Annotation algorithms	CAFA	Preditores de função que utilizam termos do GO (*molecular functionebiologicalprocess*)	Método de avaliação em larga escala que utiliza o GO.
Critical Assessment of Techniques for Protein Struture Predictions	CASP	Métodos preditores de estrutura a partir da sequência	Avaliação de métodos de modelagem de estrutura proteica, com encontros para discussão dos resultados e divulgação dos resultados.
Intelligent Systems in Molecular Biology	ISMB	Métodos computacionais avançados aplicados a problemas biológicos	Conferência em bioinformática.

QUADRO 4 - GRUPOS, ENCONTROS E AVALIAÇÕES QUE ABORDAM A PREDIÇÃO DE FUNÇÃO DE PROTEÍNAS E SUAS METODOLOGIAS. FONTE: O autor (2013).

4.3.1 Recursos com busca por sequência

A maioria dos métodos para predizer a função é feita a partir de uma dada sequência não caracterizada, retorna uma lista ou uma única opção com as funções que podem estar vinculadas aquela sequência (BLABY-HAAS; CRÉCY-LAGARD, 2011; HENRY et al., 2011; JANGA; DÍAZ-MEJÍA; MORENO-HAGELSIEB, 2011; PANDEY; KUMAR; STEINBACH, 2006; RENTZSCH; ORENGO, 2009). Em métodos que retornam uma lista fica a encargo do pesquisador eleger a função que melhor representa a sequência com função desconhecida. Isso representa uma vantagem, por agregar o conhecimento do pesquisador, mas impedem a automatização do desse processo e depende de uma escolha subjetiva, que pode acarretar em problemas na nomenclatura final da sequência (TAMAMES; VALENCIA, 2006). O Blast e PFP são exemplos desse tipo de abordagem (ALTSCHUL *et al.*, 1997; HAWKINS; LUBAN; KIHARA, D, 2006).

Tanto programas que retornam uma resposta única para cada sequência informada, quanto os que retornam uma lista, comumente apresentam uma base de dados associada. A capacidade de um sistema atribuir um nome para uma sequência desconhecida está geralmente relacionada à base de dados a qual está vinculada (FONTANA *et al.*, 2009). Essas bases de dados associadas podem existir antes do desenvolvimento do sistema, como o Blast2GO e o InterPro (CONESA *et al.*, 2005; ZDOBNOV; APWEILER, 2001) que utilizam dados de bases já existentes, ou uma base desenvolvida em conjunto com o sistema, como o Pfam e o PANTHER (SONNHAMMER; EDDY; DURBIN, 1997; THOMAS; CAMPBELL; KEJARIWAL, 2003).

Dos programas que retornam uma resposta única para cada sequência informada, se destacam os programas listados no quadro:

NOME	BASE DE DADOS ASSOCIADA	TIPO DE BUSCA	ÚLTIMA ATUALIZAÇÃO	NÚMERO DE CITAÇÕES[1]
Blast2GO	NR, GO	alinhamento pelo Blast e pontuação pelo GO	2011	736
InterPro	ProDom, PROSITE, HAMAP, PRINTS, PANTHER, PIRSF, Pfam, SMART, TIGRFAMs, Gene3D, SUPERFAMILY	Cruza dados de busca de diferentes bancos de dados	2011	861
PANTHER	Própria manualmente curada	Comparação de HMM	2012	680
Pfam	Própria manualmente curada	Comparação de HMM	2011	730
PROSITE	Própria gerada automaticamente e parte com curadoria manual	Por perfis (matriz de pesos)	2012	510

[1] dados do Google do primeiro semestre de 2012

QUADRO 5 - SUMÁRIO DAS CARACTERÍSTICAS DOS PROGRAMAS QUE RETORNAM UMA RESPOSTA PARA CADA SEQUÊNCIA INFORMADA. FONTE: O autor (2013).

4.3.1.1 Blast2GO

Programa desenvolvido em Java que trabalha sobre os resultados do Blast contra a base de NR (base de dados não redundante do NCBI) ou qualquer subconjunto de sequências no formato aceito pelos programas do pacote Blast de busca por similaridade que contenham anotações do GO. A partir dos resultados do Blast da sequência que se deseja inferir a função, são selecionadas as sequências com nomes informativos e é então realizada uma pontuação por termos do GO mapeados no resultado do Blast, elencando o nome representativo a partir dessa pontuação. Quando a sequência representa uma enzima, também é fornecida a classificação segundo o EC. Apresenta ferramentas de visualização e análises estatísticas da informação funcional gerada (FIGURA 5) (CONESA *et al.*, 2005).

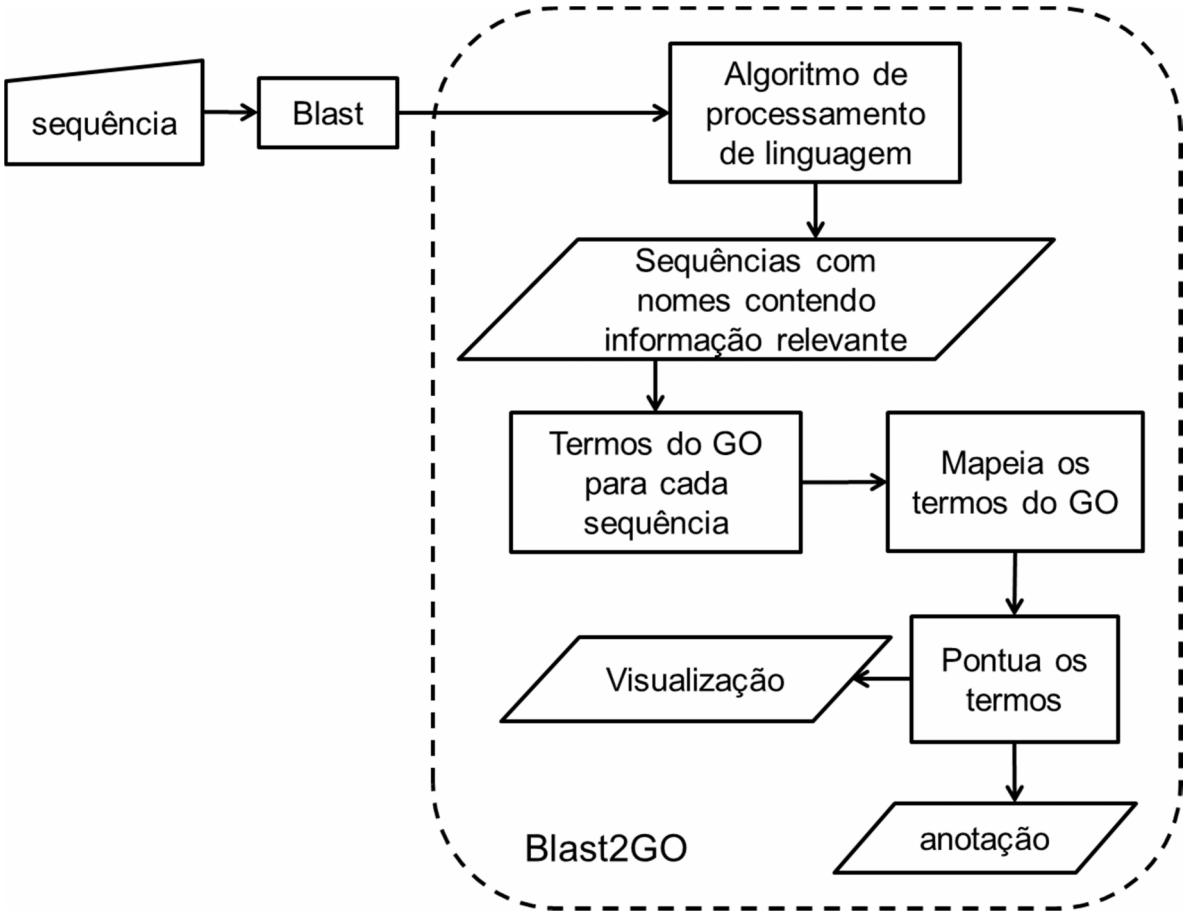


FIGURA 5 - FLUXOGRAMA SIMPLIFICADO DO PROGRAMA BLAST2GO. A partir do resultado do Blast para uma dada sequência, são mapeados e pontuados os termos do GO para cada sequência resposta, determinando através dessa pontuação a anotação mais indicada. FONTE: Adaptado de Conesa *et al.* (2005).

O Blast é utilizado para encontrar todos os homólogos das sequências com função desconhecida. Os valores de *threshold* e de *e-value* utilizados no Blast devem permitir o retorno de sequências significativas e sem alinhamentos em trechos curtos. O Blast2GO busca considerar apenas anotações que apresentam informações relevantes, desconsiderando anotações como *hypothetical protein* ou *expressed protein*, através de um algoritmo de processamento de linguagem (CONESA; GÖTZ, 2008; CONESA et al., 2005). A partir dos identificadores dos genes (*gene identifiers - gi*) e dos números do depósito dos genes (*gene accessions*) dessas sequências homólogas, o Blast2GO retoma as anotações do GO e os *evidence codes* (EC) de cada anotação. As anotações do GO são todos os termos do GO utilizados para caracterizar uma determinada sequência e o EC representa um índice de confiabilidade de cada termo em descrever a sequência (CONESA *et al.*, 2005).

O mapeamento é feito utilizando as anotações do GO e os ECs, gerando assim um conjunto de possíveis anotações para a sequência desconhecida. Nesse conjunto de anotações possíveis é aplicada a regra de anotação (*annotation rule* - AR), que busca a anotação mais específica com certo grau de confiança que pode ser fornecido pelo usuário. A anotação selecionada é aquela que apresenta a menor pontuação de anotação. A pontuação de anotação é composta pela maior similaridade somada que considera o grau de confiança para realizar ou não uma abstração que consiste em um valor proporcional ao nó parental dos nós filhos que estão presentes no conjunto de termos do GO da anotação candidata. Após a anotação o programa possibilita ao usuário uma visualização dos dados que levaram a anotação proposta, visando que o usuário avalie a anotação resultante (CONESA *et al.*, 2005).

O programa apresenta uma versão com interface gráfica e uma para execução em linha de comando. Ambas permitem as mesmas saídas de dados, porém a versão com interface gráfica é mais interativa e permite a visualização dos dados de saída. Ambas as versões foram desenvolvidas em Java, permitindo a independência de plataforma.

4.3.1.2 InterPro

InterProScan foi desenvolvido em Perl e trata-se de uma ferramenta de busca relacionada a base de dados InterPro. Para cada sequência submetida, o programa InterProScan realiza uma busca na base de dados InterPro, que resulta da mescla manual das saídas dos diferentes sistemas integrados, retornando as assinaturas InterPro e das diferentes bases de dados integradas (FIGURA 6). O programa submete a sequência a cada um dos sistemas integrados e a partir das saídas de cada programa é procurada na base InterPro a assinatura correspondente (APWEILER, *et al.*, 2001; HUNTER *et al.*, 2009; ZDOBNOV; APWEILER, 2001).

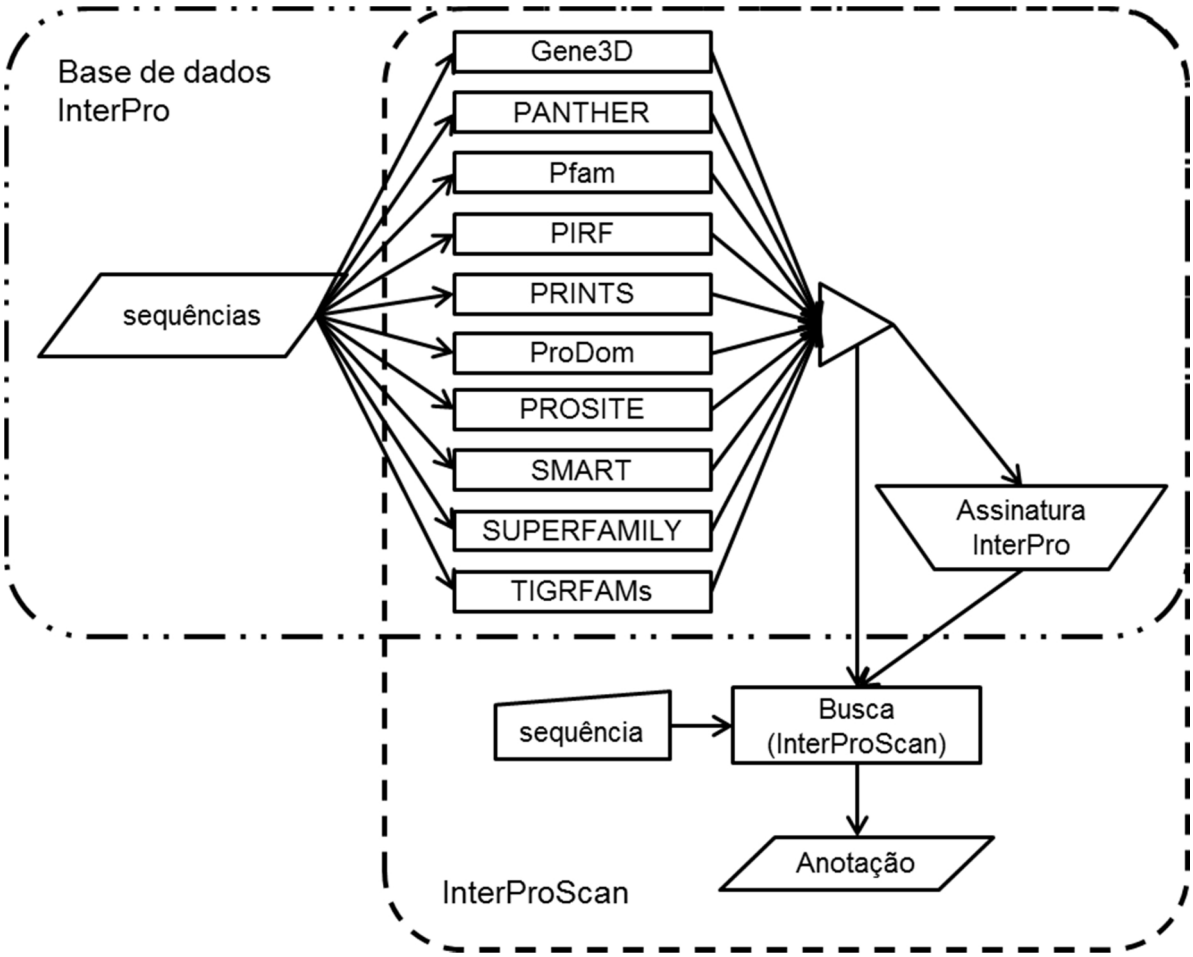


FIGURA 6 - FLUXOGRAMA SIMPLIFICADO DO SISTEMA INTERPRO. Interligando diversos sistemas o interpro apresenta uma representação das diferentes assinaturas. Em uma busca a partir de uma sequência podem ser listadas as informações presentes nos diversos sistemas a respeito daquela sequência e a assinatura do próprio InterPro. FONTE: O autor (2013).

4.3.1.3 PANTHER

O Panther Score realiza a classificação de sequências com relações a comparação com cadeias de Markov ocultas (HMMs - *Hidden Markov Models*) (THOMAS, 2011). Estas HMMs são formadas a partir da curadoria manual de sequências agrupadas em famílias e subfamílias e associadas a uma anotação de ontologia do GO por especialistas. Utiliza os programas MAFFT e SAM para gerar as HMMs, o algoritmo GIGA (*Gene tree Inference in the Genomic Age*) para construção das árvores filogenéticas (FIGURA 7) (MI *et al.*, 2010; THOMAS; CAMPBELL; KEJARIWAL, 2003).

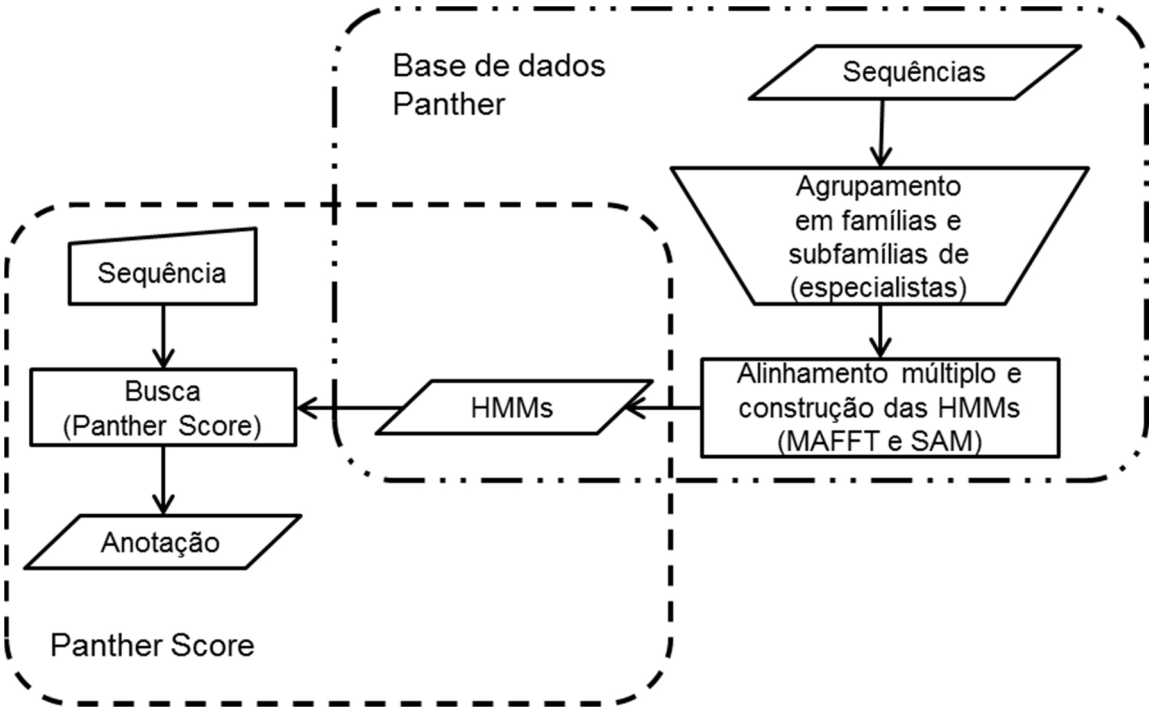


FIGURA 7 - FLUXOGRAMA SIMPLIFICADO DO SISTEMA PANTHER. As HMMs contêm as informações das famílias e subfamílias agrupadas por especialistas. Essas HMMs são utilizadas com base de busca para definir uma pista sobre a função de uma sequência de interesse. FONTE: O autor (2013).

4.3.1.4 Pfam

Pfam scan realiza a classificação de sequências com relações a comparação com cadeias de Markov ocultas (HMMs - Hidden Markov Models) utilizando o programa pfam_scan.pl (GENOME RESEARCH LTD, 2010). Estas HMMs são formadas por alinhamentos de alta qualidade curados manualmente no conjunto Pfam-A e automaticamente no conjunto Pfam-B (utilizando o algoritmo ADDA) (PUNTA *et al.*, 2012) (FIGURA 8).

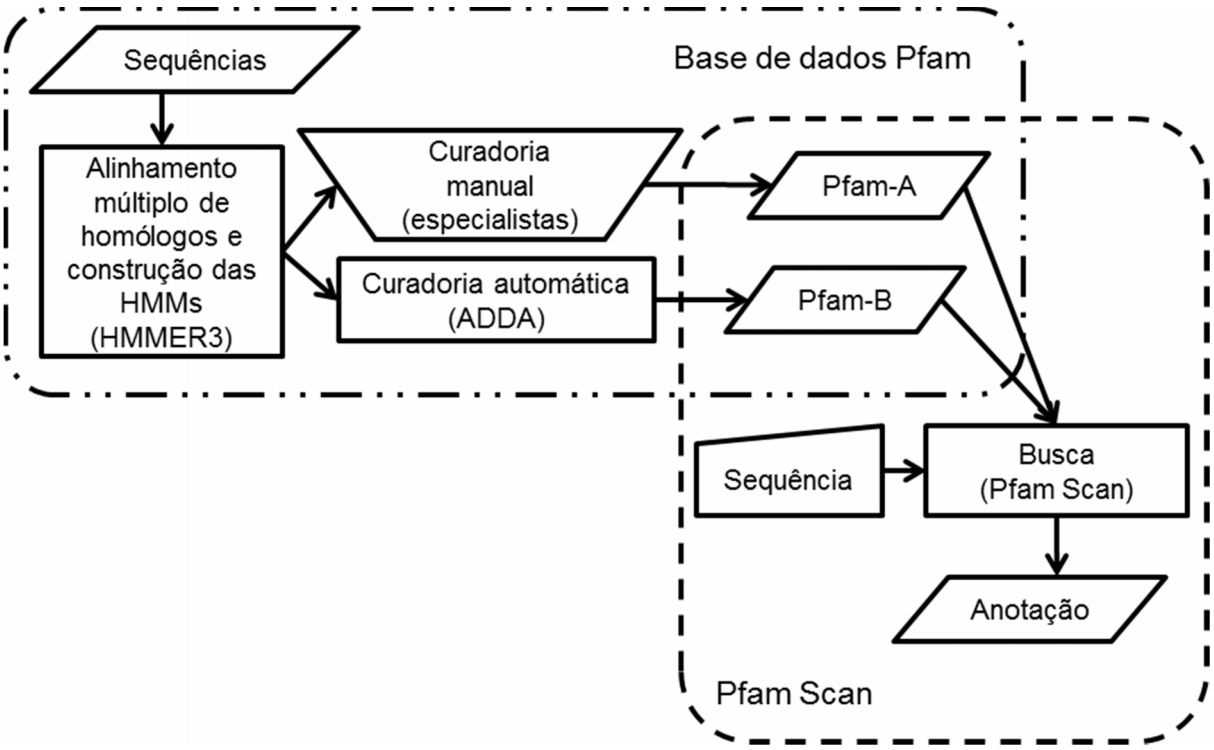


FIGURA 8 - FLUXOGRAMA SIMPLIFICADO DO SISTEMA PFAM. As bibliotecas de HMMs do Pfam são organizadas por curadoria manual e automática. Ambas as bibliotecas ou apenas uma delas podem ser utilizadas com base de busca para definir uma pista sobre a função de uma sequência de interesse. FONTE: O autor (2013).

Utiliza o programa HMMER3 para detectar as homologias, realizando a construção das HMMs, o algoritmo jackhmmer para gerar novas famílias em atualizações, e a base de dados de referência utilizada é a UniProtKB, mas também possui ligação com as bases de dados NR, PROSITE, SCOP e CAZy (PUNTA *et al.*, 2012).

Para a construção do conjunto de dados Pfam-A, as sequências são alinhadas e HMM é construída a partir desse alinhamento. Esse alinhamento passa por uma etapa manual em que é realizado o controle de qualidade, a anotação e a conexão com outras bases de dados (PUNTA et al., 2012; SONNHAMMER, E. L.; EDDY, S R; DURBIN, 1997). Assim, as famílias do Pfam-A são definidas manualmente por especialistas.

4.3.1.5 PROSITE

Caracteriza uma sequência por comparação com perfis e padrões utilizando o ps_scan.pl (execução local) ou o ScanProsite (execução por *web server*) (CASTRO *et al.*, 2006; GATTIKER; GASTEIGER; BAIROCH, 2002). Os profiles e patterns são gerados a partir do alinhamento múltiplo de sequências homólogas. Patterns correspondem a expressões consenso oriundas do alinhamento. Profiles são construídos com os programas pfw e pfmake do pacote PFTOOLS.

O programa pfw pontua os alinhamentos de forma a evitar a sobre representação e o pfmake contribui para montar os profiles com os dados dos pesos sem considerar a escala (normalizando a pontuação) utilizando a matriz BLOSUM45 para pontuar as substituições. Os profiles são então submetidos a uma edição manual por especialistas com bom conhecimento de generalizações da sintase do profile. O PROSITE adota o DAS (Distributed Annotation System) como padronização (FIGURA 9).

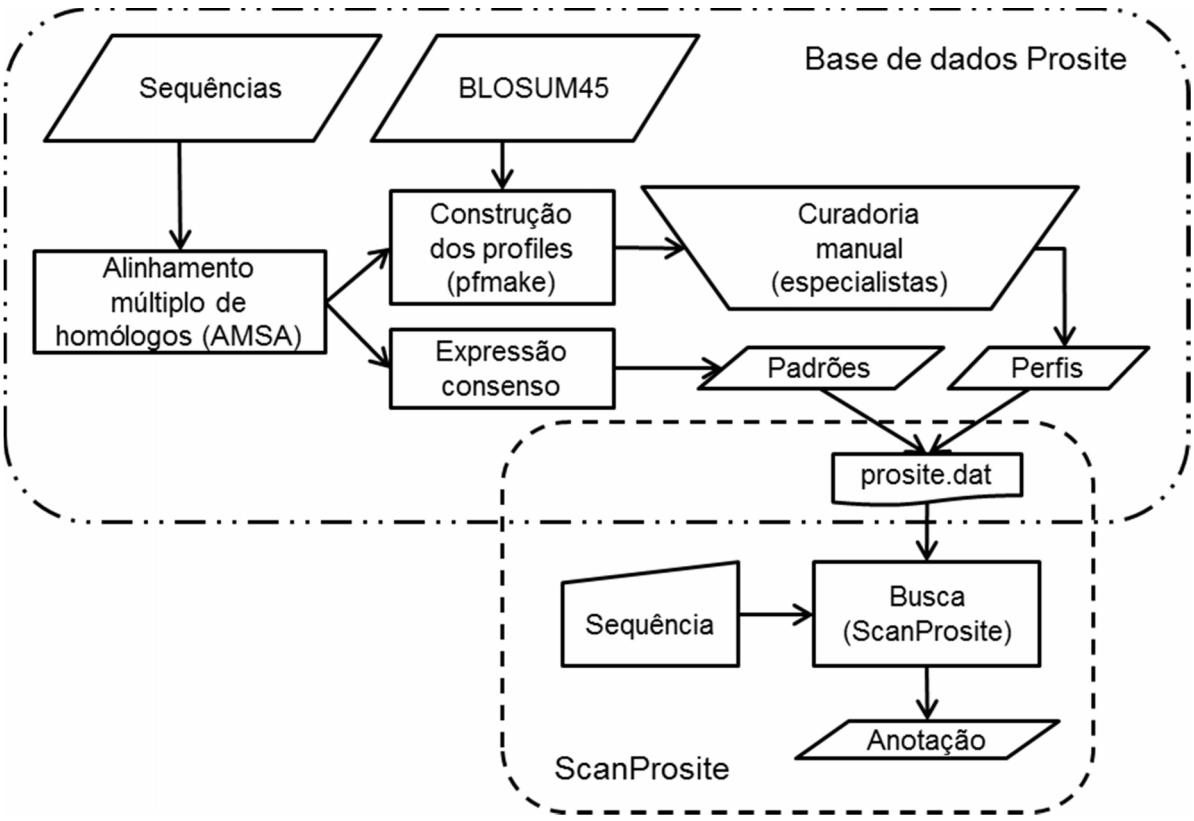


FIGURA 9 - FLUXOGRAMA SIMPLIFICADO DO SISTEMA PROSITE. Através do alinhamento múltiplo de sequências são construídos perfis (por curadoria manual) e padrões que compõe o arquivo prosite.dat, utilizado pelo sistema de busca para retornar uma indicação funcional para uma sequência de interesse. FONTE: O autor (2013).

4.3.2 Bases de dados

As bases de dados vinculadas as sequências, tanto de proteínas quanto de nucleotídeos, são repositórios das informações que podem ser diretos de dados experimentais ou curados manualmente ou automaticamente (APÊNDICE 2). Atualmente com o advento da massiva geração de dados por sequenciamento em larga escala, é observado um crescente aumento na quantidade de registros de bases de dados, de depósitos nessas diferentes bases e de erros de anotação (Gráfico 1) (BLABY-HAAS; CRÉCY-LAGARD, 2011; GALPERIN; FERNÁNDEZ-SUÁREZ, 2012).

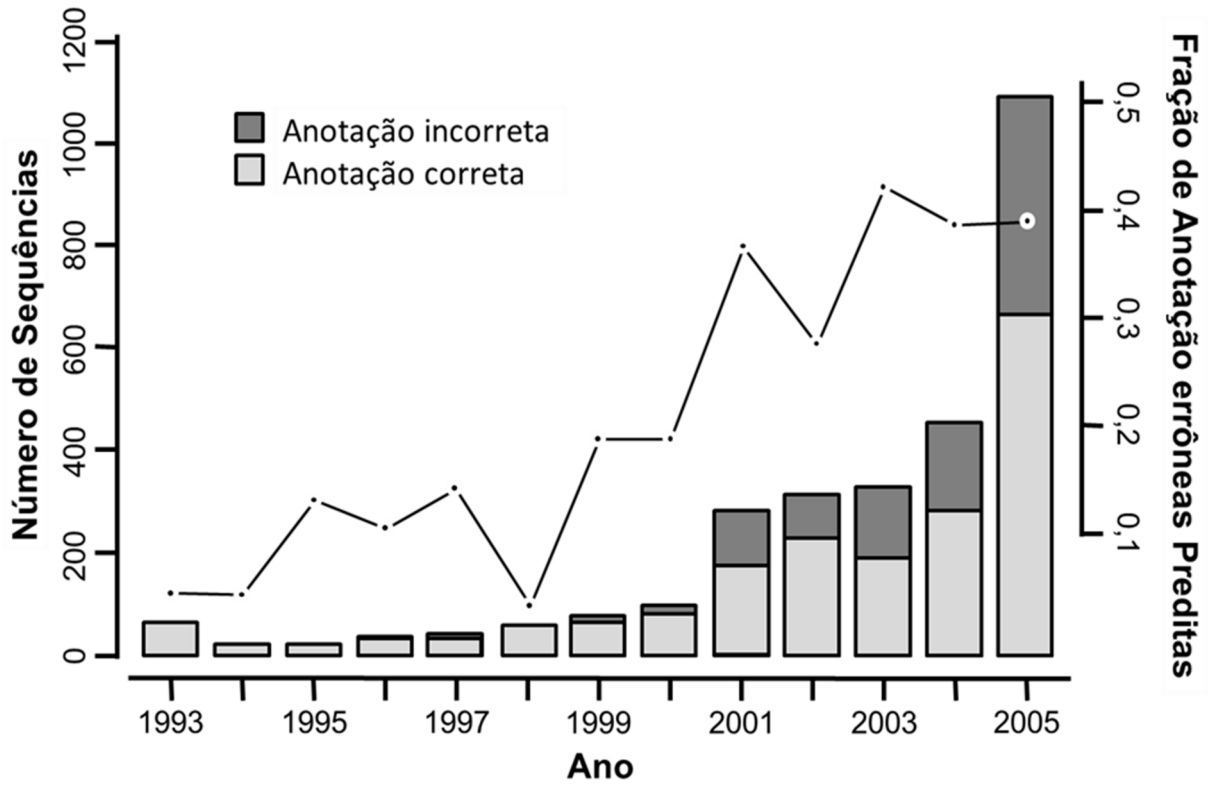


GRÁFICO 1 - ERRO DE ANOTAÇÃO NA BASE DE DADOS NR PARA 37 FAMÍLIAS NOS ANOS DE 1993 A 2005. FONTE: SCHNOES *et al.* (2009).

Em 2011, Blaby-Haas e Crécy-Lagard relataram existir por volta de 2200 genomas e 6 milhões de proteínas sequenciadas, e no NAR (Nucleic Acids Research) há o registro de mais de 1512 bases de dados de biologia molecular (Gráfico 2), sendo que ainda há a demanda para criação de mais bases de dados (APWEILER *et al.*, 2010; BLABY-HAAS; CRÉCY-LAGARD, 2011; GALPERIN;

FERNÁNDEZ-SUÁREZ, 2012). E dentro desse grande volume de bases de dados, existem diferentes bases relacionadas à armazenagem de dados referentes a diferentes características de proteínas (Quadro 6).

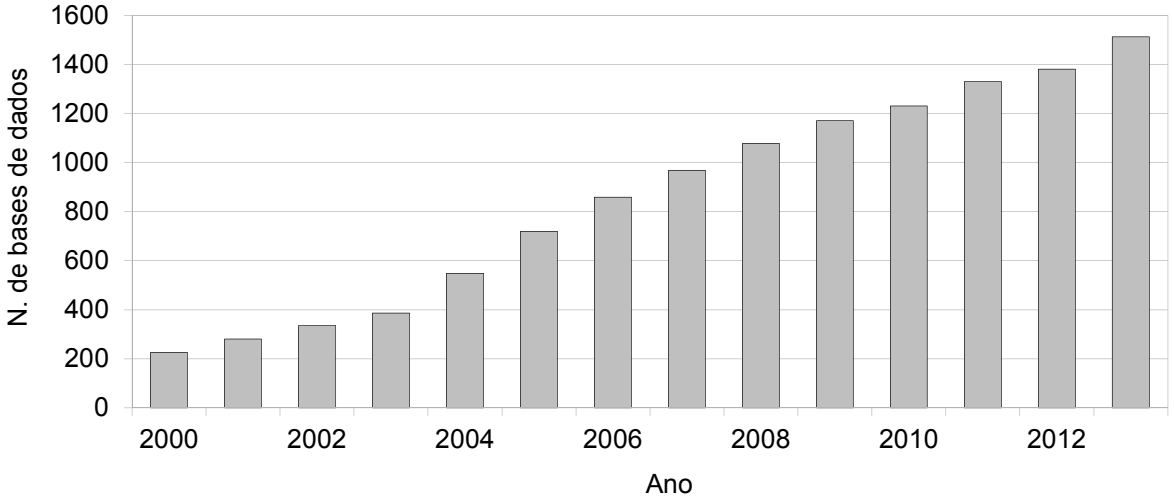


GRÁFICO 2 – CONTAGEM CUMULATIVA DE BASES DE DADOS REGISTRADAS NO NAR. FONTE: Adaptado de NAR (2013).

NOME	SIGLA	TIPO DE DADOS
Database of Interacting Proteins	DIP	Interação entre proteínas
Kyoto Encyclopedia of Genes and Genomes	KEGG	Vias metabólicas e as proteínas envolvidas
Non-Redundant	NR	Sequência de aminoácidos de proteínas
Protein Data Bank	PDB	Estrutura 3D dos aminoácidos de proteínas
PROCAT	-	Estrutura 3D do sítio ativo de proteínas
PRODOM	-	Sequência de aminoácidos de domínios de proteínas
Structural Classification Of Proteins	SCOP	Estrutura 3D de domínios de proteínas
SUrface Residues and Functions Annotated, Compared and Evaluated	SUFACE	Estrutura 3D da superfície de proteínas

QUADRO 6 - EXEMPLOS DE BASES DE DADOS PARA CARACTERÍSTICAS PROTEICAS. FONTE: Pandey *et al.* (2006).

O NAR apresenta o registro de diferentes bases de dados de biologia molecular e ele disponibiliza artigos, sumários, descrições e links dessas bases de dados. O NAR agrupa as bases de dados em categorias e subcategorias a partir dos tipos de dados que armazenam, sendo que a bases de dados que pertencem a mais que uma subcategoria.

Bases de dados de sequências genômicas são as mais numerosas, seguidas das bases de dados de sequências proteicas e de estrutura (Gráfico 3). As

bases de dados que armazenam dados de estrutura de proteínas são a subcategoria com maior número de bases de dados (APÊNDICE 3).

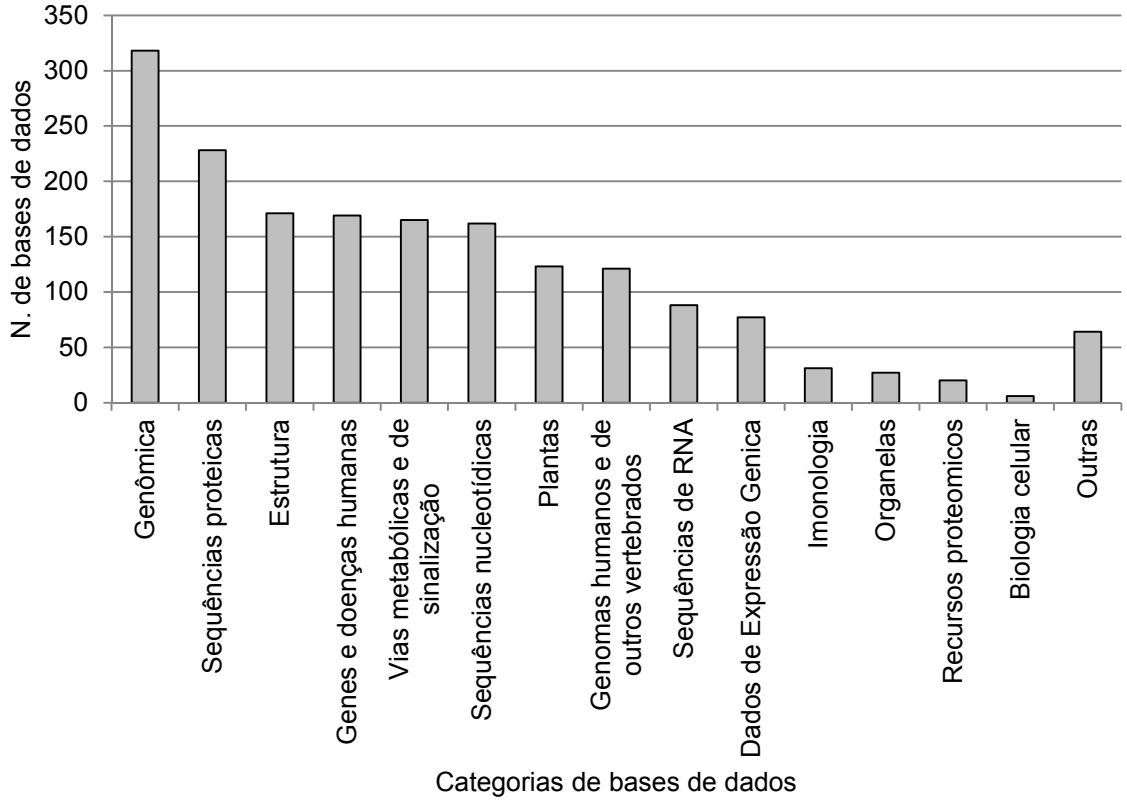


GRÁFICO 3 – NÚMERO DE BASES DE DADOS REGISTRADAS PELO NAR EM CADA CATEGORIA. AS BASES DE DADOS PODEM PERTENCER A MAIS QUE UMA CATEGORIA. FONTE: Adaptado de NAR (2013).

A base de dados UniProt é uma das bases de dados registradas pelo NAR com uma grande número de sequências proteicas e que mantém referência cruzada com outras bases de dados. Das mais de 34 milhões de sequências automaticamente anotadas (UniProtKB/TrEMBL) depositadas nessa base de dados (ANEXO 6), a maioria da referência cruzada é feita com bases de dados de famílias e domínios proteicos (Gráfico 4).

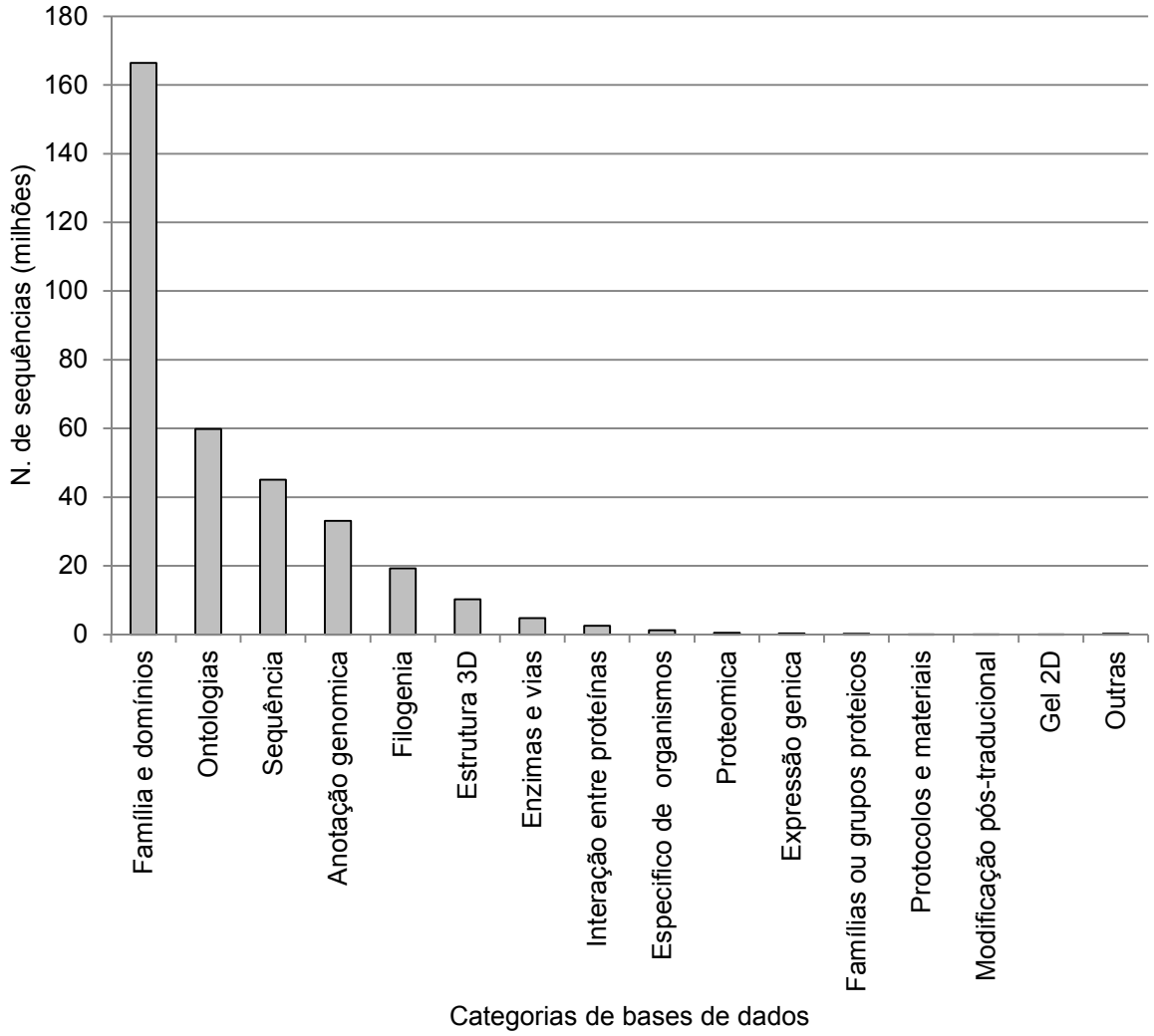


GRÁFICO 4 – NÚMERO DE SEQUÊNCIAS DEPOSITADAS NA BASE DE DADOS UNIPROTKB/TREMBL EM RELAÇÃO AS CATEGORIAS DE BASES DE DADOS RELACIONADAS (REFERÊNCIA CRUSADA). FONTE: Adaptado de UNIPROT (2013).

Essas sequências depositadas com anotação automática são 75,58% marcadas como preditas e 22,14% marcadas como inferida por homologia. O termo *similarity* é o segundo comentário mais comum nos registros de sequências, perdendo apenas para o termo *caution*.

A parte manualmente anotada da base de dados UniProt, a UniProtKB/Swiss-Prot foi a base de dados que apresentou menor porcentagem de erros de anotação em comparação a outras três bases de dados no trabalho de Schnoes *et al.* (2009). Enquanto essas três bases de dados apresentaram nível de erros de anotação entre 5% e 63%, a base de dados Swiss-Prot apresentou níveis de anotação próximos à zero (ANEXO 7).

4.4 METODOS ESTATÍSCOS

Modelos de regressão e analises de variância (ANOVA) são ferramentas fundamentais em estatística (CHRISTENSEN, 1997). Na regressão e na ANOVA um valor esperado é representado como uma combinação linear de valores de variáveis preditivas conhecidas (CHRISTENSEN, 1997). É possível realizar inferências, avaliar os fatores que podem afetar a variável resposta. Quando se tem uma resposta dependendo de apenas um fator, a regressão é uma regressão simples, já quando há mais que um fator a regressão é denominada regressão múltipla (Figura 10).

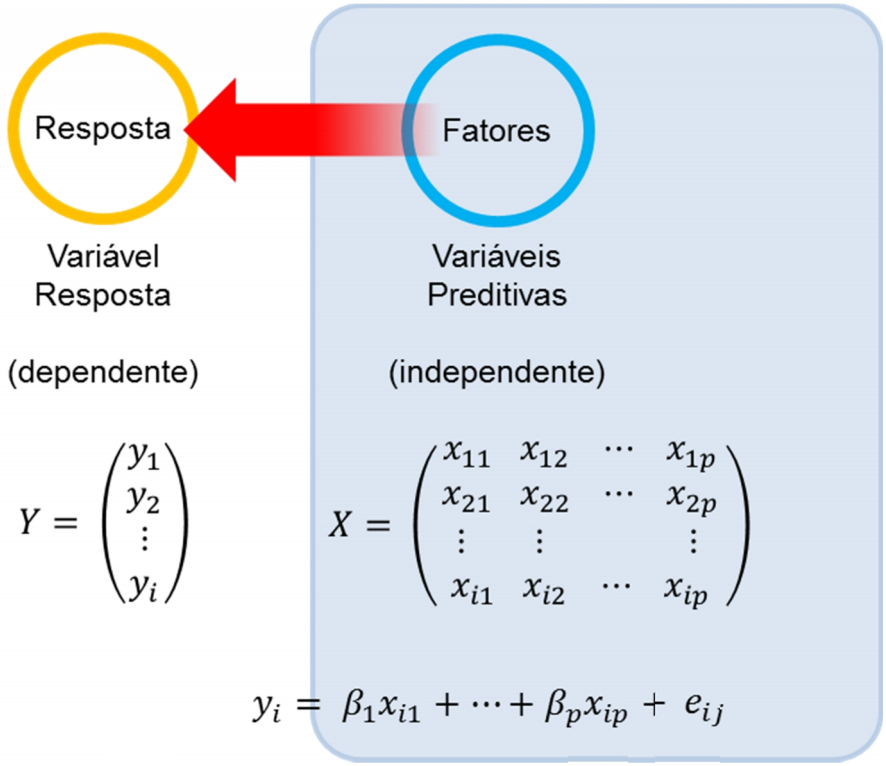


FIGURA 10 – MODELO DE REGRESSÃO MÚLTIPLA. Nas fórmulas: y é uma variável resposta e Y um vetor de respostas, x são variáveis preditivas (covariáveis, parâmetros) e X é uma matriz de covariávies, coeficientes β são os parâmetros de efeitos fixos, i é número da unidade experimental e $i = 1, \ldots, n$; p é o número de parâmetros e $p = 1, \ldots, m$; onde $E(e_i) = 0$; $x_{i1} = 1$ com β_1 sendo o intercepto. FONTE: Adaptado de Christensen (1997) e Turkman e Silva (2000).

Na regressão se deseja representar a relação causa e efeito entre variáveis, expressando matematicamente essa relação e avaliando uma possível dependência entre a variável resposta e as preditivas (CHRISTENSEN, 1997). Um exemplo de regressão é a regressão linear simples representa a relação entre duas variáveis, uma dependente (varável resposta) e outra independente (variável preditiva) (CHRISTENSEN, 1997).

O modelo linear normal que se aplica às variáveis de distribuição normal. Porém nos casos que o modelo normal não se aplica é necessária a aplicação de outros modelos

(CHRISTENSEN, 1997; RESENDE; BIELE, 2002; TURKMAN; SILVA, 2000). Modelos lineares generalizados compilam vários modelos de distribuição não normal e outros modelos, que mantêm em comum a possibilidade de expressar a variável resposta na forma de uma distribuição exponencial (Quadro 7).

MODELOS INCLUSOS NO MODELO LINEAR GENERALIZADO
• Modelo complementar log-log para ensaios de diluição • Modelo de análise de variância e covariância • Modelo de regressão de Poisson • Modelo de regressão linear clássico • Modelo de regressão logística • Modelo *probit* e *logit* para estudos de proporções • Modelos de regressão para análise de sobrevivência • Modelos log-lineares para dados de contagem • Modelos log-lineares para tabelas de contingência multidimensionais

QUADRO 7 – CASOS PARTICULARES DOS MODELOS LINEARES GENERALIZADOS. FONTE: Turkman e Silva (2000).

Desde sua introdução por Nelder e Wedderburn em 1972, com aprimoramentos dados por McCullagh e Nelder em 1989, a aplicação do modelo linear generalizado só ganhou maior popularidade 20 anos depois, quando o primeiro software (GLIM) era dirigido à aplicação desse modelo passou a ter uma utilização mais simples (CHRISTENSEN, 1997; TURKMAN; SILVA, 2000). Atualmente, muitos pacotes estatísticos apresentam módulos para o estudo desses modelos (TURKMAN; SILVA, 2000).

Para a aplicação do modelo linear generalizado Turkman e Silva (2000) listam três etapas essenciais:

1 – Formulação do modelo. Na formulação do modelo deve se considerar:

a) Escolha da distribuição da variável resposta;
b) Escolha das covariáveis e formulação da matriz de especificação (valores das covariáveis associados às respectivas respostas);
c) Escolha da função de ligação.

2 – Ajuste dos modelos. É a etapa em que os parâmetros dos modelos são estimados:

a) Coeficientes β's associados às covariáveis;
b) Parâmetro ϕ de dispersão (caso se aplique).

3 – Seleção e validação do modelo. Nessa etapa deve-se buscar um equilíbrio entre adequabilidade, parcimônia e interpretação. Como por exemplo, ao se analisar um modelo deve observar:

a) Um número moderado de parâmetros e que se adeque aos dados;
b) A consonância entre os valores preditos e os dados observados;
c) Existência de *outliers*;
d) Etc.

5 MATERIAIS

5.1 DADOS

Foram utilizados 12 conjuntos de dados de sequências proteicas no formato FASTA e contendo uma classificação funcional considerada correta para cada sequência (TABELA 1) (APÊNDICE 4). O nome dos conjuntos que correspondem a famílias e superfamília foram mantidos os mesmos utilizados na fonte do conjunto. Os conjuntos Aminergic GPCR (358 sequências), NHR (412 sequências) e Secretin-like (153 sequências), que representam três famílias, foram extraídos do material suplementar de Brown e colaboradores (2007). O material estava disponível até novembro de 2011, e os arquivos utilizados foram os contendo as sequências e com a classificação de cada sequências por especialistas.

TABELA 1 – LISTA DOS 12 CONJUNTOS DE DADOS UTILIZADOS NO PRESENTE TRABALHO. Esses conjuntos apresentam uma classificação da função considerada como correta para cada sequência. O nome de cada conjunto é o mesmo utilizado na fonte do dado. O número de sequências e o número de classificações funcionais foram retirados dos respectivos arquivos dos conjuntos.

CONJUNTOS TESTES	FONTE	NÚMERO DE SEQUÊNCIAS	NÚMERO DE CLASSIFICAÇÕES	NÚMERO DE BASES
AminergicGPCR*	Brown *et al.* (2007)	358	31	104335
NHR*	Brown *et al.* (2007)	412	27	72227
Secretin-like*	Brown *et al.* (2007)	153	15	38025
Enzimas	Dobson e Doig (2003)	690	630	226335
Não Enzimas	Dobson e Doig (2003)	487	449	88397
Bifuncional	adaptado de Dobson e Doig (2003)	60	51	22074
Enolase**	SFLD	927	25	357425
Crotonase**	SFLD	262	18	86773
Haloacid dehalogenase**	SFLD	389	22	263851
Vicinal oxygen chelate**	SFLD	145	12	39592
Radical SAM**	SFLD	145	19	52027
Padrão-ouro	Brown *et al.* (2006)	863	90	353221

* famílias
** superfamílias do Structure-Function Linkage Database (SFLD)
FONTE: O autor (2013).

Os conjuntos de sequências Enzimas (690 sequências de proteínas enzimáticas) e Não Enzimas (487 sequências de proteínas não enzimáticas) foram

extraídos do material suplementar de Dobson e Doig (2003), que consistem em proteínas com estrutura determinada experimentalmente por cristalografia de raio-X apresentando resolução menor ou igual a 2,5 Å e fator R menor ou igual a 0,25 (indicam estruturas tridimensionais com maior definições) do PDB com anotação funcional bem definida no ano de desenvolvimento do trabalho. Dobson e Doig (2003) disponibilizaram os PDB Ids (utilizado pela base de dados PDB para identificar cada sequência depositada) das sequências[1], utilizados para obter as sequências em fasta e as tabelas com dados personalizados como o identificador da sequência, nome da molécula.

O conjunto Bifuncional (60 sequências) foi selecionado do PDB seguindo as condições descritas para a seleção dos conjuntos de dados utilizados por Dobson e Doig (2003), ou seja, foram selecionadas todas as sequências proteicas presentes no PDB que apresentavam estrutura tridimensional com resolução menor ou igual a 2,5 Å e fator R menor ou igual a 0,25. Foi realizada também uma busca por título da estrutura contendo a palavra “bifunctional” e não contendo as palavras “putative”, “uncharacterized” e “unknown” para selecionar sequências bem caracterizadas.

Os conjuntos Enolase (927 sequências), crotonase (262 sequências), haloacid dehalogenase (389 sequências), vicinal oxygen chelate (145 sequências) e radical SAM (145 sequências) correspondem a superfamílias. Esses conjuntos foram retirados da base de dados *Structure-Function Linkage database* (SFLD)[2] em setembro de 2012, selecionando apenas as sequências com função conhecida e com ao menos um PDB ID vinculado. Está base de dados foi utilizada e recomendada no trabalho de Schnoes e colaboradores (2009).

O conjunto Padrão-ouro (863 sequências) foi descrito no trabalho de Brown e colaboradores (2006) como sendo um conjunto teste adequado para ferramentas de agrupamento e de classificação funcional de sequências. As sequências foram disponibilizadas como dados adicionais do artigo[3], e apresentam cada sequência com o número do GI (*GenInfo Identifier* do NCBI) correspondente.

---

[1] http://www.sciencedirect.com/science/MiamiMultiMediaURL/1-s2.0-S0022283603006284/1-s2.0-S0022283603006284-mmc01.doc/272582/html/S0022283603006284/999b7774b94f4dc086bb995ad4333cbb/mmc01.doc

[2] http://sfld.rbvi.ucsf.edu/django/

[3] http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1431709/bin/gb-2006-7-1-r8-S3.fa

5.2 HARDWARE E SOFTWARE

O computador utilizado para instalação dos programas e para a execução das análises contava com 2 GB de RAM, processador Intel Core i5 3.20 GHz, HD de 500GB. O sistema operacional utilizado foi o Ubuntu 10.04 com os pacotes g++, autoconf, csh, mysql-server, default-jre, blast2, libxml-quote-perl, R, HMMER2 e a instalação local do HMMER3.

Os programas Blast2GO, InterProScan, Panther Score, Pfam scan e ScanProsite foram instalados, assim como os recursos necessários para o funcionamento de cada programa (Quadro 8). Os critérios utilizados para a seleção dos sistemas foram: 1- compor o conjunto de sistemas relatados em revisões literárias sobre predição de função de proteína, 2- possuir número de citações superior a 500, 3- ser sistema com a última atualização em menos de 3 anos e 4- permitir análises automatizadas sem interação humana.

RECURSO		VERSÃO	DATA
Fonte de dados	arquivo de dados do ipscan	36.0	-
	arquivo de dados do Pfam	-	fev 2012
	arquivo de dados do PROSITE	20.83	13 jun 2012
	biblioteca de HMM do PANTHER	7.0	04 jan 2010
	GO	-	fev 2012
	NR	-	04 jul 2012
Programas	blastp	2.2.25+	-
	b2g4pipe	2.5.0	06 out 2011
	iprscan	4.8	-
	pantherScore.pl	1.03	04 out 2011
	pfam_scan.pl	-	18 jun 2010
	ps_scan.pl	1.79	-

QUADRO 8- RECURSOS RELACIONADOS À EXECUÇÃO DOS PROGRAMAS. FONTE: O autor (2013).

O programa Blast2GO dependia da instalação do Java para sua execução e dependia da saída em xml gerada pelo programa blastp. Os parâmetros utilizados para executar os programas estão presentes no quadro a seguir (Quadro 9). Para a execução do blastp os parâmetros utilizados foram o padrão presente no programa com interface gráfica. Os parâmetros para definição de *e-value* e *threshold* são restritivos, ou seja, retornam sequências com maior similaridade.

PROGRAMA	PARAMETRO	ARGUMENTO	DESCRIÇÃO
blastp	-query	caminho/arq.fasta	Arquivo de entrada .fasta
	-db	caminho/nr	Caminho para a base de dados no formato adequado do BLAST sem extensão
	-out	caminho/arq.xml	Arquivo de saída
	-outfmt	5	Opção de formato de saída (5= XML)
	-evalue	0,001	Limiar de *e-value*
	-num_alignments	20	Número de sequências retornadas por alinhamento
	-threshold	33	Mínimo *word score*
	-num_threads	4	Número de *threads* (CPUs) utilizada na execução
java	-Xms1024m		Tamanho (em bytes) alocado inicialmente na memória
	-Xmx2048m		Tamanho máximo (em bytes) alocado na memória
	-cp	/caminho/b2g4pipe/*: /caminho/b2g4pipe/ext/*: es.blast2go.prog. B2GAnnotPipe	Lista de arquivos ou diretórios contendo arquivos class, com o caminho para os class separados por ":"
b2g4pipe	-in		Arquivo XML com os resultados do blast no format NCBI
	-out		Arquivo de saída
	-prop		Arquivo com as configurações da aplicação (opcional: padrão=b2gPipe.properties)
	-annot		Gera um arquivo de anotação (.annot)
pfam_scan.pl	-fasta	caminho/arq.fasta	Arquivo de entrada .fasta
	-dir		Diretório com os arquivos Pfam
iprscan	-cli		Para utilizar em linha de comando
	-iprlookup		Aciona a busca pela anotação InterPro correspondente
	-goterms		Aciona a busca pela anotação InterPro GO correspondente (requer -iprlookup)
	-i	caminho/arq.fasta	Arquivo de entrada
	-o	saida.xml	Arquivo de saída (padrão: stdout)
	-format	xml	Formato de saída [raw, xml, txt, html] (default xml).
ps_scan.pl	-s	.fasta	Evita *patterns* e *profiles* comuns para maioria das sequências
	-d	prosite.dat	Caminho para o arquivo prosite.dat

Continua.

PROGRAMA	PARAMETRO	ARGUMENTO	DESCRIÇÃO
pantherScore.pl	-l	PANTHER7.0	Biblioteca PANTER com HMMs
	-D	B	Tipo de resultado: (B) maior pontuação
	-i	caminho/arq.fasta	Arquivo de entrada
	-o	caminho/arq.txt	Arquivo de saída (padrão: stdout)
	-n		Mostra os nomes das famílias e subfamilies
	-H	hmmsearch	Caminho para o binário do Hmmsearch

QUADRO 9 - PARÂMETROS UTILIZADOS NA EXECUTAÇÃO DOS PROGRAMAS. FONTE: O autor (2013).

6 MÉTODOS

6.1 ANÁLISE DOS PROGRAMAS DE PREDIÇÃO DE FUNÇÃO PROTEICA

Foram analisados cinco programas de predição de função de proteína aplicados aos conjuntos de dados elencados para os testes. Os dados resultantes das análises constituíram uma tabela com três colunas: 1- conjunto teste ao qual os dados pertencem, 2- programa, 3- classificação (correta, incorreta), e cada linha da tabela se referindo a uma sequência proteica.

A coluna de interesse é a de classificação, que contém a informação se a predição foi correta ou incorreta. Para permitir a análise dessa informação sem recorrer a métodos não paramétricos, que apresentam dados de menor confiabilidade, optou-se pela construção de modelos.

A informação contida na coluna de interesse corresponde a variável resposta no modelo. No caso, a variável resposta é uma variável de distribuição binomial, pois é gerada pela repetição do experimento de Bernoulli (experimento de resposta dicotômica) (APÊNDICE 5).

Por ser necessário um modelo que abrangesse dados com distribuição não normal, foi escolhido o ajuste dos dados ao modelo linear generalizado (MLG), que compila vários modelos de distribuição não normal, incluindo dados de distribuição binomial (TURKMAN; SILVA, 2000).

Assim, os MLGs foram construídos considerando uma variável resposta binomial. O valor dessa variável pode estar relacionado ao programa, ao conjunto e/ou ao programa e o conjunto, que correspondem aos possíveis fatores do modelo.

A função de ligação aplicada foi a função canônica (logit ou logito), pois é capaz de transformar uma probabilidade de acerto que está no intervalo entre 0 e 1 (intervalo (0,1), que não inclui nem o 0 nem o 1 no intervalo) para um com infinitas possibilidades (intervalo $(-\infty, +\infty)$, que esta entre o menos infinito até o mais infinito) (RESENDE; BIELE, 2002). O parâmetro de dispersão para dados de dispersão binomial é igual a um (TURKMAN; SILVA, 2000).

O critério de informação de Akaike (*Akaike Information Criterion* - AIC) foi utilizado para avaliar o ajuste realizado com diferentes combinações de fatores para

identificar qual MLG está mais relacionado com a resposta de predição (CHRISTENSEN, 1997; TURKMAN; SILVA, 2000). No AIC o menor valor indica o melhor ajuste. Foram construídos cinco modelos: 1- determinado por um fator constante, 2- determinado pelo conjunto, 3- determinado pelo programa, 4- determinado pelo conjunto mais o programa e 5- determinado pelo conjunto mais o programa com a interação entre conjunto e programa. Os respectivos modelos são:

Modelo 1: $resposta \sim constante$

$$y_i = \beta a$$

Modelo 2: $resposta \sim conjunto$

$$y_i = \beta_{Conj1} Conj1_i + \cdots + \beta_{Conj12} Conj12_i$$

Modelo 3: $resposta \sim programa$

$$y_i = \beta_{Prog1} Prog1_i + \cdots + \beta_{Prog5} Prog5_i$$

Modelo 4: $resposta \sim conjunto + programa$

$$y_i = \beta_{Prog1} Prog1_i + \cdots + \beta_{Prog5} Prog5_i + \beta_{Conj1} Conj1_i + \cdots + \beta_{Conj12} Conj12_i$$

Modelo 5: $resposta \sim conjunto + programa + conjunto * programa$

$$y_i = \beta_{Prog1} Prog1_i + \cdots + \beta_{Prog5} Prog5_i + \beta_{Conj1} Conj1_i + \cdots + \beta_{Conj12} Conj12_i + \beta_{Inter1} Inter1_i + \cdots + \beta_{Inter60} Inter60_i$$

,

onde y é uma variável resposta, x são variáveis preditivas (covariáveis, parâmetros), coeficiente β é o estimador, i é número da unidade experimental e $i = 1, \ldots, n$; a é uma constante qualquer.

Além da capacidade de predição correta, foi avaliado se o tempo de execução de cada programa para cada conjunto de dados difere significativamente. A ANOVA foi calculada com os tempos de execução e o teste de Tukey foi realizado para determinar quais programas apresentam tempos de execução próximos (CALLEGARI-JACQUES, 2003).

A construção da tabela com os dados de predição, do MLG, da ANOVA, do teste de Tukey e dos gráficos e tabelas representando a análise dos dados foi realizada utilizando a linguagem R. As funções e os parâmetros utilizados estão presentes no Quadro 10. Outras características analisadas foram o tamanho do arquivo, os pré-requisitos, a instalação e a usabilidade.

FUNÇÃO	PARAMETROS UTILIZADOS	EXEMPLO
glm	Função, tipo da variável dependente e dados	glm(acerto ~ conjunto + programa, family=binomial, data=tabela1)
AIC	Saída da função glm	AIC(glm0,glm1, glm2, glm3)
aov	Função e dados	aov(Tempo~Programa, data=tempoExec)
TukeyHSD	Saída da função aov	TukeyHSD(aov.tempoExec)

QUADRO 10 - FUNÇÕES DA LINGUAGEM R UTILIZADAS NA PRESENTE ANÁLISE. FONTE: O autor (2013).

6.2 WORKFLOW

Cada conjunto de sequências foi submetido aos diferentes programas, que realizam a classificação das sequências. O identificador das sequências presentes nos arquivos com a classificação considerada correta foi previamente extraído. A classificação realizada pelos diferentes programas foi extraída, para permitir a padronização das saídas das diferentes fontes para um tipo de saída (APÊNDICE 6). A extração dos identificadores e das classificações foi realizada por um algoritmo desenvolvido em C++ para cada tipo de arquivo (tanto da saída dos programas quanto da classificação correta) para gerar um arquivo de saída para cada arquivo, para que os dados presentes nos diferentes tipos de saída apresentem a mesma formatação.

Esta saída consiste em um arquivo no formato txt com a lista de classificação para cada fonte (os programas e o arquivo com a classificação considerada correta) contendo duas colunas separadas por uma tabulação (primeira com os identificadores e a segunda com as classificações).

Esses arquivos padronizados serão os dados de entrada para um algoritmo em R, que normaliza os nomes tanto dos programas quanto da classificação referência (APÊNDICE 7) e gera:

1- tabela indicando o acerto em função do programa, conjunto de dados e sequência;

2- tabela com o identificador da sequência e as classificações para aquela sequência (classificação de referência mais as classificações geradas por cada programa);

3- diagrama de Venn representando os conjuntos de sequências com predição correta para cada programa;

4- diagrama de Venn representando os conjuntos de sequências que os programas não retornaram nenhuma classificação.

O fluxo de dados proposto foi construído utilizando a linguagem Shell script (APÊNDICE 8), por permitir a construção de laços de repetição, independer de interface gráfica para execução, ser de simples confecção e permitir registrar o tempo de execução de cada programa. O fluxograma representativo do workflow se encontra na FIGURA 11.

Testou-se o uso da linguagem Swift, porém o tempo total de execução não diferiu muito do total de execução do workflow em Shell script e não foi possível executar o InterProScan com a opção de execução em paralelo (APÊNDICE 9).

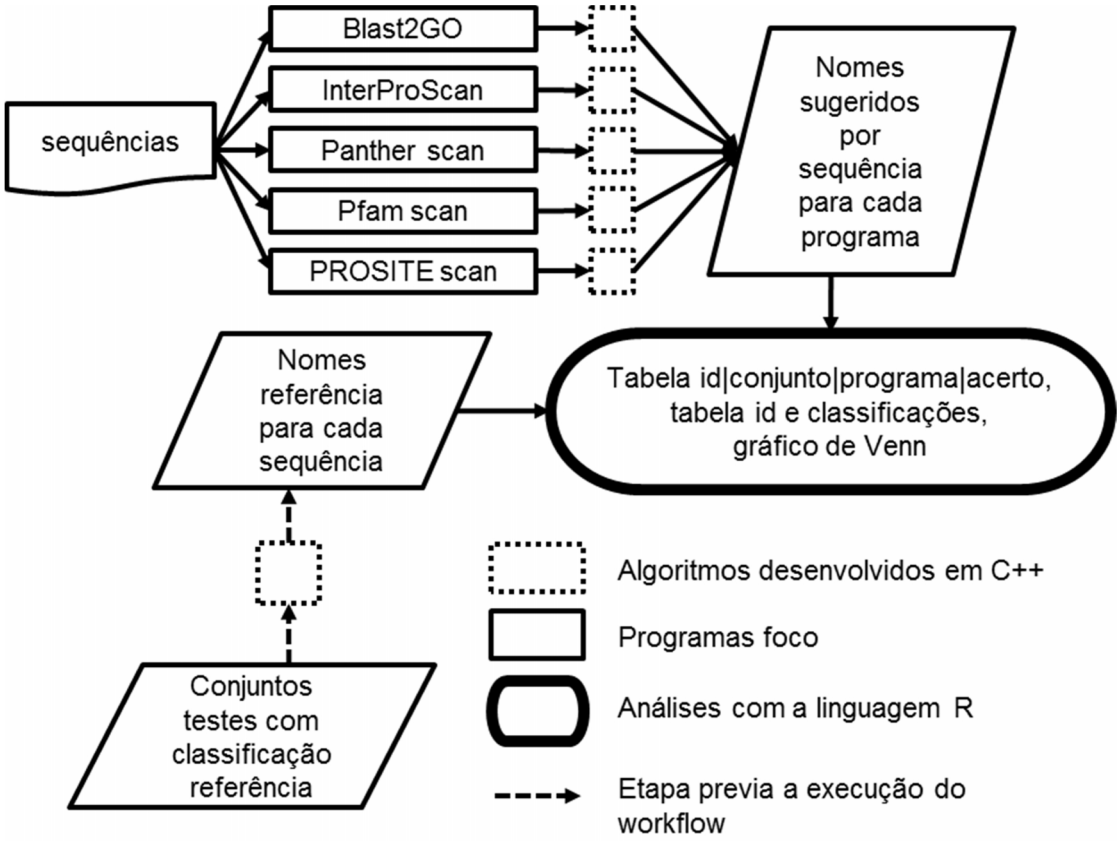


FIGURA 11 - FLUXOGRAMA DO WORKFLOW EM SHELL SCRIPT. As sequências de cada conjunto são submetidas aos diferentes programas para a classificação. Cada sequência é vinculada a cada classificação fornecida por cada programa. Esses dados e os dados da classificação padrão são os arquivos de entrada para um script desenvolvido em linguagem R que processa e analisa esses dados. FONTE: O autor (2013).

A tabela indicando o acerto em função do programa, conjunto de dados e sequência foi utilizada na construção dos MLG utilizando a função *glm* do R. Os

dados de tempo de execução dos programas foram utilizados para confeccionar uma tabela com o tempo de execução em função do programa e do conjunto. Essa tabela foi utilizada no teste estatístico de ANOVA.

6.3 EXTRAÇÃO DAS CLASSIFICAÇÕES DAS SEQUÊNCIAS PROTEICAS

Cada tipo de classificação foi submetido a um executável binário específico. O código fonte desses executáveis foi escrito em um código em C++ específico para cada tipo de classificação. Nas classificações referência a classificação alvo era a anotação funcional para cada sequência, sendo desenvolvidos quatro algoritmos para as quatro fontes de dados.

Para a extração das classificações dos programas foi desenvolvido um algoritmo para cada saída do programa, buscando a descrição funcional mais especifica que o programa pode fornecer, nos casos em que não foi possível especificar em parâmetros do programa.

Todas as saídas dos executáveis apresentam a mesma estrutura, que consiste em um arquivo txt contendo duas colunas espaçadas por uma tabulação, sendo a primeira coluna com o identificador das sequências e a segunda o nome da sequência.

7 RESULTADOS

A quantidade de sequência classificadas corretamente foi determinada pela contagem dessas sequências em função dos programas e dos conjuntos de sequências mostrando que os programas forneceram predições incorretas para mais da metade das sequências.

As semelhanças e diferenças entre os resultados dos programas analisados foram obtidas pela construção de MLGs e diagramas de Venn, onde se observou diferenças significativas entre os resultados de cada programa.

Para a análise dos tempos de execução foram aplicados os testes estatísticos ANOVA e de Tukey e verificando diferenças significativas nos tempos de execução foi possível distinguir dois grupos de programas com tempo de execução similar.

Com relação à usabilidade, foram descritas características de cada programa com relação à instalação, tipos de saída e espaço que ocupa no disco rígido.

7.1 RESULTADOS DAS ANÁLISES DOS PROGRAMAS DE PREDIÇÃO DE FUNÇÃO PROTEICA

A Tabela 2 ilustra os resultados obtidos pelos programas em estudo. Foram utilizados na avaliação dos programas de predição de proteína 12 conjuntos de sequências de proteínas. A taxa de acerto dos programas avaliados foi inferior a 35% para todos os conjuntos utilizados, exceto a taxa do InterProScan para o conjunto Vicinal oxygen chelate, onde a taxa de acerto foi superior a 50% (dado sublinhado na TABELA 2). Para quatro dos conjuntos utilizados na análise a taxa de acerto foi zero em todos os programas: 1- AminergicGPCR, 2- NHR, 3- Secretin-like, e 4- Crotonase.

TABELA 2 - NÚMERO DE PREDIÇÕES CORRETAS POR PROGRAMAS E POR CONJUNTO. O dado sublinhado se refere ao único caso em que a taxa de acerto de um programa para um conjunto de sequências de proteínas foi superior a 50%. Destacados em cinza estão os dados de contagem nula (em que não houve acerto).

CONJUNTO	N. SEQUÊNCIAS POR CONJUNTO	PROGRAMAS				
		BLAST2 GO	INTERPRO SCAN	PANTHER SCORE	PFAM SCAN	SCAN PROSITE
AminergicGPCR	358	0	0	0	0	0
NHR	412	0	0	0	0	0
Secretin-like	153	0	0	0	0	0
Enzimas	690	194	0	72	70	24
Não Enzimas	487	102	42	48	38	7
Bifuncional	60	11	4	5	2	1
Enolase	927	183	242	78	0	0
Crotonase	262	0	0	0	0	0
Haloacid dehalogenase	389	65	1	43	0	0
Vicinal oxygen chelate	145	50	91	0	0	39
Radical SAM	145	41	10	27	0	0
Padrão-ouro	863	259	203	58	9	112
Total	4891	905	593	331	119	183

FONTE: O autor (2013).

Os cinco programas estudados apresentaram divergências nos resultados. Os programas Pfam scan e ScanProsite compartilharam poucas sequências classificadas corretamente por ambos. Os programas InterProScan e Panther Score como os anteriores também compartilharam poucas predições (GRÁFICO 5).

Os programas com maior quantidade de acertos são o Blast2GO e InterProScan. Esses dois programas também compartilham mais sequências classificadas corretamente. Porém os resultados compartilhados pelos dois programas correspondem a menos que 50% dos acertos totais nos dois programas (GRÁFICO 5).

Por volta de 80% das sequências classificadas corretamente pelo programa Panther Score e mais de 50% das sequências classificadas corretamente pelo Pfam scan estão incluídas nas sequências classificadas corretamente pelo Blast2GO (GRÁFICO 5).

O ScanProsite foi o que menos compartilhou sequências classificadas corretamente com os demais programas, seguido pelo InterProScan. Nenhuma sequência foi predita corretamente por todos os programas (GRÁFICO 5).

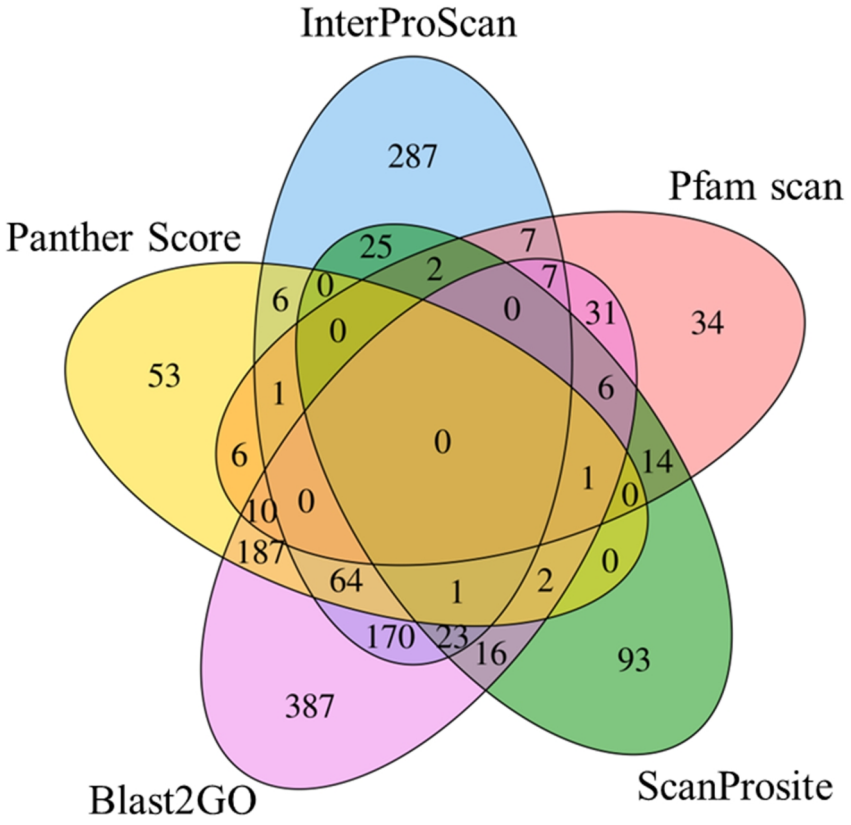


GRÁFICO 5 - DIAGRAMA DE VENN COM AS SEQUÊNCIAS CLASSIFICADAS CORRETAMENTE PELOS DIFERENTES PROGRAMAS. FONTE: O autor (2013).

Considerando apenas a quantidade de acerto houve efeito significativo ($p<0,001$) quando se elegeu para a construção dos MLGs quatro critérios: 2- apenas o efeito dos programas, 3- apenas o efeito dos conjuntos, 4- o efeito dos programas e dos conjuntos, e 5- o efeito dos programas, o efeito dos conjuntos e o efeito da interação entre cada programa e cada conjunto. O modelo linear que considerou o critério 5 foi o que melhor representou os resultados de acerto, segundo o AIC (Figura 12) (TABELA 3).

TABELA 3 - AIC PARA CADA MLG PROPOSTO.

MODELO LINEAR GENERALIZADO	GRAUS DE LIBERDADE	AIC
Nenhum fator	1	13.227
Conjunto	8	12.998
Programa	5	12.130
Conjunto e programa	12	11.885
Conjunto e programa, com interação entre o conjunto e o programa	40	10.887

FONTE: O autor (2012)

$$resposta \sim conjunto + programa + conjunto * programa$$

$$y_i = \beta_{Prog1} Prog1_i + \cdots + \beta_{Prog5} Prog5_i + \beta_{Conj4} Conj4_i + \cdots + \beta_{Conj7} Conj7_i + \beta_{Conj9} Conj9_i + \cdots + \beta_{Conj12} Conj12_i + \beta_{Inter1} Inter1_i + \cdots + \beta_{Inter60} Inter40_i$$

FIGURA 12 – MODELO QUE MELHOR REPRESENTOU OS RESULTADOS DE ACERTO. Modelo que considera o efeito dos programas, o efeito dos conjuntos e o efeito da interação entre cada programa e cada conjunto. FONTE: O autor (2013).

Utilizando os modelos construídos pode-se observar a probabilidade de acerto de cada programa em cada modelo e nos diferentes conjuntos. O MLG que considera apenas o efeito dos programas apresentou a probabilidade de acerto para todos os programas é inferior a 30% (Gráfico 6a).

A probabilidade de acerto dos programas e o desvio variam dependo do conjunto de dados nos MLGs que consideram o efeito do conjunto, chegando em alguns casos ser superior ao valor de 30% (Gráfico 6b, 6c e Gráfico 7).

Destaca-se que o MLG que considera a interação entre programa e conjunto de dados foi o único a apresentar uma probabilidade de acerto superior a 50% (Gráfico 6c e Gráfico 7).

Os programas Panther Score, Pfam scan e ScanProsite apresentam um desempenho inferior aos programas Blast2GO e InterProScan nos três modelos em que se considerou o efeito dos programas (Gráfico 6).

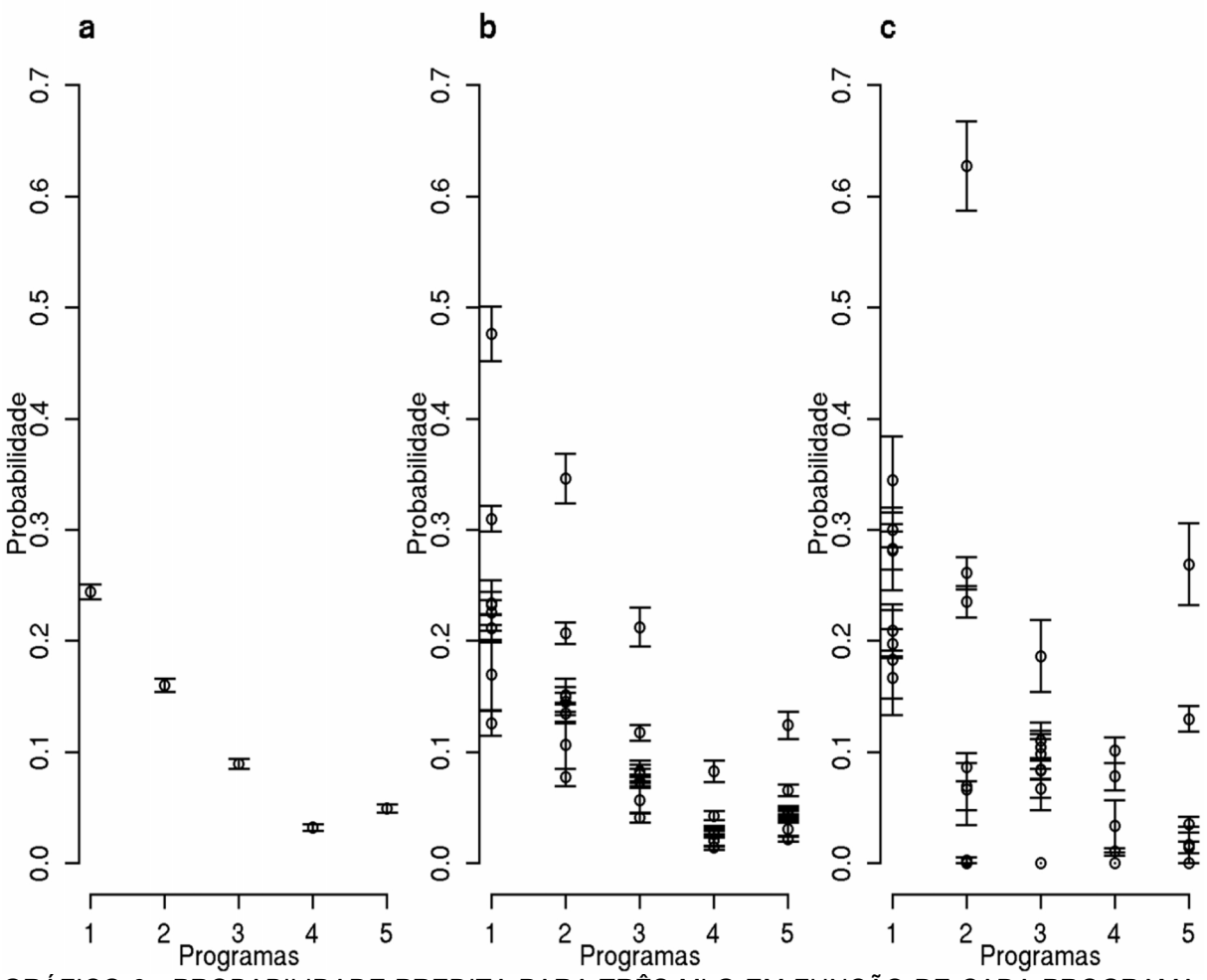


GRÁFICO 6 - PROBABILIDADE PREDITA PARA TRÊS MLG EM FUNÇÃO DE CADA PROGRAMA. Probabilidade predita para os programas (1)- Blast2GO, (2)- InterProScan, (3)- Panther Score, (4)- Pfam scan e (5)- ScanProsite para os MLG (a)- apenas o efeito dos programas, (b)- o efeito dos programas e dos conjuntos, e (c)- o efeito dos programas, o efeito dos conjuntos e o efeito da interação entre cada programa e cada conjunto. Para os MLG (b) e (c) ponto representa um conjuntos de dados teste. FONTE: O autor (2013).

O programa Blast2GO se destaca em todos os MLGs que consideram o efeito dos programas na predição de função de proteínas. Mas a MLG 4 foi selecionada por apresentar o melhor ajuste entre os modelos (Tabela 3). Considerando esse modelo 4, pode-se notar que o programa InterProScan apresentou uma probabilidade maior de acerto apenas nos conjuntos Vicinal oxygen chelate e o Enolase (Gráfico 7).

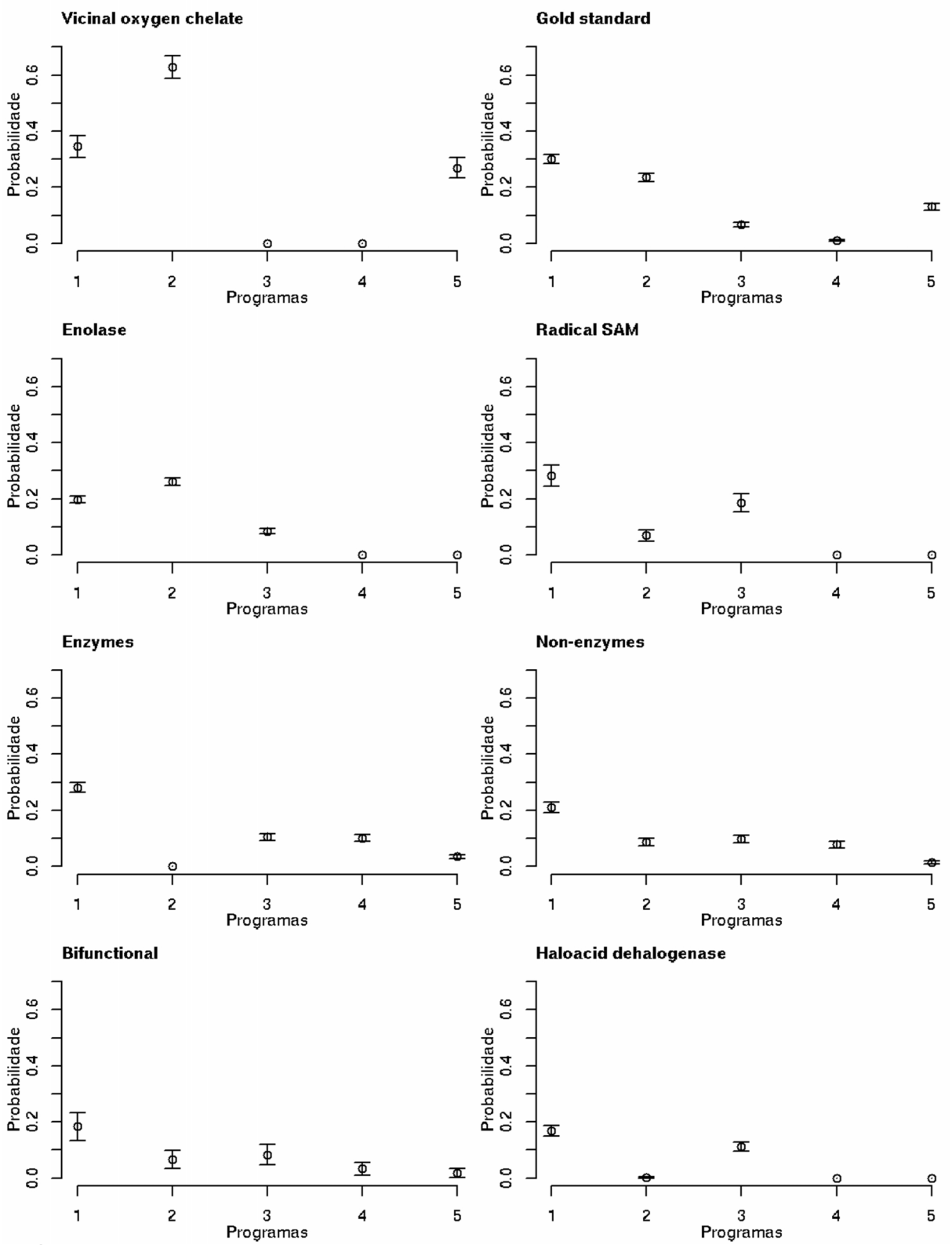


GRÁFICO 7 - PROBABILIDADE PREDITA DO MLG 4 EM FUNÇÃO DE CADA PROGRAMA PARA CADA CONJUNTO. Probabilidade predita para os programas (1)- Blast2GO, (2)- InterProScan, (3)- Panther Score, (4)- Pfam scan e (5)- ScanProsite para o MLG considerando o efeito dos programas, o efeito dos conjuntos e o efeito da interação entre cada programa e cada conjunto para cada conjunto apresentado. FONTE: O autor (2013).

Houve diferenças significativas entre os conjunto e programas em relação ao conjunto de dados Padrão-ouro com Blast2GO na maioria dos casos, principalmente para os programas InterProScan e Panther Score (Tabela 4).

Foi observada a contagem nula do InterProScan para o conjunto Enzimas, do Pfam scan e do ScanProsite para os conjuntos Enolase, Haloacid dehalogenase e Radical SAM, e do Panther score e do Pfam scan para o conjunto Vicinal oxygen chelate (Tabela 2). Essas contagens nulas não disponibilizam a informação necessária para a avaliação desses programas nesses conjuntos de dados. As contagens nulas são responsáveis pela ocorrência de desvios padrões elevados no MLG e de não haver diferenças significativas ($p>0,1$) em relação ao conjunto de dados Padrão-ouro com Blast2GO.

TABELA 4 - RESULTADOS DO MLG 4 EM RELAÇÃO AO BLAST2GO E O CONJUNTO PADRÃO-OURO. Notar as interações significativas entre os conjuntos e programas (p<0,05).

CONJUNTO	PROGRAMA	PARÂMETRO	DESVIO PADRÃO	SIGNIFICÂNCIA
Enzimas	InterProScan	-18,29506	409,3983	-
	Panther Score	0,57255	0,21605	**
	Pfam scan	2,46345	0,37532	***
	ScanProsite	-1,32836	0,25713	***
Não Enzimas	InterProScan	-0,69985	0,22453	**
	Panther Score	0,89863	0,24395	***
	Pfam scan	2,56479	0,39843	***
	ScanProsite	-1,84344	0,41608	***
Bifuncional	InterProScan	-0,81284	0,62541	-
	Panther Score	0,87968	0,59456	-
	Pfam scan	1,83259	0,86392	*
	ScanProsite	-1,52745	1,0696	-
Enolase	InterProScan	0,69436	0,15607	***
	Panther Score	0,79886	0,21168	***
	Pfam scan	-14,45755	353,2083	-
	ScanProsite	-17,10735	353,20816	-
Haloacid dehalogenase	InterProScan	-4,02236	1,01637	***
	Panther Score	1,30477	0,26195	***
	Pfam scan	-14,25375	545,25023	-
	ScanProsite	-16,90355	545,25013	-
Vicinal oxygen chelate	InterProScan	1,49602	0,26831	***
	Panther Score	-17,14056	893,0722	-
	Pfam scan	-15,21825	893,07225	-
	ScanProsite	0,69814	0,28527	*
Radical SAM	InterProScan	-1,33958	0,39163	***
	Panther Score	1,23962	0,32174	***
	Pfam scan	-14,92929	893,07225	-
	ScanProsite	-17,57909	893,0722	-

Significância: (-) > 0,1; (.) <0,1; (*) <0,05; (**) <0,01; (***) <0,001
FONTE: O autor (2013).

No presente trabalho, as predições incorretas puderam ser classificadas em cinco casos: 1- totalmente divergente do padrão, 2- grafias diferentes de um mesmo nome, 3- classificação padrão ser mais especifica que a dada pelo programa, 4- classificação do padrão ser de um subtipo diferente, 5- a classificação padrão ser mais genérica que a do programa. Exemplos de cada um dos casos estão no Quadro 11.

CLASSIFICAÇÃO PADRÃO	EXEMPLOS DE CLASSIFICAÇÕES	
	PADRÃO	PROGRAMA
Mais específica	Purine nucleoside phosphorylase	Nucleoside phosphorylase
Em um nível mais específico	Pyruvate kinase (família)	Pyruvate kinase, C-terminal-like (domínio)
Em um subtipo diferente	L-phenylalanine dehydrogenase	Leucine dehydrogenase
Mais genérica	DNA polymerase	DNA-directed DNA polymerase, family B
Com grafia diferente	Dethiobiotin synthetase	Dethiobiotin synthase
Totalmente diferente	Integrase	Polymerase polyprotein

QUADRO 11 - TIPO DE DIFERENÇAS ENTRE A CLASSIFICAÇÃO PADRÃO E A DADA POR UM DOS PROGRAMAS TESTADOS. FONTE: O autor (2013).

Além das sequências com predições incorretas, houve sequências que os programas não retornaram predições. Os programas que apresentaram resultados com mais sequências sem classificação foram o Panther Score e o ScanProsite, e o que apresentou menos foi o Pfam scan (Tabela 5).

TABELA 5 – NÚMERO DE SEQUÊNCIAS POR CONJUNTO EM QUE OS PROGRAMAS NÃO RETORNARAM PREDIÇÕES. Em destaque os resultados com mais sequências sem classificação (Panther Score e o InterProScan), e o que apresentou menos (Pfam scan).

CONJUNTO	N. SEQUÊNCIAS	BLAST2GO	INTERPRO SCAN	PANTHER SCAN	PFAM SCAN	SCAN PROSITE
AminergicGPCR	358	4	0	358	0	9
NHR	412	3	412	388	3	405
Secretin-like	153	0	4	8	3	38
Enzimas	690	14	13	397	7	463
Não Enzimas	487	46	15	324	17	367
Bifuncional	60	2	1	39	1	32
Enolase	927	33	16	623	1	396
Crotonase	262	2	0	262	0	65
Haloacid dehalogenase	389	0	46	331	0	292
Vicinal oxygen chelate	145	2	0	102	0	61
Radical SAM	145	0	1	110	1	120
Padrão-ouro	863	12	160	673	3	211
Total	4891	118	668	3615	36	2459

FONTE: O autor (2013).

Também houve divergências entre os programas com relação às sequências que cada um não foi capaz de predizer a função (Gráfico 8). Os programas Panther scan e ScanProsite, que tiveram o maior número de sequências sem classificação.

Eles compartilham a maior quantidade de sequências que sem classificação, totalizando 1.755 sequências não classificadas pelos dois programas (Gráfico 8).

O Pfam scan foi o programa que apresentou menor número de sequências sem classificação, sendo que todas as sequências não classificadas por esse programa também não foram classificadas por outros programas (Gráfico 8).

O InterProScan foi o terceiro programa que apresentou o menor número de sequências sem classificação. Ele foi segundo programa com mais sequências não classificadas compartilhadas com outros programas, compartilhando mais sequências sem classificação com os programas Panther scan e ScanProsite (Gráfico 8).

Das sequências sem classificação, dez não foram classificadas simultaneamente pelos cinco programas. Dessas dez sequências, uma é do conjunto teste de NHR, três são do conjunto teste de enzimas e seis restantes são do conjunto teste de não enzimas (Quadro 12).

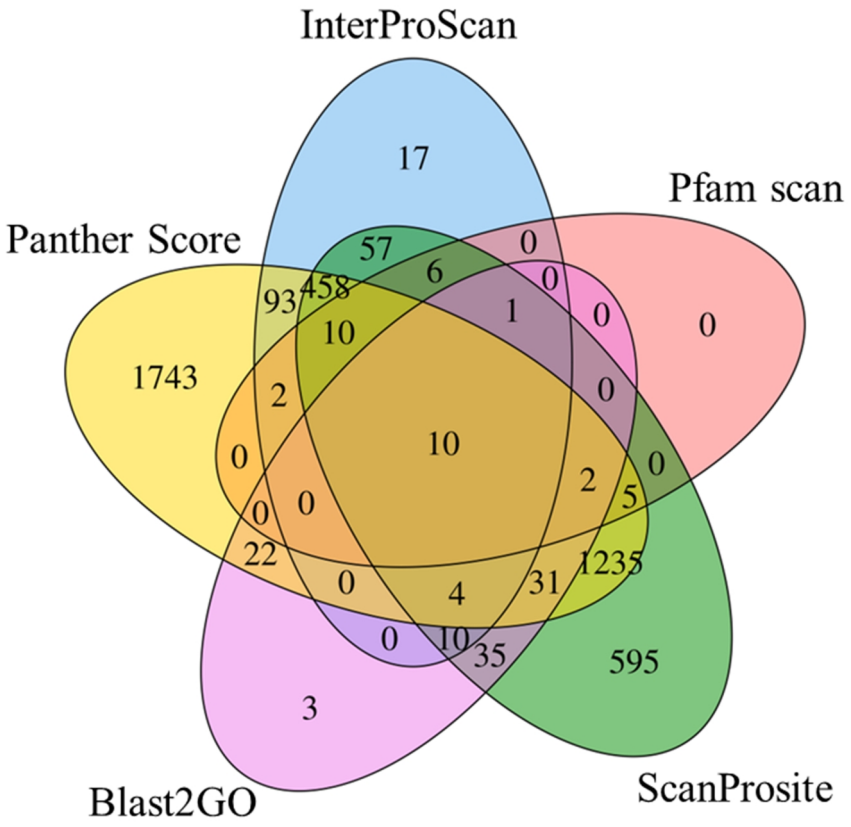


GRÁFICO 8 - DIAGRAMA DE VENN COM AS SEQUÊNCIAS SEM CLASSIFICAÇÃO NOS RESULTADOS DOS DIFERENTES PROGRAMAS. FONTE: O autor (2013).

CONJUNTO TESTE	NOME PADRÃO
NHR	5AFushitarazuF1likeFTZF1FTFSF1A
Enzimas	RIBONUCLEASE PANCREATIC
	TETRAPEPTIDE (GLY SER ASN SER)
	GAMMA CHYMOTRYPSIN
Não enzimas	RETRO-GCN4 LEUCINE ZIPPER
	COLLAGEN-LIKE PEPTIDE
	FUSION PROTEIN BETWEEN THE HYDROPHOBIC POCKET OF HIV GP41 AND GCN4-PIQI
	CELL DIVISION PROTEIN FTSZ
	CYCLOSPORIN A
	RIGHT-HANDED COILED COIL TETRAMER

QUADRO 12 – AS 10 SEQUÊNCIAS QUE NÃO FORAM CLASSIFICADAS SIMULTANEAMENTE PELOS CINCO PROGRAMAS. FONTE: O autor (2013)

Avaliando o tempo de execução, os programas apresentaram tempos de execução significativamente distintos ($p<0,01$), sendo que os programas Panther Score, Pfam scan e ScanProsite apresentam tempos de execução inferiores ao dos programas Blast2GO e InterProscan (Gráfico 9) (APÊNDICE 10).

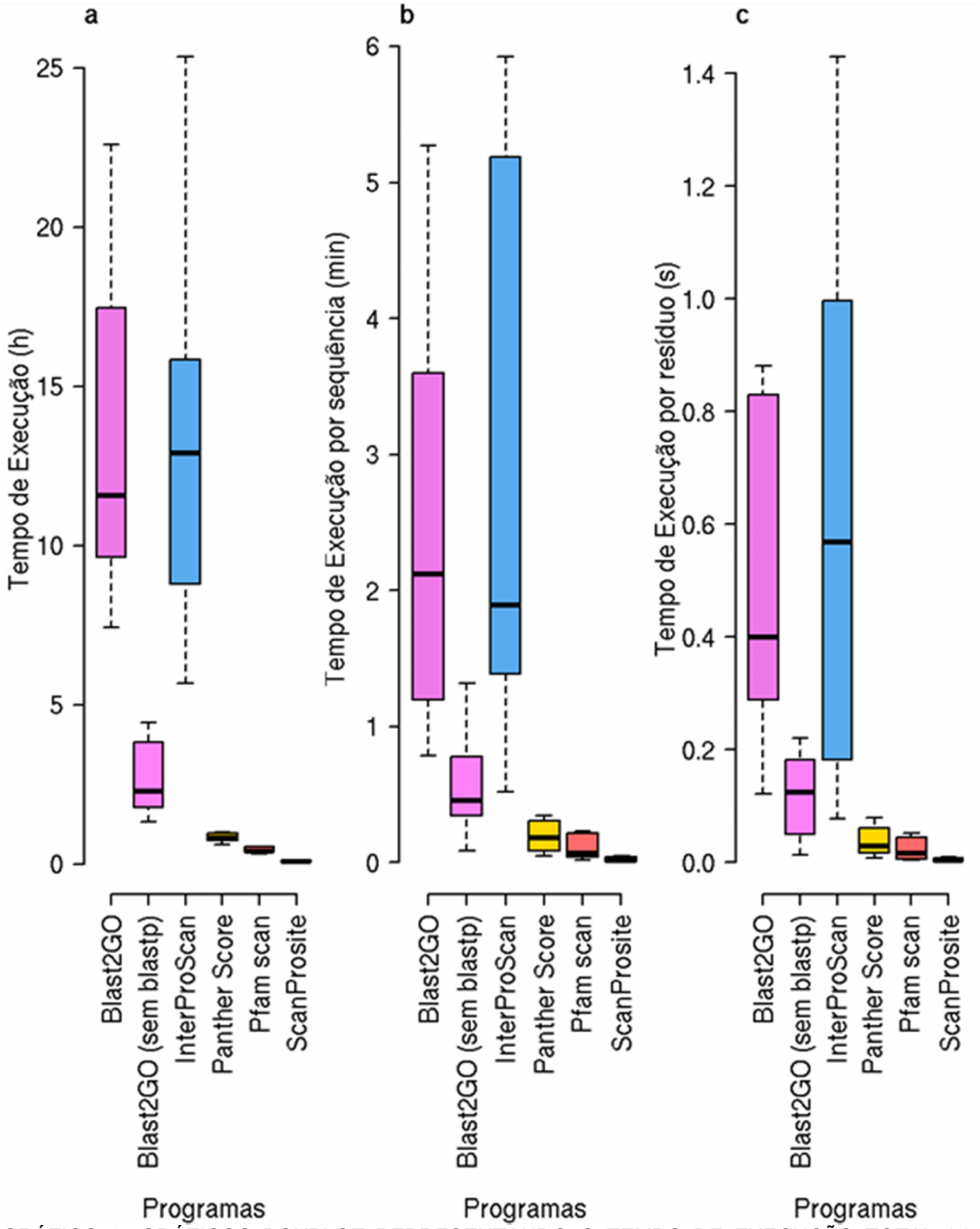


GRÁFICO 9 -GRÁFICOS BOXPLOT REPRESENTANDO O TEMPO DE EXECUÇÃO TOTAL (**a**), POR SEQUÊNCIA (**b**) E POR AMINOÁCIDO (**c**). FONTE: O autor (2013).

7.2 CARACTERÍSTICAS

As características dos programas estudados estão apresentadas no QUADRO 11.

Nome	Prompt	GUI	Inclui genes hipotéticos*	Pré-requisitos	Espaço utilizado	Tipo de busca	Tipo de saída
Blast2GO	Sim	Sim	Sim	JRE, mySQL server	≈40GB	Alinhamento pelo Blast e pontuação por termos do GO	txt, dat, annot
InterProscan	Sim	Não	Não	Biblioteca xml do perl	≈39GB	Procura a assinatura no banco de dados próprio manualmente curado que corresponde a combinação de saídas das diferentes bases integradas	raw, xml, ebixml, txt, html
Panther Score	Sim	Não	Não	HMMER2, blast+	≈20GB	Comparação de HMM	csv
Pfam scan	Sim	Não	Não	HMMER3	≈7GB	Comparação de HMM	txt
Scan Prosite	Sim	Não	Não	-	≈32MB	Por perfis (matriz de pesos)	txt, fasta, psa, msa, pff

*embora exista um algoritmo buscando evitar nomes pouco informativos (como nomes contendo "*hypothetical*"), sequências hipotéticas podem ser inclusas caso o nome da sequência não seja filtrado.

QUADRO 13 - AMBIENTE DE EXECUÇÃO, PRÉ-REQUISITOS, ESPAÇO EM DISCO, BASES DE DADOS ASSOCIADA E FORMATOS DE ARQUIVOS DOS PROGRAMAS BLAST2GO, INTERPROSCAN, PANTHER SCORE, PFAM SCAN E SCANPROSITE. FONTE: O autor (2013).

Os programas InterProScan, Panther Score, Pfam scan e ScanProsite apresentam instalações mais simples e de fácil compreensão. Elas não demandam conhecimentos mais aprofundados do sistema operacional. Por sua vez o programa Blast2GO apresentou instalação mais complexa demandando conhecimentos avançados do sistema operacional por depender da instalação local da base de dados do GO. A necessidade do Blast2GO de possuir localmente a base de dados do GO agrega um complicador a mais pelo fato de exigir mais espaço em disco.

8 DISCUSSÃO

Com o advento de novas tecnologias de sequenciamento genômico os experimentos em laboratório para caracterização funcional dos produtos polipeptídicos não ocorrem num ritmo compatível com o ritmo de descoberta de novas sequências, tornando-se necessário o desenvolvimento de métodos mais ágeis que a anotação de proteínas com dados experimentais (GERLT, J. A. et al., 2012).

Anotações utilizando sistemas computacionais para predição de função de proteínas são uma valiosa alternativa às anotações baseadas em dados experimentais e têm contribuído para acelerar o processo de caracterização de sequências proteicas (RENTZSCH; ORENGO, 2009).

Diferentes artigos de revisão avaliaram sistemas computacionais preditores de função de proteínas, compilando dados dos artigos originais que descrevem cada sistema. As metodologias utilizadas nestes artigos de revisão avaliaram características como acurácia, sensibilidade, especificidade e termos do GO. Os métodos de avaliação são os gráficos ROC, tabela de contingência e similaridade entre termos do GO (BLABY-HAAS; CRÉCY-LAGARD, 2011; HENRY et al., 2011; PANDEY; KUMAR; STEINBACH, 2006; RENTZSCH; ORENGO, 2009) (ANEXO 8).

Não há relatos de avaliações comparando mais que um preditor de função de proteína com dados da nomenclatura de função, apenas comparando classificações segundo termos do GO. O presente trabalho realizou uma comparação com dados da nomenclatura de função utilizando MLGs.

Através de MLGs foi possível analisar quais características afetaram positivamente a capacidade de predição de função de proteínas. Possibilita considerar simultaneamente todos os programas, todos os conjuntos de dados e todas as classificações de função de proteínas preditas, sem perda de informação.

A avaliação simultânea sem perda de informação citada no parágrafo anterior não seria possível com metodologias que empregam dados de tabelas de contingência e focam em características como sensibilidade e especificidade (APÊNDICE 10), pois estas metodologias, além de analisarem apenas um programa e um conjunto de dados por vez, exigiriam:

1 – Reduzir as classificações de função de proteínas preditas em duas categorias (THOMAS *et al.*, 2003) ou;

2 – Comparar o resultado de uma classificação contra as demais classificações (PRATI; BATISTA; MONARD, 2008).

As análises dos MLGs possibilitaram determinar:

1 – os conjuntos de dados utilizados que apresentaram um efeito significativo na capacidade de realizar uma predição correta;

2 – os programas que apresentaram um efeito significativo na capacidade de realizar uma predição correta e;

3 – a interação entre um programa e um determinado conjunto que afeta a capacidade de predição.

Além disso, os MLGs permitem trabalhar com os dados de contagem de acertos sem recorrer a métodos não paramétricos, compilam vários modelos de distribuição não normal e outros modelos e vários programas de computador permitem a sua utilização de maneira simples (TURKMAN; SILVA, 2000).

O presente trabalho utilizou MLG para analisar os cinco sistemas computacionais Blast2GO, InterProScan, Panther Score, Pfam scan e ScanProsite, aplicados aos 12 conjuntos AminergicGPCR, NHR, Secretin-like, Enzimas, Não Enzimas, Bifuncional, Enolase, Crotonase, Haloacid dehalogenase, Vicinal oxygen chelate, Radical SAM e Padrão-ouro.

Dentre os MLGs construídos, aquele que melhor representou os dados de classificação analisados foi o que considerou os sistemas de predição, os conjuntos de teste e a interação sistema/conjunto. Ou seja, a capacidade de realizar uma predição de função correta apresentou influência do conjunto de dados.

Essa relação entre sistema/conjunto de dados pode explicar diferenças observadas em relação ao desempenho de programas entre artigos. Como por exemplo, o programa Blast2GO é descrito como apresentando acurácia de 70% para anotação de proteínas de *Arabidopsis* sp no trabalho de Conesa e Götz (2008) e 47,7% de F-score[4] no trabalho de Björne e Salakoski (FRIEDBERG; LINIAL; RADIVOJAC, 2011), ambos os dados extraídos de análises utilizando comparação entre termos do GO.

Na presente análise, o Blast2GO foi o programa que apresentou diferenças significativas com relação aos demais e que apresentou maior capacidade de

[4] F-score: media harmônica entre precisão e sensibilidade

predizer corretamente a função das sequências dos conjuntos de dados testados, segundo o MLG 4, que melhor representou os dados de classificação analisados.

Embora o F-score do InterProScan possa ser de 48%[5] (FRIEDBERG; LINIAL; RADIVOJAC, 2011), que é um valor próximo ao de 47,7% obtido pelo Blast2GO (FRIEDBERG; LINIAL; RADIVOJAC, 2011), no presente trabalho o capacidade preditiva do Blast2GO foi significativamente superior ao do InterProScan e aos demais programas.

O Blast2GO depende dos arquivos de saída da ferramenta Blast, que geralmente é vinculada com a base de dados NR. Era esperado que apresentasse uma maior taxa de acerto, pois a base de dados NR apresenta 18.972.433 sequências (dados da versão do NR utilizada no presente trabalho), que incluem algumas as sequências presentes nos conjuntos de dados utilizados (a bases de dados NR inclui sequências depositadas na base de dados PDB[6]).

O Blast2GO através da pontuação de termos GO prioriza anotação com base em dados experimentais (CONESA; GÖTZ, 2008), o que também pode ter contribuído para a maior capacidade preditiva.

Mas seu tempo de execução foi maior que os tempos de execução do Panther Score, Pfam scan e ScanProsite. O sistema do Blast2GO depende do sistema de busca do Blast e de dados recuperados da base de dados do GO, aumentando assim o tempo de execução e o espaço ocupado no disco rígido.

Durante a instalação o Blast2GO necessitou de mais pré-requisitos e conhecimento sobre bases de dados que os demais programas. O Blast2GO demanda a instalação de uma base de dados MySQL, depende da instalação do Java e para a execução sem instalação local necessita da instalação do Java Web Start.

Embora o Blast2GO tenha se destacado por sua capacidade de predição ser superior aos demais sistemas, a taxa de acerto dos sistemas para os conjuntos utilizados na construção dos modelos foi inferior a 35% para os sistemas nos

---

[5] Valor calculado com base em dados de um gráfico

[6] http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=ProgSelectionGuide#db

diferentes conjuntos teste, excetuando o InterProScan no conjunto Vicinal oxygen chelate.

Houve pouca coincidência entre as classificações realizadas pelos sistemas. Como por exemplo, o InterProScan compartilhou poucas sequências preditas corretamente com o Panther Score, Pfam scan e ScanProsite, mesmo o InterProScan apresentndo uma integração com as bases de dados Panther, Pfam e ProSite (ZDOBNOV; APWEILER, 2001).

Embora os Blast2GO e o InterProScan sejam os programas que mais compartilham sequências preditas corretamente, essas sequências ainda representam menos da metade dos total de acertos tanto do Blast2GO quanto do InterProScan. Não houve nenhuma sequência em que a predição foi feita corretamente por todos os cinco sistemas.

Houve uma maior sobreposição das sequências sem classificação. Houve dez sequências em que todos os sistemas não foram capazes de realizar a predição. E cada programa compartilha com pelo menos um programa a maioria das sequências sem classificação.

Um exemplo de como os programas apresentam contribuições diferentes para a predição de função é o fato do InterProScan ter apresentado uma probabilidade de acerto superior ao Blast2GO no conjunto teste Vicinal oxygen chelate e no conjunto Enolase.

Isso pode ser decorrente de características comuns às sequências desses conjuntos. A superfamília Vicinal oxigen chelate apresenta motivos pareados característicos, embora apresentem baixa similaridade entre sequências (ARMSTRONG, 2000; HE; MORAN, 2011). A superfamília Enolase apresenta um sítio ativo comum, mas com grande diversidade de funções (BABBITT *et al.*, 1996; GERLT *et al.*, 2012).

Dessa forma a caracterização por similaridade entre pares de sequências, como o Blast2GO, apresentaria uma capacidade menor de predizer corretamente a função para essas famílias. Buscas por motivos e sítios ativos apresentariam maior capacidade de predição correta, como o InterProScan, que associa informações de bases de dados de motivos, sítios ativos e entre outras.

Assim, o Blast2GO provavelmente terá uma probabilidade menor de classificar corretamente sequências que apresentam mesma função, mas com baixa similaridade. Ou sequências com funções diferentes com sequência similares, pois

estudos vêm indicando que a transferência de anotação com dados de similaridade pode não ser a melhor opção, mesmo com alta similaridade entre as sequências (BLABY-HAAS; CRÉCY-LAGARD, 2011; CHITALE; KIHARA, 2011; FRIEDBERG, 2006).

7 CONCLUSÃO

O Blast2GO e o InterProScan se destacaram por apresentar uma quantidade predições corretas maior em relação aos demais programas. Porém foram os programas que demandaram maior tempo de execução.

O Blast2GO se destacou com uma probabilidade de acerto significativamente maior que os demais programas na maioria dos conjuntos. Mas foi o programa que apresentou uma instalação mais complexa.

Essa probabilidade maior de acertar a predição de função de uma proteína apresentada pelo Blast2GO pode ser decorrente da presença da sequência dos conjuntos teste na base de dados do NR, que é a base de dados padrão para a realização do Blastp.

Por trabalhar sobre os dados do Blast, o resultado de predição do Blast2GO está vinculado à sequências similares e já conhecidas. Isso pode levar à um baixa taxa de acerto para situações em que a similaridade entre sequências não está relacionada à função, como foi observado nos casos dos conjuntos teste Vicinal oxigen chelate e Enolase.

Dessa forma o InterProScan se torna uma alternativa interessante para a predição de função de proteínas por: 1- ter se destacado juntamente com o Blast2GO na quantidade de acertos, 2- ter apresentado uma probabilidade de acerto maior que o Blast2GO para dois dos conjuntos testados (Vicinal oxigen chelate e Enolase), e 3- por não se basear na similaridade entre sequências proteicas.

Iniciativas como a do InterProScan, para unificar diferentes esforços de caracterização de sequências podem melhorar a qualidade da predição de função de proteínas. Pode-se agregar outras abordagens como a curadoria automática além dos dados de curadoria manual do InterPro, e abordagens considerando nomes de sequência existentes, como a proposta do BioSOM por Otemaier (2012)[7].

---

[7] Otemaier, K. R. BioSOM: Metodologia para identificação de sinônimos de genes utilizando Self-Organizing Maps. 138 p. Dissertação (Mestrado em Bioinformática) – Setor de Educação Profissional e Tecnológica, Universidade Federal do Paraná. Curitiba, 2012

O acerto dos programas apresentado no presente trabalho (por volta de 35%) e os dados de acerto encontrados na literatura (de no máximo 80%) expõe que o desempenho desses sistemas computacionais para a predição de função de proteínas utilizando dados da sequência de aminoácidos apresentam uma probabilidade de erro geralmente maior que 20%, o que resulta em uma alta probabilidade de erro.

Com essa capacidade de predição correta abaixo de desejado e como outras informações além da sequência de aminoácidos podem não estar disponíveis para as sequências de proteínas recém-descobertas, sistemas que permitem a predição de função de proteínas a partir apenas de dados da sequência de aminoácidos ainda são uma demanda.

REFERÊNCIAS

ALTSCHUL, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. **Nucleic acids research**, v. 25, n. 17, p. 3389-402, 1997.

APWEILER, R et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. **Nucleic acids research**, v. 29, n. 1, p. 37-40, 2001.

APWEILER, ROLF et al. A large-scale protein-function database. **Nature chemical biology**, v. 6, n. 11, p. 785, 2010. Nature Publishing Group.

ARMSTRONG, R. N. Current Topics Mechanistic Diversity in a Metalloenzyme Superfamily †. , v. 39, n. 45, 2000.

BABBITT, P C et al. The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids. **Biochemistry**, v. 35, n. 51, p. 16489-501, 1996.

BLABY-HAAS, C. E.; CRÉCY-LAGARD, V. DE. Mining high-throughput experimental data to link gene and function. **Trends in biotechnology**, v. 29, n. 4, p. 174-82, 2011.

BROWN, D.; KRISHNAMURTHY, N.; SJÖLANDER, K. Automated protein subfamily identification and classification. **PLoS computational biology**, v. 3, n. 8, p. 1526-1538, 2007.

BROWN, S. D.; GERLT, JOHN A; SEFFERNICK, J. L.; BABBITT, PATRICIA C. A gold standard set of mechanistically diverse enzyme superfamilies. **Genome biology**, v. 7, n. 1, p. R8, 2006.

CALLEGARI-JACQUES, S. M. **Bioestatística: princípios e aplicações**. Porto Alegre: Artmed, 2003.

CAMPOS, D.; MATOS, S.; LEWIN, I.; OLIVEIRA, J. L.; REBHOLZ-SCHUHMANN, D. Harmonization of gene/protein annotations: towards a gold standard MEDLINE. **Bioinformatics (Oxford, England)**, v. 28, n. 9, p. 1253-61, 2012.

CASTRO, E. et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. **Nucleic acids research**, v. 34, n. Web Server issue, p. W362-5, 2006.

CHEN, F.; MACKEY, A. J.; VERMUNT, J. K.; ROOS, D. S. Assessing performance of orthology detection strategies applied to eukaryotic genomes. **PloS one**, v. 2, n. 4, p. e383, 2007.

CHEN, L.; LIU, H.; FRIEDMAN, C. Gene name ambiguity of eukaryotic nomenclatures. **Bioinformatics (Oxford, England)**, v. 21, n. 2, p. 248-56, 2005.

CHITALE, M.; KIHARA, DAISUKE. Computational protein function prediction: framework and challenges. In: Daisuke Kihara (Ed.); **Protein Function Prediction for Omics Era**. p.1-17, 2011. Dordrecht: Springer Netherlands.

CHRISTENSEN, R. **Log-linear models and logistic regression**. 2nd ed. New York: Springer, 1997.

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

CONESA, A.; GÖTZ, S. Blast2GO: A comprehensive suite for functional analysis in plant genomics. **International journal of plant genomics**, v. 2008, p. 619832, 2008.

DOBSON, P. D.; DOIG, A. J. Distinguishing Enzyme Structures from Non-enzymes Without Alignments. **Journal of Molecular Biology**, v. 330, n. 4, p. 771-783, 2003.

FONTANA, P.; CESTARO, A.; VELASCO, R.; FORMENTIN, E.; TOPPO, S. Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. **PloS one**, v. 4, n. 2, p. e4619, 2009.

FRIEDBERG, I. Automated protein function prediction--the genomic challenge. **Briefings in bioinformatics**, v. 7, n. 3, p. 225-42, 2006.

FRIEDBERG, I.; JAMBON, M.; GODZIK, A. New avenues in protein function prediction. **Protein science**, v. 15, p. 1527-1529, 2006.

FRIEDBERG, I.; LINIAL, M.; RADIVOJAC, P. Automated Function Prediction Featuring a Critical Assessment of Function Annotations (AFP/CAFA) 2011. ISMB Special Interest Group Meeting. **Anais...** p.79, 2011. Vienna.

GALPERIN, M. Y.; FERNÁNDEZ-SUÁREZ, X. M. The 2012 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. **Nucleic acids research**, v. 40, n. Database issue, p. D1-8, 2012.

GATTIKER, A.; GASTEIGER, E.; BAIROCH, A. ScanProsite: a reference implementation of a PROSITE scanning tool. **Applied bioinformatics**, v. 1, n. 2, p. 107-8, 2002.

GENOME RESEARCH LTD. Pfam Scan version 1.3. ,2010. Disponível em: <ftp://ftp.sanger.ac.uk/pub/databases/Pfam/Tools>.

GERLT, J. A.; BABBITT, PATRICIA C.; JACOBSON, M. P.; ALMO, S. C. Divergent evolution in enolase superfamily: strategies for assigning functions. **The Journal of biological chemistry**, v. 287, n. 1, p. 29-34, 2012.

GERSTEIN, M. B. et al. What is a gene, post-ENCODE? History and updated definition. **Genome research**, v. 17, n. 6, p. 669-81, 2007.

HAWKINS, T.; LUBAN, S.; KIHARA, D. Enhanced automated function prediction using distantly related sequences and contextual association by PFP. **Protein Science**, v. 15, p. 1550-1556, 2006.

HE, P.; MORAN, G. R. Structural and mechanistic comparisons of the metal-binding members of the vicinal oxygen chelate (VOC) superfamily. **Journal of inorganic biochemistry**, v. 105, n. 10, p. 1259-72, 2011. Elsevier Inc.

HENRY, C. S. et al. Connecting genotype to phenotype in the era of high-throughput sequencing. **Biochimica et biophysica acta**, v. 1810, n. 10, p. 967-77, 2011. Elsevier B.V.

HUNTER, S. et al. InterPro: the integrative protein signature database. **Nucleic acids research**, v. 37, n. Database issue, p. D211-5, 2009.

JANGA, S. C.; DÍAZ-MEJÍA, J. J.; MORENO-HAGELSIEB, G. Network-based function prediction and interactomics: the case for metabolic enzymes. **Metabolic engineering**, v. 13, n. 1, p. 1-10, 2011.

JEFFERY, C J. Moonlighting proteins. **Trends in biochemical sciences**, v. 24, n. 1, p. 8-11, 1999.

JEFFERY, CONSTANCE J. Moonlighting proteins--an update. **Molecular bioSystems**, v. 5, n. 4, p. 345-50, 2009.

KEGG. Kyoto Encyclopedia of Genes and Genomes. Disponível em: <http://www.genome.jp/kegg/>.

LAIDLER, K. J. The development of theories of catalysis. **Archive for History of Exact Sciences**, v. 35, n. 4, p. 345-374, 1986.

LIU, H.; HU, Z.-Z.; ZHANG, JIAN; WU, C. BioThesaurus: a web-based thesaurus of protein and gene names. **Bioinformatics (Oxford, England)**, v. 22, n. 1, p. 103-5, 2006.

LOEWENSTEIN, Y. et al. Protein function annotation by homology-based inference. **Genome biology**, v. 10, n. 2, p. 207, 2009.

MARCOTTE, E. M. et al. Detecting Protein Function and Protein-Protein Interactions from Genome Sequences. **Science**, v. 285, n. 5428, p. 751-753, 1999.

MI, H. et al. PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. **Nucleic acids research**, v. 38, n. Database issue, p. D204-10, 2010.

MOUNT, D. W. **Bioinformatics: Sequence and Genome Analysis**. 2 ed. ed. New York: Cold Spring Harbor Laboratory Press, 2004.

NATURE PUBLISHING GROUP. The name game. **Nature cell biology**, v. 5, p. 1-2, 2003.

NCBI. National Center for Biotechnology Information. Disponível em: <http://www.ncbi.nlm.nih.gov/>.

NIKOLOSKI, Z.; GRIMBS, S.; KLIE, S.; SELBIG, J. Complexity of automated gene annotation. **Bio Systems**, v. 104, n. 1, p. 1-8, 2011. Elsevier Ireland Ltd.

PANDEY, G.; KUMAR, V.; STEINBACH, M. **Computational approaches for protein function prediction: A survey**. Minneapolis, 2006.

PILLET, V.; ZEHNDER, M.; SEEWALD, A. K.; VEUTHEY, A.-L.; PETRAK, J. GPSDB: a new database for synonyms expansion of gene and protein names. **Bioinformatics (Oxford, England)**, v. 21, n. 8, p. 1743-4, 2005.

PRATI, R. C.; BATISTA, G. E. A. P. A.; MONARD, M. C. Curvas ROC para avaliação de classificadores. **Revista IEEE América Latina**, v. 6, n. 2, p. 1-8, 2008.

PUNTA, M. et al. The Pfam protein families database. **Nucleic acids research**, v. 40, n. Database issue, p. D290-301, 2012.

RENTZSCH, R.; ORENGO, C. A. Protein function prediction--the power of multiplicity. **Trends in biotechnology**, v. 27, n. 4, p. 210-9, 2009.

RESENDE, M. D. V. DE; BIELE, J. Estimação e predição em modelos lineares generalizados mistos com variáveis binomiais. **Revista de Matemática e Estatística**, v. 20, p. 39-65, 2002.

SCHNOES, A. M.; BROWN, S. D.; DODEVSKI, I.; BABBITT, PATRICIA C. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. **PLoS computational biology**, v. 5, n. 12, p. e1000605, 2009.

SIKIC, K.; CARUGO, O. Protein sequence redundancy reduction : comparison of various methods Bioinformation. **Bioinformation**, v. 5, n. 6, p. 234-239, 2010.

SONNHAMMER, E. L.; EDDY, S R; DURBIN, R. Pfam: a comprehensive database of protein domain families based on seed alignments. **Proteins**, v. 28, n. 3, p. 405-20, 1997.

TAMAMES, J.; VALENCIA, A. The success (or not) of HUGO nomenclature. **Genome biology**, v. 7, n. 5, p. 402, 2006.

THOMAS, PAUL. PANTHER HMM scoring tools version 1.03. ,2011. Disponível em: <http://www.pantherdb.org/downloads/>.

THOMAS, PD; CAMPBELL, M.; KEJARIWAL, A. PANTHER: a library of protein families and subfamilies indexed by function. **Genome research**, v. 13, p. 2129-2141, 2003.

TIPTON, K; BOYCE, S. History of the enzyme nomenclature system. **Bioinformatics (Oxford, England)**, v. 16, n. 1, p. 34-40, 2000.

TSURUOKA, Y.; MCNAUGHT, J.; ANANIADOU, S. Normalizing biomedical terms by minimizing ambiguity and variability. **BMC bioinformatics**, v. 9, n. Suppl 3, p. S2, 2008.

TURKMAN, M. A. A.; SILVA, G. L. **Modelos Lineares Generalizados: da teoria à prática**. Lisboa: Edições SPE, 2000.

VALENCIA, A. Automatic annotation of protein function. **Current opinion in structural biology**, v. 15, n. 3, p. 267-74, 2005.

ZDOBNOV, EVGENI M; APWEILER, ROLF. InterProScan - an integration platform for the signature-recognition methods in InterPro. **Bioinformatics Applications Note**, v. 17, n. 9, p. 847-848, 2001.

APÊNDICES

APÊNDICE 1 – RECURSOS DE PREDIÇÃO DE FUNÇÃO DE PROTEÍNAS PRESENTES EM DIFERENTES REVISÕES. Linhas com texto em cinza indicam recursos que não apresentaram funcionamento adequado do site ou do sistema de busca, ou não estão mais disponíveis.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
3DCoffee	O'Sullivan *et al.* 2004	Alinhamento múltiplo de sequências e estruturas	http://igs-server.cnrs-mrs.fr/Tcoffee/
AFAWE	Jocker *et al.* (2008)	Integração de diferentes ferramentas e workflows para a predição automática gerando um resultado condensado das informações	http://bioinfo.mpiz-koeln.mpg.de/afawe/
AutoFACT	Koski *et al.* 2005	Sistema automático de análise e anotação de sequência baseado base de dados de genomas	
BETE	Sjolander 1997	Método - Construção de árvores filogenéticas a partir de genomas	
Bio-Dictionary	Rigoutsos *et al.* 1999	Dicionário de partes de sequências que cobrem uma sequência de entrada, podendo ser usada para anotar a proteína com uma função	
Blast	Altschul *et al.* 1990	Ferramenta de alinhamento de sequências	http://blast.ncbi.nlm.nih.gov/
BLAST2GO	Conesa *et al.* (2005)	Pontua termos do GO para elencar o nome da lista que o BLAST retorna que melhor representa a sequência de interesse	http://www.blast2go.com
CE	Shindyalov e Bourne 1998	Alinhamento par a par de estrutura	http://source.rcsb.org/jfatcatserver/ceHome.jsp
CHUGO	Eisner *et al.* 2005	Sistema de classificação binária que classifica uma proteína em um ou mais termos do GO	http://webdocs.cs.ualberta.ca/~bioinfo/CHUGO/
CODENSE	Hu *et al.* 2005	Algoritmo que utiliza o algoritmo MODES para construir um subgrupo de grafos densos de um grafo representando a interação entre as proteínas	http://zhoulab.usc.edu/CODENSE/
CombFunc	Wass e Sternberg (2008)	Combinação de diversos métodos (ConFunc, BLAST, InterPro, Pfam, Phyre2, MINT, IntAct, COXPRESdb e 3DLigandSite) usando SVM para realizar a predição baseada no GO	http://www.sbg.bio.ic.ac.uk/~mwass/combfunc/
CSC	Stoica e Hearst 2007	Metodologia - predição de função de gene a partir de texto usando uma abordagem entre espécies	

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
Dali	Holm e Sander 1995	Metodologia - mensura a relação estrutural pela similaridade de matrizes de distâncias intramoleculares	
DaliLite	Holm e Park 2000	Alinhamento par a par de estrutura	http://www.ebi.ac.uk/Tools/dalilite/
EFICAz	Tian *et al.* (2004)	Sistema de classificação de enzimas seguindo o padrão EC que utiliza busca por resíduos discriminantes de função, busca por padrões do PROSITE, comparação de sequências e SVM	http://cssb.biology.gatech.edu/skolnick/webservice/EFICAz2/index.html
eggNOG	Jensen *et al.* (2008)	Constrói grupos não-supervisionados de ortólogos por alinhamentos de Smith-Waterman através identificação de melhores alinhamentos recíprocos e agrupamento por triangulação	http://eggnog.embl.de/version_3.0/
EUCLID	Tamames *et al.* 1998	Sistema usado pelo ProtFun para associar categorias funcionais as palavras-chave do Swiss-Prot	http://www.gredos.cnb.uam.es/EUCLID
EVEREST	Portugaly *et al.* (2006)	Método automático não supervisionado que identifica regiões conservadas recorrentes com base em similaridades locais e busca por padrões interativos	http://www.everest.cs.huji.ac.il/
EzyPred	Shen e Chou (2007)	Classificação por k-nearest neighbor para fundir as abordagens de FunD (Functional Domain) e Pse-PSSM (Pseudo Position-Specific Scoring Matrix) para uma predição em 3 camadas (se é enzima, classe funcional e sub-classe funcional)	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/
FAT-CAT	Berkeley Phylogenomic Group	Identifica ortólogos e realiza a predição de função usando HMMs para inferência filogenética	http://phylogenomics.berkeley.edu/phylofacts/fatcat/
FCANAL	Suzuki *et al.* 2005	Utiliza a definição de 2 a 4 resíduos com importância funcional e uma matriz de similaridade local com os demais resíduos nas comparações de estrutura 3D	http://atgc002.ps.noda.tus.ac.jp/contents/fcanal/
FFF	Fetrow e Skolnick 1998	Método - Fuzzy Functional Form retorna assinatura funcional por comparação de estruturas e por estrutura de motifs utilizando a comparação de descritores 3D fuzzy de funções especificas e a geometria, identidade do resíduo e sítios ativos confirmados.	
FFPred		Classificação por máquinas de vetores de suporte (Support Vector Machines - SVM) de proteínas em termos de processo biológico e função molecular do GO	http://bioinf.cs.ucl.ac.uk/ffpred
FIGENIX*	Gouret *et al.* (2005)	Constrói uma árvore de ortólogos utilizando resultados do Blast, Clustaw, HMMPfam, Tree-Puzzle e PAUP	http://figenix2.up.univ-mrs.fr/Figenix/

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
Finger-PRINTScan	Scordis *et al.* 1999	Métodos – busca por partes em sequências	
FunctionalFlow	Nabieva *et al.* 2005	Algoritmo que utiliza grafos para representar as interações entre proteínas através de uma adaptação do algoritmo max-flow min-cut (West 2001), em que as proteínas são classificadas como tendo ou não uma função defenida	http://compbio.cs.princeton.edu/function/
GAIN	Karaoz *et al.* 2004	Método – algoritmo que extrai informações a respeito de função da interação entre conjuntos de dados de *S. cerevisiae* e *H. sapiens*	
GeneFIND	Wu *et al.* 1998	Identificação de famílias de proteicas e recuperação de informação	http://diana.uthct.edu/
GeneParser	Snyder e Stormo 1995	Identificação de genes em organismo eucariotos, utilizando programação dinâmica	http://beagle.colorado.edu/~eesnyder/GeneParser.html
GeneQuiz	Andrade *et al.* 1999	Sistema de análise semiautomático de inferência de função e anotação de sequências	http://cbdm.mdc-berlin.de/ma_genequiz
GenMultiCut	Vazquez *et al.* 2003	Algoritmo para determinar classes funcionais com base em uma rede de interações físicas com o número mínimo de interações entre diferentes categorias funcionais	
GenScan	Burge e Karlin 1997	Identificação de genes em organismo eucariotos, utilizando HMM	http://genes.mit.edu/GENSCAN.html
GO Engine	Xie *et al.* 2002	Sistema que integra homologia de sequência com dados textuais (resumos do PubMed) para a predição de função	
GoAnno	Chalmel *et al.* 2005	Sistema que usa a informação de alinhamentos múltiplos para recuperar a informação evolucionaria e anota utilizando termos do GO	http://bips.u-strasbg.fr/GOAnno/GOAnno.html
Goblet	Hennig *et al.* 2004	Anotação de sequências por similaridade utilizando dados do GO	http://goblet.molgen.mpg.de/cgi-bin/goblet2008/goblet.cgi
GOFigure	Khan *et al.* 2003	Utiliza resultados do BLAST para realizar a anotação automática da sequência com termos do GO	http://udgenome.ags.udel.edu/frm_go.html/

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
GOPET	Vinayagam *et al.* 2006	Sistema baseado em SVM para anotação automatica da sequência com termos do GO	http://genius.embnet.dkfz-heidelberg.de/menu/cgi-bin/w2h-open/w2h.open/w2h.startthis?SIMGO=w2h%2ewelcome
GOSLING	Jones *et al.* (2008)	Utiliza resultados do BLAST para retornar um conjunto de termos do GO para uma sequência (selecionados por conjunto de árvore de decisão) com pontuação indicativa de acurácia	https://www.sapac.edu.au/gosling
GOtcha	Martin *et al.* (2004)	Usa a similaridade de sequência para buscar termos do GO associados a ela, retornando uma lista de termos pontuada.	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php
GTREE	Berkeley Phylogenomic Group	Programa de várias etapas envolvendo dados de bancos de dados e de outros programas	http://phylogenomics.berkeley.edu/software/
InParanoid	Remm *et al.* (2001)	Método automático de geração de grupos ortólogos e inparálogos a partir de resultados do Blast de cada par de sequências	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
InterProScan	Hunter *et al.* (2009)	Busca em uma base de dados que integra as assinaturas das diferentes bases de dados em uma única assinatura	http://www.ebi.ac.uk/interpro/
iProClass	Wu *et al.* 2003	Sistema de classificação funcional do PIR que integra várias bases de dados	http://pir.georgetown.edu/pirwww/dbinfo/iproclass.shtml
JAFA	Friedberg *et al.* 2006	Sistema que integra resultados de algoritmos baseados em sequências para predizer a função (como GOFigure, GOtcha, GOblet, Phydbac2 e InterProScan	http://jafa.burnham.org
LSQMAN	Kleywegt 1996	Alinhamento par a par de estrutura	http://xray.bmc.uu.se/usf/mol_morph.html
Majority-rule	Schwikowski *et al.*, 2000	Método - predição de função de proteínas com base em rede de interação	
MAGIC	Troyanskaya *et al.* 2003	Sistema estruturado em redes Bayesian que incluí dados de expressão genica, co-localização, fator de transcrição, associação funcional e associação de genes	http://genome-www.stanford.edu/magic/
MATRAS	Kawabata 2003	Programa compara estruturas 3D de proteínas para realizar a classificação	http://strcomp.protein.osaka-u.ac.jp/matras/

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
MEME	Bailey *et al.* 1997	Algoritmo usado para extrair cerca de 30 motifs que foram usados na construção de vetores para cada sequência	http://meme.nbcr.net/meme/
MFGO	Sun *et al.* 2006	Realiza a predição de função com dados de interação de proteína-proteína	http://www.bioinfo.org.cn/MFGO/
MILANO	Rubinstein e Simon 2005	Sistema que a partir de um conjunto de identificadores de genes e de um termos, liga ao PubMed e GeneRIF pelo LocusLink	http://bioinformatics.ekmd.huji.ac.il/milano/
MultiProt	Shatsky *et al.* 2004	alinhamento múltiplo de estrutura	http://bioinfo3d.cs.tau.ac.il/MultiProt/
Neighborhood	Hishigaki *et al.* 2001	Método - predição da localização subcelular, papel celular e função bioquímica	
OntoBlast	Zehetner (2003)	A partir de uma sequência identifica os termos do GO uma lista de sequências fornecidas pelo Blast propondo uma anotação funcional	http://functionalgenomics.de/ontogate/
Operon	Strong *et al.* 2003	Método - comparação genômica (vizinhança genica, padrões filogenéticos e fusão de genes) e validação pela combinação de aproximadamente 10 métodos	
OrthoMCL	Li *et al.* (2003)	Método automático de agrupamento de ortólogos que utiliza uma pesquisa de um grupo de proteínas Blast de todos contra todos para o proteoma de cada espécie, seguida de uma normalização por agrupamentos de Markov	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi
PANTHER		Cadeias de Markov para comparar a sequência de interesse com a base de dados	http://www.pantherdb.org/
PathBLAST	Kelley *et al.* 2003	Algoritmo que trabalha com redes de vias metabólicas, permitindo o alinhamento de redes de uma espécie pouco conhecida com uma mais conhecida permitindo elucidar a função de proteínas da espécie pouco conhecida	http://www.pathblast.org/
PEDANT	Riley *et al.* 2005	Sistema automático de análise e anotação de sequência baseado base de dados de genomas	http://pedant.gsf.de/
Pfam	Sonnhammer *et al.* (1997)	Base de dados de famílias proteicas na forma de bibliotecas de HMM, onde uma sequêmcia pode ser mais a uma dessas bibliotecas	http://pfam.sanger.ac.uk/

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
PFP	Hawkings, T. *et al.* (2008)	Lista de nome do PSI-BLAST com pontuações por termos GO e dados de PPI	http://kiharalab.org/web/pfp.php
PHUNCTIONER	Pazos e Sternberg 2004	Método - alinhamento estrutural da base de dados FSSP para encontrar as posições na estrutura proteica que são as funcionalmente mais importantes para uma categoria particular do GO para realizar a classificação	
Phydbac	Enault *et al.* 2005	Sistema que realiza a predição de função de proteínas bacteriana no contexto genomico	http://www.igs.cnrs-mrs.fr/phydbac/indexPS.html
PIRSF	Wu *et al.* (2004)	Sistema de classificação de sequências em níveis: família homeomórfica, subfamílias e superfamílias	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml
PLEX	Date e Marcotte 2005	Sistema interativo de busca por proteínas com perfis filogenéticos similares, com ligações com fusão de genes e vizinhança de genes	http://bioinformatics.icmb.utexas.edu/plex
PolyFARM	Clare e King 2003	Método - algoritmo que usa dados de expressão genica, fenótipo, homologia de sequência e predição da estrutura secundária para a classificação funcional	
PRED-CLASS	Pasquier *et al.* 2001	Sistema de classificação funcional de proteínas em transmembrana, fibrosa e globular de três níveis	http://athina.biol.uoa.gr/PRED-CLASS/
PredictProtein	ROSTLAB	Sistema de predição automática que trabalha com alinhamento multiplo de sequências, motivos do PROSITE e regiões de baixa complexidade	http://www.predictprotein.org/
PRIAM	Claudel-Renard *et al.* (2003)	Método de detecção de enzimas a partir de um genoma gerado a partir de todas as sequêmcias contidas na base de dados ENZYME e o programa MKDOM	http://priam.prabi.fr/
ProCANS	Wu *et al.* 1992	Método - *Protein classification artificial neural system*. sistema que utiliza redes neurais para classificar sequências proteicas (que são entendidas como um vetor de caracteres - n-gram), porém o desempenho é afetado pela correta representação do vetor de caracteres	
PRODISTIN	Brun *et al.* 2003	Calcula a distância entre duas proteínas e utiliza o algoritmo BioNJ para agrupá-las ou algoritmo baseado em densidade conforme a sua função celular (seria a mais relevante para predição da função)	http://crfb.univ-mrs.fr/webdistin/

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
ProFound	Zhang e Chait 2000	Método - algoritmo para identificação de sequências de proteínas por dados de espectrometria de massa, utilizando métodos comparativos como fragmentação de peptídeos e massa peptídica finger-printing	
Prolinks	Bowers *et al.* 2004	Inferência de interação e função de proteínas	http://prl.mbi.ucla.edu/prlbeta/prolinks.jsp
PropSearch	Hobohm e Sander (1995)	Método que busca a sequência com menor distância euclidiana entre as propriedades da sequências e as do banco de dados (144 propriedades, como hidrofobicidade, peso molecular e ponto isoelétrico), quando buscas por similaridade não foram satisfatórias	http://abcis.cbs.cnrs.fr/propsearch/propsearch.html
PROSITE	Sigrist *et al.* (2002)	Base de dados de domínios proteicos, famílias e sítios funcionais são que uma sequência pode conter	http://prosite.expasy.org/
ScanProsites	Scordis *et al.* 1999	Métodos de busca por partes em sequências	http://prosite.expasy.org/
ProFAT	Bradshaw *et al.* (2006)	Combina busca por similaridade com busca textual para realizar a anotação funcional de proteínas	http://cluster-1.mpi-cbg.de/profat/profat.html
ProFAL	Couto *et al.* 2003	Método - *PROtein Functional Annotation through Literature*. sistema que liga documentos do PubMed com GenBank, Swiss-Prot e PDB, a enzima de interesse é anotada segundo termos do GO presentes nos documentos obtidos pelas ligações	
ProtFun	Jensen *et al.* 2002	Rede neural que utiliza 14 atributos para predizer a função de proteínas. Utiliza características de modificação pós-traducionais, sinais direcionadores de proteínas e propriedades físico-químicas. Foca especialmente para predição de função de sequências sem sequências homólogas.	http://www.cbs.dtu.dk/services/ProtFun/
ProtoNet	Sasson *et al.* (2003)	Conjunto de famílias proteicas geradas por agrupamento aglomerativo automático de 94 mil sequências de proteínas do UniProt	http://www.protonet.cs.huji.ac.il/
PSI-BLAST	Altschul *et al.* 1997	Busca por sequências homólogas	http://www.ebi.ac.uk/Tools/sss/psiblast/

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
pvSOAR	Edelsbrunner *et al.* 1998	Similaridade da superfície de para predizer a função biológica de proteínas	http://pvsoar.bioengr.uic.edu/
RAST	Aziz *et al.* 2008	Anotação de genomas de bactérias e arqueobactérias	http://rast.nmpdr.org/
RIO	Zmasek e Eddy (2002)	Análise automática de filogenomas que possibilita a detecção automática de representantes de novas subfamílias	http://rio.janelia.org/
PSI-BLAST	Altschul *et al.* 1997	Métodos de busca de sequências posição específico	http://www.ebi.ac.uk/Tools/sss/psiblast/
SIFTER	Engelhardt *et al.* 2005	Transfere a função molecular de um nó pai para um nó filho em um árvore filogenética usando propagação probabilística.	http://sifter.berkeley.edu/
SMART	Schultz *et al.* (1998)	Domínios proteicos manualmente curados a partir de bibliotecas de HMM e de alinhamentos de sequências	http://smart.embl-heidelberg.de/
SPASM	Kleywegt 1999	Retorna assinatura funcional por comparação de estruturas e por estrutura de motifs	http://xray.bmc.uu.se/usf/spasm.html
SPLASH	Califano 2000	Algoritmo que caracteriza sequências com base na presença ou ausência de motifs	ww.research.ibm.com/splash/
STRING	Snel *et al.* (2000)	Método de representação interativa de interações de um grupo de proteínas na forma de rede	http://string-db.org/
SUPERFAMILY	Gough *et al.* (2001)	Base de dados de domínios baseados na classificação estrutural de proteínas (SCOP) na forma de bibliotecas de HMM. A busca de uma sequência é pela sua semelhança com alguma desses registros	http://supfam.cs.bris.ac.uk/SUPERFAMILY/
SVM-Prot	Cai *et al.* (2003)	Classificação por máquinas de vetores de suporte (Support Vector Machines - SVM) de proteínas em famílias funcionais utilizando dados como volume normalizado de Van der Waals, polaridade, carga e tensão superficial de todos os aminoácidos da proteína	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi
SynFPS	Li *et al.* 2007	Sistema de classificação SVM (support vector machine) que a partir de grupos de genes	http://www.synteny.net/
TEIRESIAS	Rigoutsos e Floratos 1998	Método - algoritmo usado para descobrir quais subsequências a sequência de entrada contém	

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO	SITE
TESS	Wallace *et al.* 1997	Método - algoritmo que retorna assinatura funcional por comparação de estruturas e por estrutura de motifs	
T-HMM	Qian e Goldstein 2003	Constrói um HMM para cada nó da árvore usando alinhamento múltiplo de sequência	por correspondência com o autor - richard.goldstein@nimr.mrc.ac.uk
TribeMCL	Pereira-Leal *et al.* 2003	Método - algoritmo que usa interações presentes no DIP e medidas de entropia para formar grupos homogêneos	
VIRGO	Massjouni *et al.* 2006	Programa de predição de função que utiliza redes de interação com dados de microarranjo	http://whipple.cs.vt.edu:8080/virgo
WILMA	Prlic *et al.* 2004	Sistema que integra várias bases de dados tanto as de proteína (Swiss-prot, IPI e WORMPEP) como as de subsequências (PROSITE, Pfam e PRINTS) para anotação funcional em larga escala	http://www.came.sbg.ac.at/wilma/
WIT	Overbeek *et al.* 2000	Análise comparativa de sequência de genomas e reconstrução metabólica baseada em base de dados de sequências cromossômicas e módulos metabólicos	http://wit.mcs.anl.gov/WIT2/

APÊNDICE 2 - BASES DE DADOS CITADAS EM DIFERENTES REVISÕES

SIGLA	CITAÇÃO	DESCRIÇÃO
ArrayExpress	Parkinson *et al.* 2005	Expressão genica
Barley DB	Shen *et al.* 2005	Expressão genica
BGED	Matoba *et al.* 2000	Expressão genica
BIND	Alfarano *et al.* 2005	Interações e complexos proteína-proteína de uma serie de organismos
BodyMap	Sese*et al.* 2001	Expressão genica
BRENDA	Schomburg *et al.* 2004	Dados experimentais de enzimas
CATH	Orengo *et al.* 1997	Sistema de classificação hierárquica de domínios proteicos
CAZy	Coutinho e Henrissat	Enzima
CDD	Marchler-Bauer *et al.* 2005	Domínios proteicos
CGED	Kato *et al.* 2005	Expressão genica
COGs	Tatusova *et al.* 1997	Genes ortólogos de uma série de genomas
CuraGen	Uetz *et al.* 2000	Interação física proteína-proteína de S. cerevisiae
CYGD	Guldener *et al.* 2005	Interação física, genética e complexos proteína-proteína de S. cerevisiae

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO
DIP	Xenarios *et al.* 2002	Interação física proteína-proteína de *S. cerevisiae*, *D. melogaster* e *E. coli*
Drosophila	Neal *et al.* 2003	Expressão genica
DSMP	Guruprasad *et al.* 2000	Estrutura 3D de motifs
eF-site	Kinoshita *et al.* 2001	Superfície de um conjunto de estruturas de proteínas do PDB extraídas com o programa MSP, anotando o potencial eletrostático e a hidrofobicidade
EnsEMBL	Hubbard *et al.* 2005	Alguns mamíferos
FlyBase	FlyBase Consortium 2003	Sequências proteicas de *D. melanogaster*
FunCat	Rueppe*t al.* 2004	Antigo MIPS/PEDANT (Mewes *et al.* 2002), baseado em fenômenos biológicos gerais de uma ampla variedade de espécies
GenBank	Benson *et al.* 2004	Sequência de genes
Gene3D	Buchan *et al.* 2002	Base de dados de domínios proteicos do CATH com assinaturas para genomas do ENSEMBL e sequências do Uniprot
GeneNote	Safran *et al.* 2003	Expressão genica
GeneRIF	Mitchell *et al.* 2003	Resumo de artigos sobre genes conhecidos
GenProtEC	Serres *et al.* 2004	Sequências proteicas de *E. coli*
GEO	Barrett *et al.* 2005	Expressão genica
GOA	Camon *et al.* 2003	Anotações do GO e id de documentos do MEDLINE com suporte para anotação
GPCRDB	Horn *et al.* 2003	Receptores acoplados de proteína G
GRID	Breitkreutz *et al.* 2003	Interação física proteína-proteína de *S. cerevisiae*, *D. melogaster* e *C. elegans*
GSDB	Harger *et al.* 1998	Genome Sequence DataBase of the Nacional Center for Genome Resources
GXD	Hill *et al.* 2004	Expressão genica
HMS-PCI	Ho *et al.* 2002	Complexos proteína-proteína de *S. cerevisiae*
IPI	Kersey*et al.* 2004	Sequências proteicas
KEGG	Kanehisa*et al.* 2004	Rotas metabólicas
LGIdb	Novre e Changex 2001	Canais iônicos dependentes de ligante
LIGAND	Goto *et al.* 2002; Vert 2002	Interação vias metabólicas proteína-proteína de S. cerevisiae
MEPD	Henrich *et al.* 2003	Expressão genica
MEROPS	Rawlings e Barret 1999	Manualmente curada de famílias proteicas - Peptidases
MGI Mouse	Blake *et al.*2003	informações sobre camundongo
MIPS	Mewes *et al.* 2002	interação física, genética e complexos proteína-proteína de 10 mamíferos
MultiFun	Serres e Riley 2000	Sistema de classificação para função celular de *Escherichia coli* K-12
NuclearRDB	Horn *et al.* 2001	Receptores nucleares
PathCASE	Krishnamurthy *et al.* 2003	Sistema de base de dados de vias metabólicas
PDB	Berman *et al.* 2000	Estrutura 3D de proteínas obtidas experimentalmente (por experimentos como cristalografia de raio-X, microscopia eletrônica e ressonância magnética nuclear). Apresenta ferramentas de análise de estruturas

Continua.

SIGLA	CITAÇÃO	DESCRIÇÃO
PhylProM	Thoren 2000	Perfis filogenéticos
PIM	Rain *et al.* 2001	Interação física proteína-proteína de H. pylori
PIR	Wu *et al.* 2003	Sequências proteicas
PRINTS	Attwood*et al.* 2003	Sequências de motivos
Predictome	Mellor *et al.* 2002	Associações funcionais preditas entre proteínas
PROCAT	Wallace *et al.* 1997	Construída a partir de modelos 3D de sítios ativos extraídos utilizando o algoritmo TESS
PRODOM	Sonnhammer e Kahn 1994	Registros de domínios proteicos gerados automaticamente
ProDom	Servant *et al.* 2002	Domínios proteicos
ProKnow	Pal e Eisenberg2005	Integra informação da sequência (sequência e motivos) e estrutura (“folds” e motivos 3D) para predizer a função proteica usando uma abordagem probabilística, apresentando ligação com DIP através de categorias do GO
PROSITE	Hulo *et al.* 2006	Manualmente curada de motivos
SBASE	Vlahovicek *et al.* 2002	Domínios proteicos construídos a partir do algoritmo nearestneighbor e support vector machines
SCOP	Andreeva *et al.* 2004	Estrutura proteicas organizadas de forma hierárquica, mantendo relações evolutivas, na forma de famílias, superfamílias e “fold”
SGD	Dwight *et al.* 2002	Sequências proteicas de *S. cerevisiae*
SMD	Sherlock *et al.* 2001	Expressão genica
SMoS	Chakrabarti*et al.* 2003	Estrutura 3D de motivos
SURFACE	Ferre *et al.* 2004	Construída a partir partes da superfície extraídas pelo algoritmo SURFNET, classificados em códigos do GO usando PROSITE e foi considerada a similaridade estrutural e de resíduos para gerar a medida de RMSD e a matriz PAM
Swiss-Prot	Boeckmann *et al.* 2003	Manualmente curada contendo informação como anotação funcional, sequência de aminoácidos e palavras-chave
TAIR	Huala *et al.* 2001	Sequências proteicas de *A. thaliana*
TCDB	Saier Jr 2000	Proteínas de transporte de membrana
TIGRFAMs	Haft *et al.* 2001	Cadeias de Markov de alinhamento multiplo de sequências.
TrEMBL	Boeckmann*et al.* 2004	Curada automaticamente e suplementar ao Swiss-prot, contendo a tradução de todos as sequências de nucleotídeos presentes nas bases EMBL/GenBank/DDBJ e classificação e anotação automática
TubercuList	Camus *et al.* 2002	Sequências proteicas de *M. tuberculosis*
WormBase	Harris *et al.* 2004	Sequências proteicas de *C. elegans*
yMGV	Lelandais *et al.* 2004	Expressão genica
YPD	Costanzo *et al.* 2000	Interação física proteína-proteína de *S. cerevisiae* classificada em 3 categorias (localização subcelular, papel celular e função bioquímica)

APÊNDICE 3 – NÚMERO DE BASES DE DADOS POR SUBCATEGORIA DO NAR. FONTE: NAR (2013).

CATEGORIA	SUBCATEGORIA	TOTAL
Biologia celular	Biologia celular	6
Genomica	Genoma de fungos	37
	General genomics databases	45
	Temos de anotação, ontologias e nomenclaturas	22
	Genomas de não-vertebrados	2
	Genomas de invertebrados	61
	Genomas de procariotos	83
	Taxonomy and identification	11
	Unicellular eukaryotes genome databases	22
	Viral genome databases	35
Genomas humanos e de outros vertebrados	Genoma humano, mapas e vizualizadores	15
	ORFs humanas	27
	Organismos modelos e genomas comparativos	79
Genes e doenças humanas	Genes do cancer	37
	Especificos de genes, sistemas ou doenças	67
	Genética humana de propósito geral	20
	Polimorfismo	43
	Genes e doenças humanas	2
Imonologia	Imunologia	31
Vias metabólicas e de sinalização	Enzimas e nomenclatura de enzimas	17
	Vias metabólicas e de sinalização	1
	Vias metabólicas	37
	Interação entre proteínas	98
	Vias de sinalização	12
Dados de Expressão Genica	Dados de Expressão Genica	77
Sequências nucleotídicas	DNA codificante e não-codificante	49
	Estrutura de genes, introns, exons e sítios de splice	28
	Base de dados com colaboração internacional	9
	Sítios reguladores de transcrição e fatores de transcrição	76
Organelas	Genes e proteínas mitocondriais	19
	Organelas	8
Outras bases de dados de biologia molecular	Drogas e modelagem de drogas	35
	Primes e probes	11
	Outras de biologia molecular	18
Plantas	Arabidopsis thaliana	28
	Bases de dados de plantas com propósito geral	53
	Outras plantas	21
	Plantas	1
	Arroz	20
Sequências proteicas	Família proteica	92
	Bases de dados genéricas de sequências proteicas	16
	Domínios e classificação de proteínas	41
	Localização de proteínas e alvo de ação da proteína	25
	Propriedades de proteínas	22
	Sequências de motivos e sítios ativos	32

Continua.

CATEGORIA	SUBCATEGORIA	TOTAL
Recursos proteômicos	Recursos proteômicos	20
Sequências de RNA	Sequências de RNA	88
Estrutura	Carboidratos	12
	Estrutura de ácido nucleica	22
	Estrutura de proteínas	114
	Moléculas pequenas	23
	Total	1770

APÊNDICE 4 - TRECHOS DOS ARQUIVOS DOS CONJUNTOS UTILIZADOS.

Trecho do arquivo do material suplementar do trabalho de Brown *et al.* (2007):

```
ADA1A_BOVIN     N630 AlphaAdrenoceptorstype1A
ADA1A_CAVPO     N630 AlphaAdrenoceptorstype1A
ADA1A_HUMAN     N630 AlphaAdrenoceptorstype1A
ADA1A_MOUSE     N630 AlphaAdrenoceptorstype1A
ADA1A_ORYLA     N630 AlphaAdrenoceptorstype1A
```

Trecho do arquivo .doc do material suplementar do trabalho de Dobson e Doig (2003):

Dobson, P. D.; Doig, A. J. (2003) Distinguishing enzyme structures from non-enzymes without alignments. v. 330, n. 4, p. 771–783.

Supplementary Materials

1178 proteins in the dataset. Note that the split into enzymes and non-enzymes is only as accurate as the annotations in the databases (PDB, Medline abstracts) and as such may contain errors.

691 Enzymes

11AS 1A26 1A2J 1A2P 1A33 1A49 1A4M 1A4S 1A59 1A5V 1A5Z 1A69 1A77 1A7U 1A82 1A8D 1A8P 1A8R 1A95 1A9U 1A9X 1AA6 1AA8 1ABO 1AD4 1ADE 1ADO 1AE1 1AFW 1AGJ 1AI4 1AJ5 1AJA 1AK2 1AL8 1ALN 1AQ2 1AQY 1ARC 1AUR 1AUW 1AV4 1AW9 1AY5 1AYD

[...]

Trecho do arquivo .tsv do SFLD (2012) que contém 28 colunas ("Enzyme Functional Domain ID", "Name", "Superfamily", "Superfamily Name", "Superfamily Evidence

Code", "Subgroup", "Subgroup Name", "Family", "Family Name", "Family Evidence Code", "Taxonomy Ids", "Species Name", "Type of Life", "Reaction ID", "Reaction Name", "GI", "Genbank Accession Number", "RefSeq Accession Number", "Uniprot", "SwissProt", "TrEMBL", "MicrobesOnline", "The SEED", "EFI Target ID", "DES Target ID", "Length domain", "Length full" e "PDB ID"):

Enzyme	Functional Domain ID	Name	Superfamily	Superfamily Nam
1	OSBS / NSAR 1	enolase	ISS 180	muconate cyclo
2	OSBS / NSAR 1	enolase	ISS 180	muconate cyclo
3	o-succinylbenzoate synthase 1	enolase	ISS 180	
4	o-succinylbenzoate synthase 1	enolase	ISS 180	
5	o-succinylbenzoate synthase 1	enolase	ISS 180	

[...] [...]

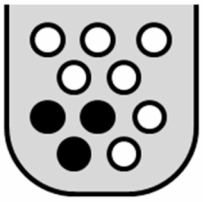
Trecho do arquivo .xls do material suplementar do trabalho de Brown *et al.* (2006):

GI Number[1]	Gold Std. Family[3]	Gold Std. SF
9837270	1-aminocyclopropane-1-carboxylate deaminase	amidohydrolase
16078516	adenine deaminase	amidohydrolase
1518868	adenosine deaminase	amidohydrolase
3892028	adenosine deaminase	amidohydrolase
4557249	adenosine deaminase	amidohydrolase

[...]

APÊNDICE 5 – EXPERIMENTO COM RESPOSTA DICOTÔMICA. FONTE: Adaptado de Melo (2007)[8].

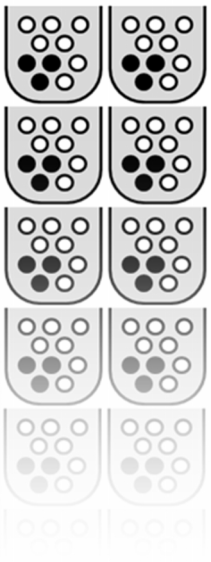
MODELO OU DISTRIBUIÇÃO DE BERNOULLI



- Em um experimento (ensaio) de Bernoulli tem-se associada a ocorrência ou não de um evento desejado
- O sucesso (ocorrência do evento desejado) ou fracasso (não ocorre o evento desejado) é representado por uma variável aleatória que pode assumir o valor 1 (sucesso) ou 0 (fracasso):
- Essa variável aleatória apresenta a distribuição de Bernoulli, e também pode ser chamada de variável aleatória de Bernoulli.
- Exemplo de experimento de Bernoulli:
  - Sorteio de um paciente de um conjunto de pacientes que apresentam ou não um sintoma.
  - Sorteio de uma bolinha em uma urna que contêm bolinhas brancas e pretas.

$$X = \begin{cases} 0, \text{fracasso} \\ 1, \text{sucesso}, \end{cases} \text{onde } X \text{ é uma variável aleatória de Bernoulli}$$

MODELO OU DISTRIBUIÇÃO DE BINOMIAL



- Repetições independentes do experimento de Bernoulli
- O número de vezes que se tem sucesso pode ser de zero até o número de vezes que o experimento foi repetido
- Esse número de vezes que se teve de sucesso em relação a todas as repetições do experimento de Bernoulli é a variável de distribuição binomial
- Dessa forma as características de variável binomial são:
  - Ser resultado de uma contagem
  - Os experimentos que a determina devem ser independentes
  - A probabilidade de sucesso é a mesma a cada repetição do experimento
- Exemplo de repetições independentes do experimento de Bernoulli:
  - 20 sorteios de um paciente de um conjunto de pacientes que apresentam ou não um sintoma. O número total de sucessos possíveis para esses 20 sorteios é dado por uma variável com distribuição binomial
  - 15 sorteios de uma bolinha em uma urna que contêm bolinhas brancas e pretas. O número total de sucessos possíveis para esses 15 sorteios é dado por uma variável com distribuição binomial

$$Y \sim B(n,p), \text{ onde } \begin{cases} n = \text{nº total de repetições do experimento} \\ p = \text{probabilidade de sucessos do experimento}, \end{cases}$$

onde Y é uma variável determinada por uma função de distribuição binomial do número total de repetições do experimento de Bernoulli e da probabilidade de sucesso desse experimento

[8] Melo, A. C. Estatística de probabilidade para ciências exatas (notas de aula). INE – Departamento de informática e de estatística da Universidade Federal de Santa Catarina. 2007. Disponível em: <http://www.inf.ufsc.br/~anaclaudia/ine5108/notas_aula/texto_Bernoulli_Bin.pdf>. Acesso em: 14/03/2013.

APÊNDICE 6 – TRECHO DOS ARQUIVOS DE SAÍDA DOS PROGRAMAS UTILIZADOS.

Trecho do arquivo de saída do Blast2GO com destaque para a informação de interesse:

```
ADA1A_BOVIN alpha-1a adrenergic receptor          271 20  0.0 92.45%  31  P:activation of p
ADA1A_CAVPO alpha-1a adrenergic receptor          271 20  6.8672E-173 92.15%  31  P:activat
ADA1A_HUMAN alpha-1a adrenergic receptor isoform 2 271 20  0.0 94.0%   31  P:activat
ADA1A_MOUSE alpha-1a adrenergic receptor          271 20  0.0 92.3%   32  ; P:activation of
ADA1A_ORYLA alpha-1a adrenergic receptor-like     270 20  2.59616E-174    86.25%  31  P
```

Trecho do arquivo de saída do InterProScan com destaque para a informação de interesse:

```
Sequence "ADA1A_BOVIN" crc64 checksum: 7148B1ACB375D24E length: 271 aa.

InterPro        IPR000276       7TM GPCR, rhodopsin-like
Biological Process: G-protein coupled receptor protein signaling pathway (GO:0007186), Cellular Component:
method          AccNumber       shortName                               location
FPrintScan      PR00237         GPCRRHODOPSN                            T[19-40] 6.6e-45 T[64-86] 6.6e-45 T[1
HMMPfam         PF00001         7tm_1                                   T[1-271] 1.9e-82

InterPro        IPR002233       Adrenergic receptor
Molecular Function: adrenergic receptor activity (GO:0004935), Biological Process: G-protein coupled recept
method          AccNumber       shortName                               location
FPrintScan      PR01103         ADRENERGICR                             T[39-50] 1.5e-06 T[92-100] 1.5e-06 T|

InterPro        NULL            NULL
method          AccNumber       shortName                               location
HMMPanther      PTHR24248       FAMILY NOT NAMED                        T[1-271] 3e-214
HMMPanther      PTHR24248:SF89  SUBFAMILY NOT NAMED                     T[1-271] 3e-214

Sequence "ADA1A_CAVPO" crc64 checksum: D47757A479F5DB3D length: 271 aa.

InterPro        IPR000276       7TM GPCR, rhodopsin-like
Biological Process: G-protein coupled receptor protein signaling pathway (GO:0007186), Cellular Component:
method          AccNumber       shortName                               location
FPrintScan      PR00237         GPCRRHODOPSN                            T[19-40] 7.9e-45 T[64-86] 7.9e-45 T[1
```

Trecho do arquivo de saída do Panther Score com destaque para a informação de interesse:

```
ADA1A_BOVIN PTHR24248:SF89  SUBFAMILY NOT NAMED 3e-214  722.6   1-271,
HRH1_BOVIN  PTHR24249:SF29  SUBFAMILY NOT NAMED 1.1e-201    680.9   1-209,252-321,
ACM5_HUMAN  PTHR24249:SF36  SUBFAMILY NOT NAMED 1.3e-205    693.9   1-217,256-323,
5HT1B_HUMAN PTHR24247:SF67  SUBFAMILY NOT NAMED 8.1e-244    820.8   1-299,
ADRB1_FELCA PTHR24248:SF43  SUBFAMILY NOT NAMED 1.3e-213    720.5   1-296,
ADA1B_HUMAN PTHR24248:SF87  SUBFAMILY NOT NAMED 1.2e-222    750.5   1-272,
Q6TLJ0_MUSPF    PTHR24249:SF50  SUBFAMILY NOT NAMED 1.5e-179    607.3   1-262,
ADA1D_HUMAN PTHR24248:SF88  SUBFAMILY NOT NAMED 7.3e-227    764.5   1-273,
```

Trecho do arquivo de saída do Pfam scan com destaque para a informação de interesse:

```
ADA1A_BOVIN       1    271      1    271 PF00001.16  7tm_1          Family     1   257   257
ADA1A_CAVPO       1    271      1    271 PF00001.16  7tm_1          Family     1   257   257
ADA1A_HUMAN       1    271      1    271 PF00001.16  7tm_1          Family     1   257   257
ADA1A_MOUSE       1    271      1    271 PF00001.16  7tm_1          Family     1   257   257
ADA1A_ORYLA       1    270      1    270 PF00001.16  7tm_1          Family     1   257   257
ADA1A_RABIT       1    271      1    271 PF00001.16  7tm_1          Family     1   257   257
ADA1A_RAT         1    271      1    271 PF00001.16  7tm_1          Family     1   257   257
```

Trecho do arquivo de saída do ScanProsite com destaque para a informação de interesse:

```
>ADA1A_BOVIN : PS00237 G_PROTEIN_RECEP_F1_1 G-protein coupled receptors family 1 signature.
     70 - 86  ASImGLCIISIDRYIgV
>ADA1A_CAVPO : PS00237 G_PROTEIN_RECEP_F1_1 G-protein coupled receptors family 1 signature.
     70 - 86  ASImSLCIISIDRYIgV
>ADA1A_HUMAN : PS00237 G_PROTEIN_RECEP_F1_1 G-protein coupled receptors family 1 signature.
     70 - 86  ASImGLCIISIDRYIgV
>ADA1A_MOUSE : PS00237 G_PROTEIN_RECEP_F1_1 G-protein coupled receptors family 1 signature.
     70 - 86  ASImGLCIISIDRYIgV
>ADA1A_ORYLA : PS00237 G_PROTEIN_RECEP_F1_1 G-protein coupled receptors family 1 signature.
     70 - 86  ASImSLCVISVDRYIgV
```

APÊNDICE 7 – ETAPAS DE NORMALIZAÇÃO DOS NOMES EXTRAÍDOS DAS SAÍDAS DOS PROGRAMAS E DOS ARQUIVOS COM AS DENOMINAÇÕES CORRETAS PARA AS SEQUÊNCIAS.

1) Substitui (! " # $ % & ' ( ) * + , - . / : ; < = > ? @ [\ ] ^ _ ` { | } ~) por " " através de um vetor de expressões regulares
2) Passa número romano para arábico
3) Separa caracteres maiúsculos precedidos de minúsculos
4) Separa caracteres maiúsculos precedidos de minúsculos
5) Passa para caixa baixa
6) Separa caracteres precedidos de dígitos
7) Separa dígitos precedidos de caracteres
8) Retira palavra sem relevância para a comparação (com base em Tsuruoka *et al.*, 2008) ('protein', 'gene', 'type', 'superfamily', 'subfamily', 'family', 'like', 'domain', 'region','isoform ','variant','specific','precursor')
9) Retira "s" do final das palavras (retira plural apenas de palavras no meio do nome)

10) Retira "s" do final do nome (retira plural apenas de palavras no fim do nome)
11) Passa de nome de caractere grego para latino
12) Passa de caractere grego para latino
13) Retira espaços múltiplos
14) Retira espaços do final
15) Retira espaços do começo
16) Ordena os nomes (retira o efeito de posição diferente)

APÊNDICE 8 – ALGORITMO EM LINGUAGEM SHELL SCRIPT DO FLUXO DE DADOS.

```
#!/bin/bash

### Workflow que executa em série vários programas de predição de função de
proteínas
### Recebe 6 argumentos, todos obrigatórios:
# $1 = n. de conjuntos (inteiro)
# $2 = caminho para os conjuntos de sequências em fasta (caminho para a
pasta que contêm as sequências)
# $3 = caminho para as saídas (caminho para a pasta que receberá os
resultados) e onde se encontram as tabelas com a classificação padrão
# $4 = caminho para os programas que geram tabelas (classificação, de tempo
de execução e de n. de bases)
# $5 = caminho para script em R (que realiza as análises)
# $6 = caminho para saída do script em R#
#
### Exemplo de linha de comando para sua execução:
# ./exuta_programas.sh 12 /home/usr/proj/sequencias
/home/usr/proj/saidas_programas /home/usr/proj/processamento
/home/usr/proj/analiseProgramas.R /home/usr/proj/saidaR

echo
"_____________________________________________________________________________"
echo "Inicio da execucao do worflow:" $(date)
echo
"_____________________________________________________________________________"
echo ""

export PERL5LIB=/home/usr/progGrupoGenes/PANTHER/pantherScore1.03/lib

### execucao dos programas
i=0
while test $i != $1
do
      echo "Inicio blastp:" $(date)
      blastp -query $2/seq000$i.fasta -db /media/arquivos/dados-
progGrupoGenes/NR/nr -out  $3/blastp000$i.xml -outfmt 5 -evalue 0.001 -
num_alignments 20 -threshold 33 -num_threads 4
      echo "Termino blastp:" $(date)
      echo ""
      echo "Inicio blast2go:" $(date)
      java -Xms1024m -Xmx2048m -cp
/home/bnr/progGrupoGenes/b2g4pipe/*:/home/bnr/progGrupoGenes/b2g4pipe/ext/*
: es.blast2go.prog.B2GAnnotPipe -in  $3/blastp000$i.xml -out $3/b2g000$i -
prop /home/bnr/progGrupoGenes/b2g4pipe/b2gPipe.properties -v -annot
```

```
	echo "Termino blast2go:" $(date)
	echo ""
	echo "Inicio PfamScan:" $(date)
	cd /media/arquivos/PfamScan
	./pfam_scan.pl -fasta $2/seq000$i.fasta -dir /media/arquivos/dados-
progGrupoGenes/pfam_dados > $3/pfam000$i.txt
	cd
	echo "Termino PfamScan:" $(date)
	echo ""
	echo "Inicio iprscan:" $(date)
	/media/arquivos/iprscan/bin/iprscan -cli -iprlookup -goterms -i
$2/seq000$i.fasta -o $3/ipr000$i.xml -format xml
	echo "Termino iprscan:" $(date)
	echo ""
	echo "Inicio prosite:" $(date)
	/media/arquivos/ps_scan/ps_scan.pl -s $2/seq000$i.fasta -d
/media/arquivos/ps_scan/prosite.dat > $3/pst000$i.txt
	echo "Termino prosite:" $(date)
	echo ""
	echo "Inicio PANTHER:" $(date)
	/home/bnr/progGrupoGenes/PANTHER/pantherScore1.03/pantherScore.pl -l
/media/arquivos/dados-progGrupoGenes/PANTHER7.0 -D B -V -i
$2/seq000$i.fasta -o $3/panther000$i.txt -n -H /usr/bin/hmmsearch
	echo "Termino PANTHER:" $(date)
	echo ""
	i=`expr $i + 1`
done

### gerar tabela de nomes
i=0
aux=""
while test $i != $1
do
	$4/geratbl_b2g $3/b2g000$i.annot > $3/tableb2g000$i.txt
	$4/geratbl_ipr $3/ipr000$i.xml > $3/tableipr000$i.txt
	$4/geratbl_panther $3/panther000$i.txt > $3/tablepanther000$i.txt
	$4/geratbl_pfam $3/pfam000$i.txt
/media/arquivos/PfamScan/pdb_pfam_mapping.txt > $3/tablepfam000$i.txt
	$4/geratbl_pst $3/pst000$i.txt > $3/tablepst000$i.txt

	aux="$aux"$($4/geratbl_numBases
$2/seq000$i.fasta)"\t$3/tableb2g000$i.txt,$3/tableipr000$i.txt,$3/tablepant
her000$i.txt,$3/tablepfam000$i.txt,$3/tablepst000$i.txt,$3/tablepadrao000$i
.txt\n"
	i=`expr $i + 1`
done
echo -e $aux > $3/caminhos_tbls.txt

echo "Análise em R"
echo "Inicio:" $(date)
Rscript --vanilla $5 $6 $4/caminhos_tbls.txt > $6/controleDeExecucaoR.txt
echo "Termino:" $(date)

echo ""
echo
"_____________________________________________________________________________"
echo "Termino da execucao do worflow:" $(date)
echo
"_____________________________________________________________________________"
```

APÊNDICE 9 – EXECUÇÃO DO WORKFLOW UTILIZANDO A LINGUAGEM SWIFT E UTILIZANDO A LINGUAGEM SHELL SCRIPT.

A linguagem Swift também apresenta as características linguagem Shell script, embora apresentasse mais linhas de código. Porém a linguagem Shell permitiria a execução do workflow em paralelo.

Ao executar os programas com a opções próprias de execução em paralelo, a implementação do programa InterProScan não permitia que o Swift detectasse que ele ainda estava em execução.

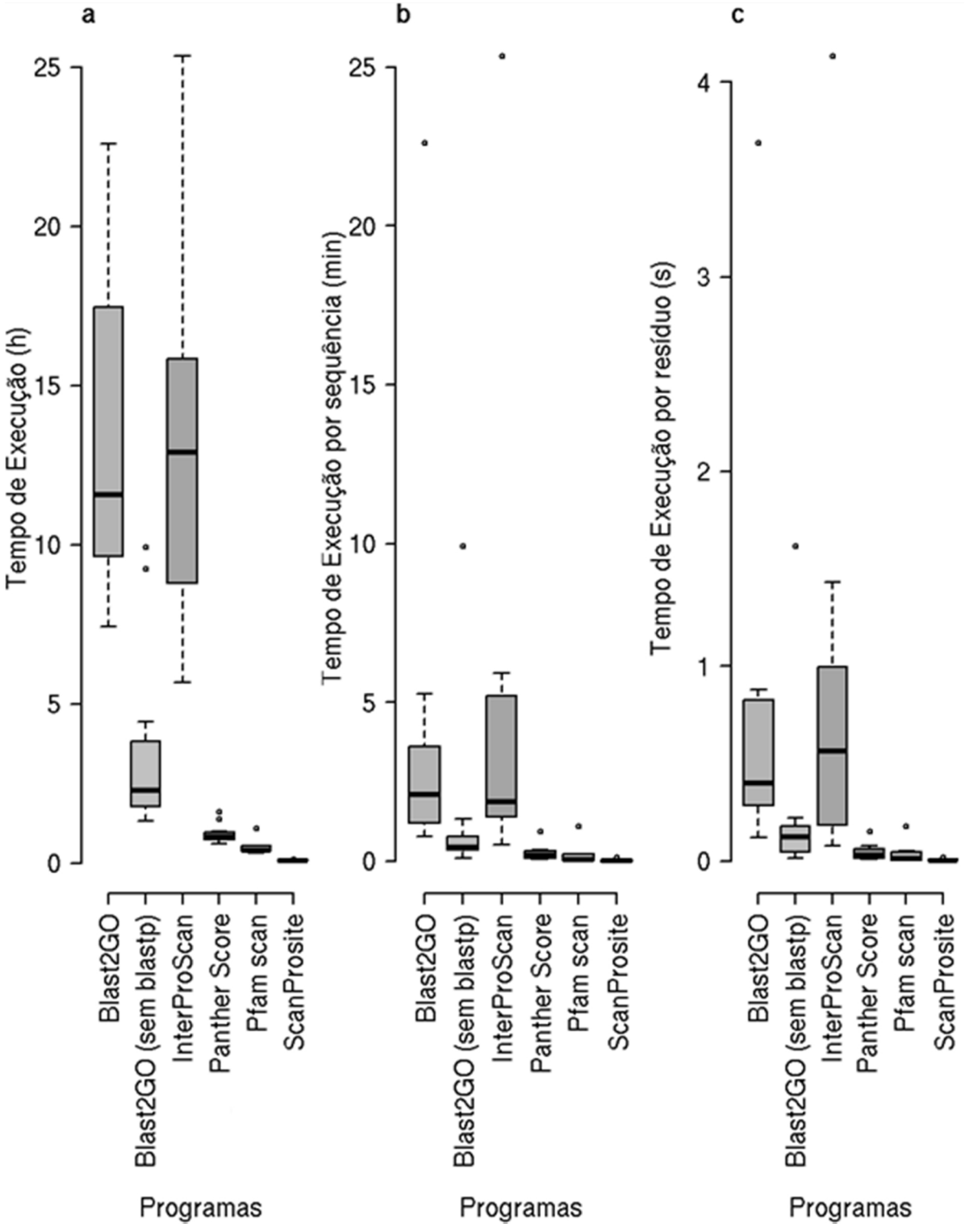
Dessa forma, era requisitado o uso de um espaço na memória RAM que ainda estava sendo ocupado, interrompendo a execução do workflow. Quando se executava sem a paralelização, o tempo de execução total do workflow era próximo ao tempo total de execução do workflow em Shell script com apenas paralelização oferecida por cada programa:

Tempo de execução total do workflow para 2160 sequências proteicas

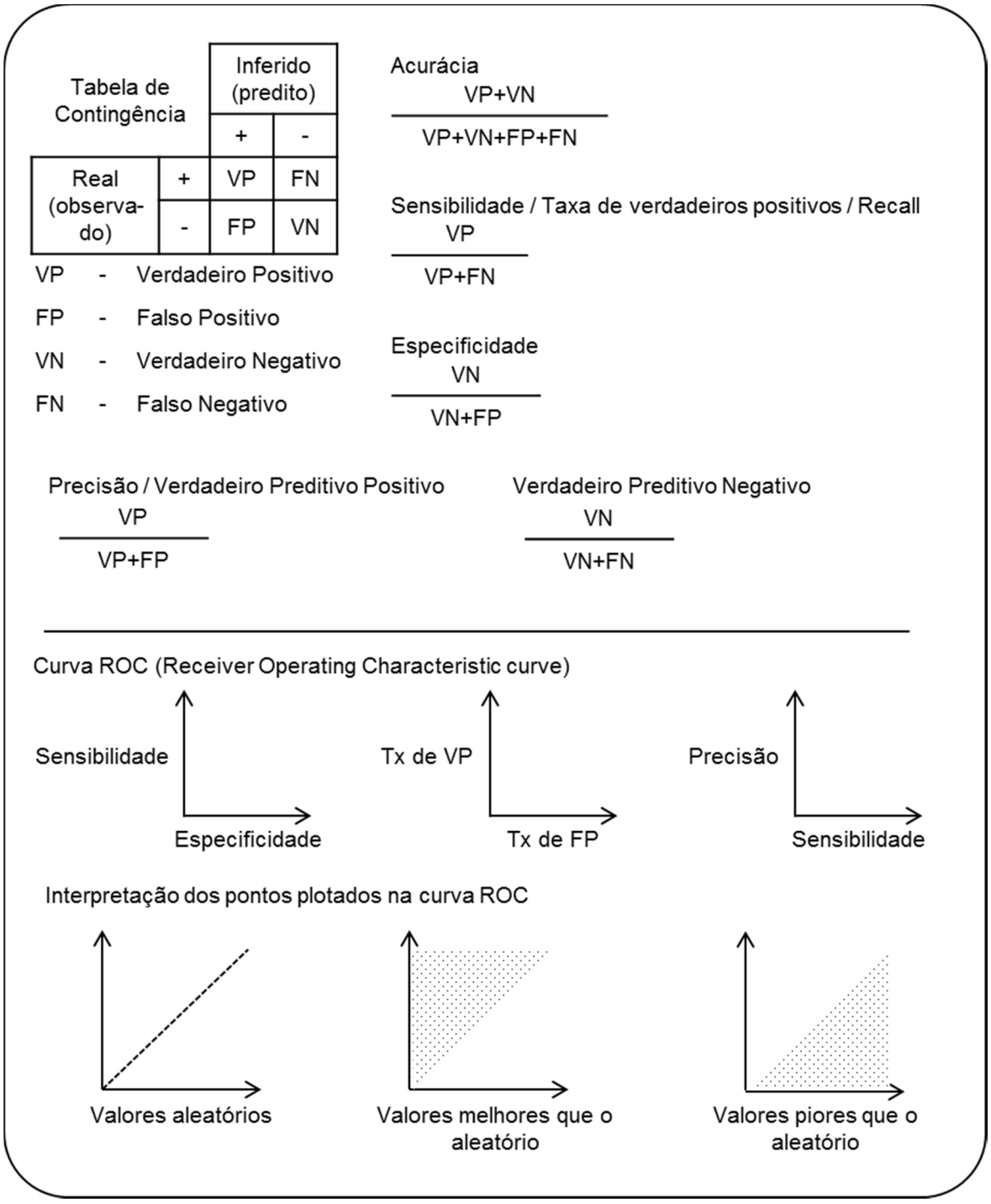
LINGUAGEM	TEMPO DE EXECUÇÃO (hh:mm:ss)
Shell script*	19:01:34
Swift**	18:04:04

* Execução com a opção de paralelização oferecida pelos programas
** Execução sem a paralelização oferecida pelos programas

APÊNDICE 10 – GRÁFICOS BOXPLOT DO TEMPO DE EXECUÇÃO COM OUTLINE. Gráficos boxplot representando o tempo de execução total (a), por sequência (b) e por aminoácido (c).



APÊNDICE 11 – TABELAS DE CONTINGÊNCIA E CURVA ROC. Descrição simplificada de construção e análise empregando tabelas de contingência e curva ROC. FONTE: Adaptado de Prati, Batista e Monard (2008) e Estatcamp (2012)[9].



[9] Estatcamp. Portal Action. São Carlos, 1997-2012. Disponível em: http://www.portalaction.com.br/989-predicao

ANEXOS

ANEXO 1 – EXEMPLO DE CLASSIFICAÇÃO DO EC. FONTE: http://www.chem.qmul.ac.uk/iubmb/enzyme/

EC 2.TRANSFERASES

EC 2.1 **Transferring One-Carbon Groups**

EC 2.1.1 Methyltransferases

EC 2.1.2 Hydroxymethyl-, Formyl- and Related Transferases

EC 2.1.3 Carboxy- and Carbamoyltransferases

EC 2.1.4 Amidinotransferases

EC 2.2 **Transferring Aldehyde or Ketonic Groups**

EC 2.2.1 Transketolases and Transaldolases

EC 2.3 **Acyltransferases**

EC 2.3.1 Transferring groups other than amino-acyl groups

EC 2.3.1 Transferring groups other than amino-acyl groups

EC 2.3.2 Aminoacyltransferases

EC 2.3.3 Acyl groups converted into alkyl on transfer

EC 2.4 **Glycosyltransferases**

EC 2.4.1 Hexosyltransferases

EC 2.4.1.1 phosphorylase

EC 2.4.1.2 dextrin dextranase

EC 2.4.1.3 deleted, included in EC 2.4.1.25

EC 2.4.1.4 amylosucrase

EC 2.4.1.5 dextransucrase

EC 2.4.1.6 deleted

EC 2.4.2 Pentosyltransferases

EC 2.4.99 Transferring other glycosyl groups

ANEXO 2 – EXEMPLO DE CLASSIFICAÇÃO DO FUNCAT. FONTE: http://mips.helmholtz-muenchen.de/proj/funcatDB/

MIPS **Fun**ctional **Cat**alogue

01 METABOLISM

02 ENERGY

04 STORAGE PROTEIN

10 CELL CYCLE AND DNA PROCESSING

11 TRANSCRIPTION

12 PROTEIN SYNTHESIS

14 PROTEIN FATE (folding, modification, destination)

16 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic)

18 REGULATION OF METABOLISM AND PROTEIN FUNCTION

20 CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES

20.01 transported compounds (substrates)

20.01.01 ion transport

20.01.03 C-compound and carbohydrate transport

20.01.03.01 sugar transport

20.01.03.03 C4-dicarboxylate transport (e.g. malate, succinate, fumarate)

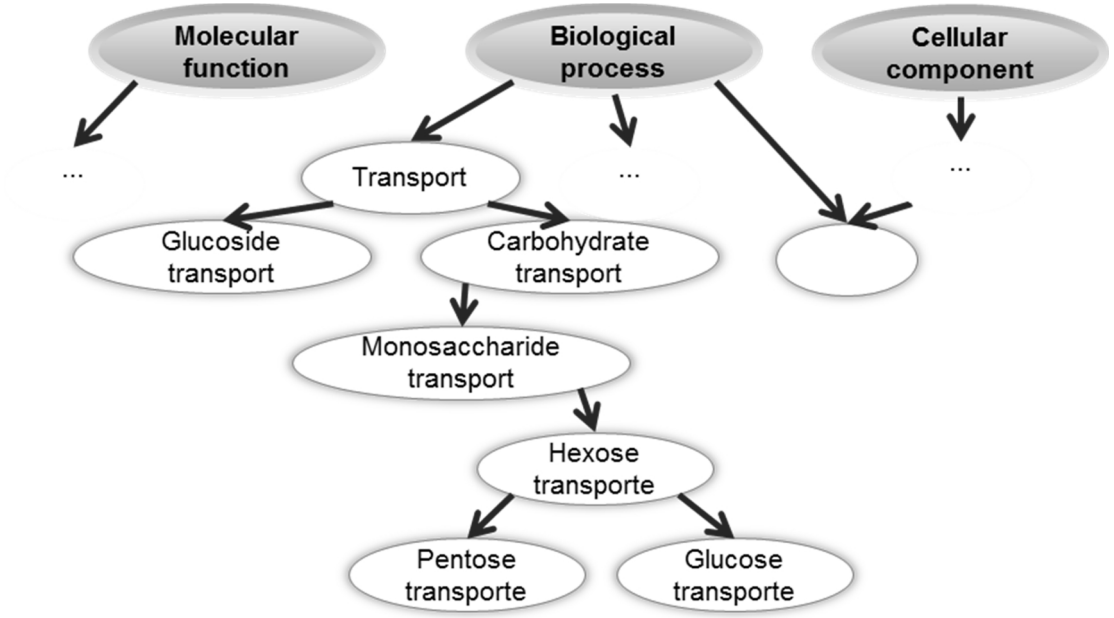
20.01.07 amino acid/amino acid derivatives transport

20.01.09 peptide transport

20.01.10 protein transport

20.01.11 amine / polyamine transport

ANEXO 3 - EXEMPLO DE CLASSIFICAÇÃO DO GO. FONTE: Chitale e Kihara (2011).



ANEXO 4 - ESCOPO DO GO. O GO realiza a caracterização de produtos gênicos por seus atributos, seguindo um contexto celular. Dessa forma, a nomenclatura de proteínas não é abrangida pelo GO. FONTE: GO (2012).

www.geneontology.org/GO.doc.shtml#not

The Scope of GO

It is important to clearly state the scope of GO, and what it does and does not cover. The preceding section explained the domains covered by GO; the following areas are outside the scope of GO, and terms in these domains would not appear in the ontologies.

Gene products: e.g. cytochrome c is not in the ontologies, but attributes of cytochrome c, such as oxidoreductase activity, are.

Processes, functions or components that are unique to mutants or diseases: e.g. **oncogenesis** is not a valid GO term because causing cancer is not the normal function of any gene.

Attributes of sequence such as intron/exon parameters: these are not attributes of gene products and will be described in a separate sequence ontology (see the OBO website for more information).

Protein domains or structural features.

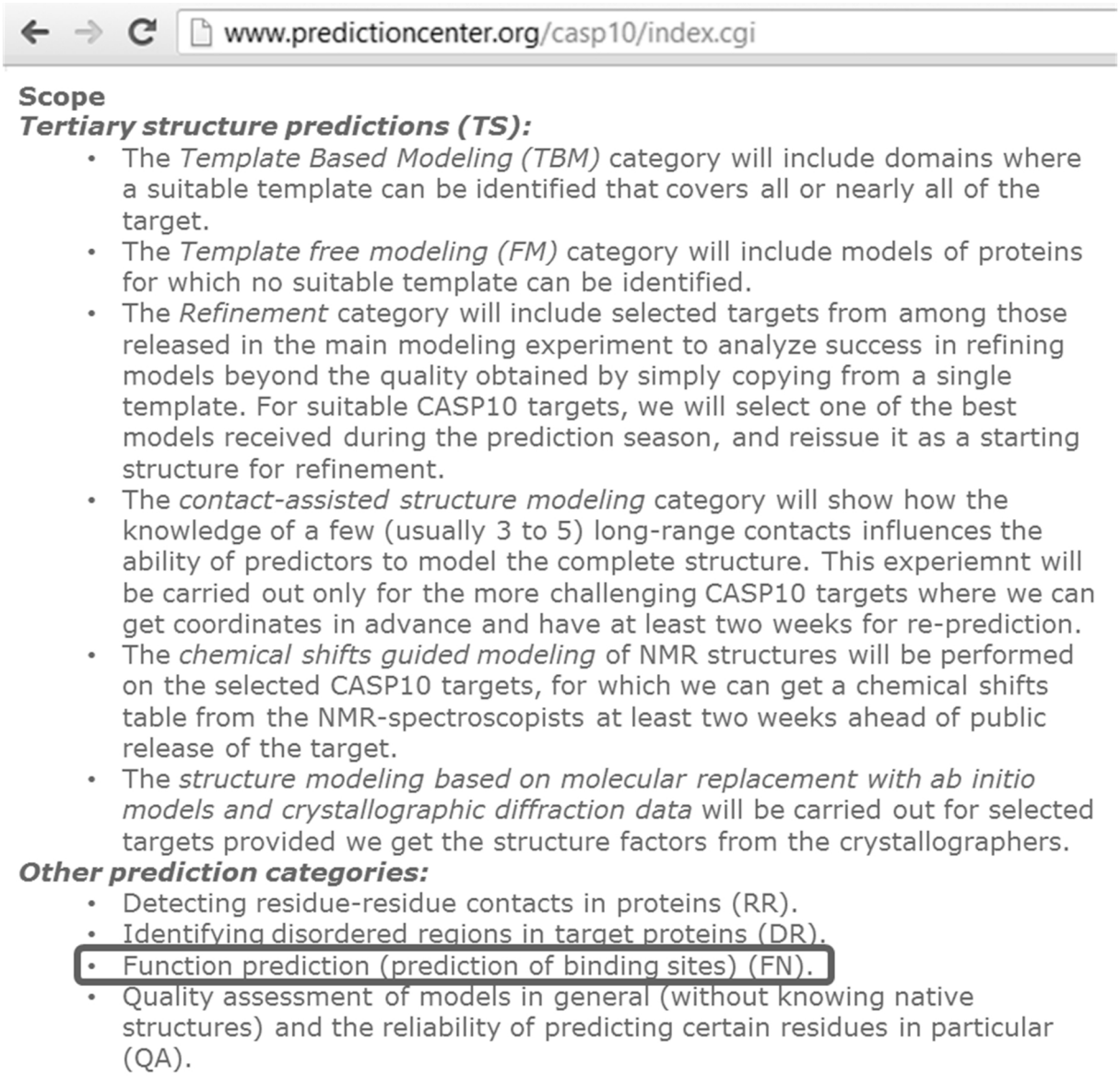
Protein-protein interactions.

Environment, evolution and expression.

Anatomical or histological features above the level of cellular components, including cell types.

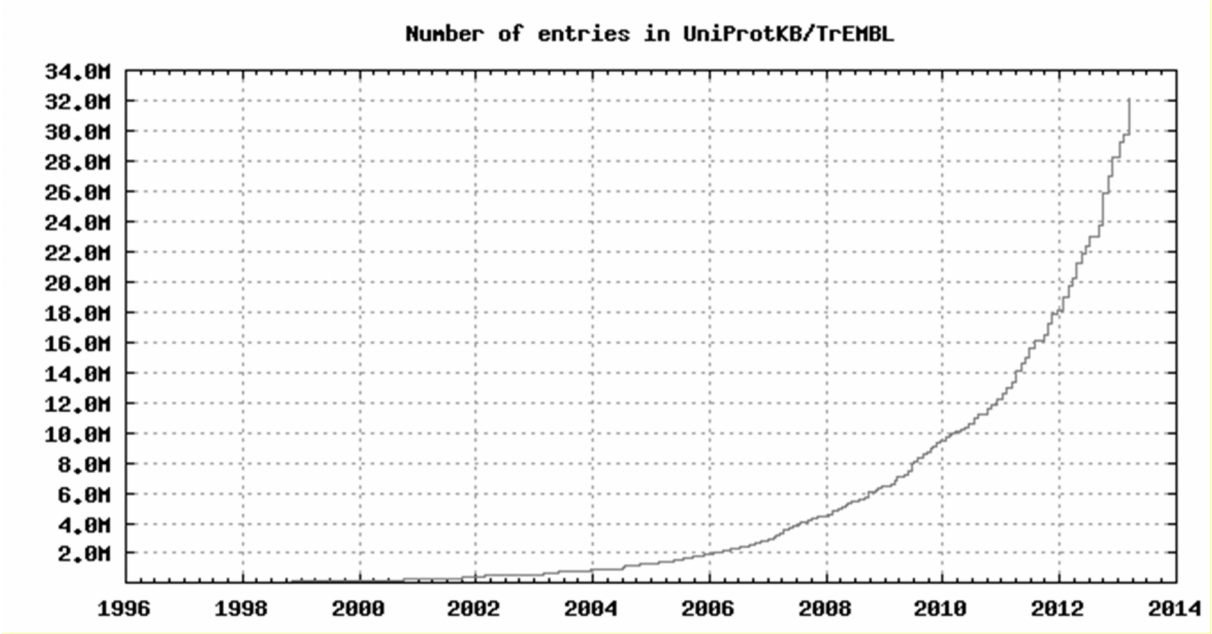
GO is not a database of gene sequences, nor a catalog of gene products. Rather, GO describes how gene products behave in a cellular context.

ANEXO 5 – ESCOPO DO CASP (Critical Assessment of protein Structure Prediction). FONTE: Protein Structure Prediction Center (2012) [10].

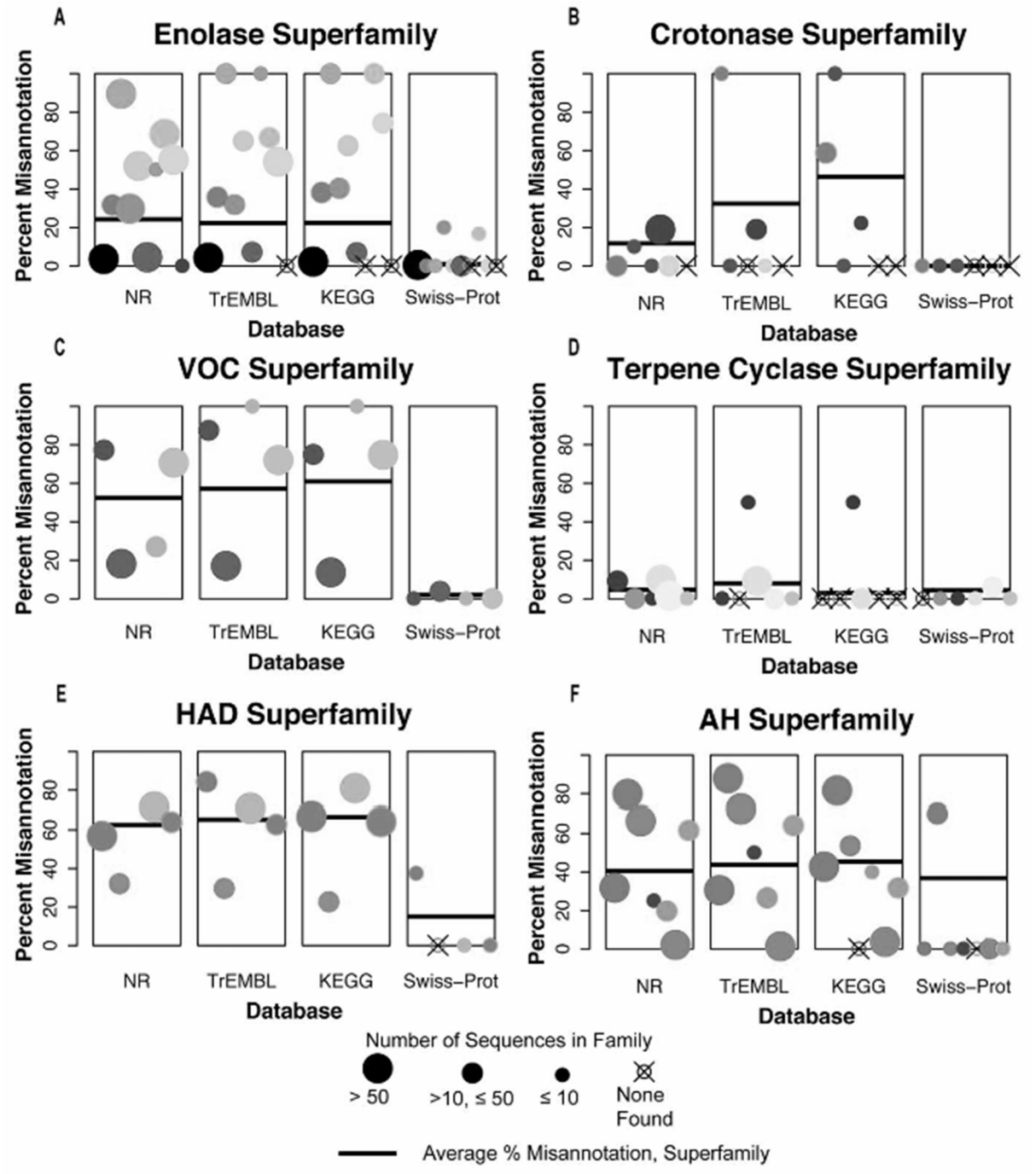


[10] Protein Structure Prediction Center. US National Institute of General Medical Sciences (NIH/NIGMS). University of California, Davis. 2007-2012. Disponível em: <http://www.predictioncenter.org/casp10/index.cgi>.

ANEXO 6 – NÚMERO DE SEQUÊNCIAS DEPOSITADAS NA BASE DE DADOS UNIPROT DE 1996 ATÉ 2013. FONTE: UNIPROT (2013).



ANEXO 7 – PORCENTAGEM DE ERRO DE ANOTAÇÃO EM FAMÍLIAS E SUPERFAMÍLIAS TESTADAS NO TRABALHO DE SCHNOES *et al.* (2009). Os resultados estão organizados por superfamília: A: enolase, B: crotonase, C: vicinal oxigen chelate, D: terpene cyclase, E: haloacid dehalogenase e F: amidohidrolase. Cada circulo representa uma família e o tamanho do circulo se refere ao número de sequências da família. A barra horizontal se refere à média de erro de anotação. Notar a baixa porcentagem de erros de anotação encontrada na base de dados Swiss-Prot para a maioria das superfamílias.



ANEXO 8 – SIMILARIDADE SEMÂNTICA ENTRE DOIS TERMOS DO GO UTILIZANDO A METODOLOGIA DE LORD *et al.* (2003).

Cada termo recebe um valor de p equivalente a sua frequência no UniProt. Isso leva o valor do termo ancestral apresentar um valor de p igual a soma dos valores de p dos termos filhos, em que o termo raiz apresenta p = 1. Assim os termos His-NH3 lyase e Ser-NH3 lyase estão mais próximos entre si que os termos His-NH3 lyase e C-S lyase. O calculo da pontuação de similaridade é realizada pelo logaritmo natural negativo da menor soma de p entre dois termos, ou seja: sim(c1, c2)= -ln(Pms(c1, c2)), onde c1 e c2 são os dois termos, Pms(c1, c2) é soma da probabilidade mínima entre esses termos e sim(c1, c2) é a pontuação de similaridade. Extraído de Friedberg (2006).