UTILIZAÇÃO DE REDES NEURAIS ARTIFICIAIS COMBINADAS COM ALGORITMOS DE EXTRAÇÃO DE CARACTERÍSITCAS COMO CLASSIFICADORES DE FAMÍLIAS DE PROTEÍNAS

FREDERICO DE MENEZES*+, ANDRÉ NASCIMENTO*++, EDUARDO ANDRADE*, ADRIANO OLIVEIRA*

** Departamento de Sistemas Computacionais – UPE, Recife, PE, Brasil*

+ Departamento de Química Fundamental – UFPE, Recife, PE, Brasil

++Centro de Informática – UFPE, Recife, PE, Brasil

E-mails: fredquim22@yahoo.com.br, andrecamara@gmail.com, ezoby@yahoo.com.br, adriano@dsc.upe.br

Abstract— The continuous and exponential growth of DNA and protein databases has generated a need for advanced computational tools to analyze these data and to find solutions to problems like protein family classification. This paper describes the use of a Multi-layer perceptron (MLP) Artificial Neural Net (ANN) combined with six data extraction algorithms to classification of proteins. The algorithms were implemented in JAVA and the neural net was simulated by tool WEKA. The results were analyzed considering the accuracy rate and the area under the ROC curve. The results obtained in the experiments using ANNs for the protein classification problem, demonstrated reasonably good results, considering the simple algorithms implementations for information pre-processing.

Keywords— Artificial Neural Networks, protein family classification, feature extraction, bioinformatics.

Resumo— O crescimento contínuo e exponencial de bases de dados sobre DNA e proteínas tem gerado uma necessidade de ferramentas computacionais avançadas para a análise destes dados e resolução de problemas como classificação de famílias de proteínas. Este artigo descreve o uso de uma Rede Neural Artificial (RNA) do tipo Multi-layer perceptron (MLP) combinada com seis algoritmos de extração de informações para a classificação de proteínas. Os algoritmos foram implementados em JAVA e a rede neural foi simulada na ferramenta WEKA. Os resultados foram analisados considerando-se a taxa de acerto das classificações e a área sob a curva ROC. Os resultados obtidos nos experimentos utilizando RNAs para o problema de classificação de proteínas demonstraram resultados bastante satisfatórios, considerando-se a simplicidade dos algoritmos implementados para o pré-processamento das informações.

Palavras-chave— Redes Neurais Artificiais, classificação de famílias de proteínas, extração de características, bioinformática.

1 Introdução

O crescimento contínuo e exponencial de bases de dados que armazenam informações sobre o seqüenciamento de DNA e proteínas de diversos organismos tem gerado uma grande necessidade de novos recursos computacionais para a análise e gerenciamento das informações contidas nestas bases de dados [Rost (1996), Wu (1992), Baldi (1998)]. Novas ferramentas capazes de extrair informações relevantes sobre proteínas de interesse científico, ou capazes de realizar a predição de propriedades físico-químicas de novas proteínas, baseando-se em informações disponíveis em diferentes bases de dados, vem sendo desenvolvidas por diferentes grupos de pesquisa no mundo.

Para um número considerável de proteínas, o conhecimento básico de sua seqüência de aminoácidos (AAs) é suficiente para que seja possível obter informações a respeito de suas funções biológicas, estrutura tridimensional, etc. Porém, várias técnicas computacionais atualmente utilizadas para esses fins possuem a desvantagem de possuir uma complexidade computacional de ordem $O(n^2)$ [2], no que diz respeito ao tamanho das seqüências de AA analisadas. Isto limita a utilização destas ferramentas para a análise de grandes massas de dados, visto que várias proteínas podem ser constituídas de centenas de AAs.

Felizmente, há também diversas técnicas computacionais de análise, baseadas em métodos de aprendizado de máquinas, cuja complexidade computacional é bem inferior a $O(n^2)$ [Wu (1992)]. É o caso das técnicas de árvore de decisão(AD) e redes neurais artificiais(RNAs). As RNAs oferecem uma arquitetura computacional singular, visto que o processamento realizado sobre uma massa de dados é realizado de forma dinâmica e rápida, isto é, as RNAs são capazes de se adaptar a novos dados que lhe sejam apresentados, em tempo real [Wu (1992), Baldi (1998), Wu (1997), Ding (2001), Uberbacher (1991), Murvai (2001)]. Esta adaptação ocorre através da atualização dos pesos relativos ao processamento de cada neurônio da RNA. O objetivo final destas atualizações é a minimização da soma quadrática dos erros da RNA, considerando-se cada padrão de treinamento apresentado à rede.

Entretanto, um fator crucial para o bom desempenho de uma RNA, ou de qualquer outra técnica computacional de análise, é a forma como os dados a serem processados são apresentados à ferramenta. Um bom pré-processamento dos dados "brutos" garante a confiabilidade das respostas geradas pela RNA, assim como pode diminuir consideravelmente a quantidade de dados que a RNA irá realmente processar. Devido a este fato, é comum o uso de algoritmos que sejam capazes de extrair informações, das mais diversas naturezas, a respeito dos dados que se deseja analisar. Com isso gera-se uma nova base de dados que contém informações representativas sobre os dados originais, isto é, geram-se "metadados", isto é, informações sobre informações, que podem ser utilizados, finalmente, para o treinamento e teste da RNA utilizada [Wu (1992), Ding (2001)]. Este é foco principal dos métodos utilizados para a classificação de famílias de proteínas, a partir da análise das suas seqüências extraídas das diversas bases de dados disponíveis para domínio público.

Por fim, vale ressaltar que vários métodos de classificação de proteínas já são bastante utilizados, sendo que o mais famoso dentre eles é o BLAST [Altshul (1990)]. Entretanto, estes métodos baseiam-se basicamente na análise de similaridade entre a seqüência de AAs das proteínas analisadas. Como famílias diferentes de proteínas podem conter em sua estrutura múltiplos domínios funcionais em comum, porém com funções bastante distintas, o simples compartilhamento destes domínios (denominados comumente de "domínios promíscuos") podem confundir estes métodos baseados em alinhamentos simples de seqüências de AAs [Murvai (2001)]. O que se pretende com o desenvolvimento de novas técnicas computacionais é disponibilizar ferramentas mais simples de processamento de informações, que necessitem de poucos parâmetros para a configuração de experimentos e que as respostas sejam mais confiáveis e fáceis de serem interpretadas.

Este artigo descreve a utilização de seis algoritmos de extração de informações combinados com a técnica de RNA do tipo Multi-layer perceptron (MLP) com a finalidade de classificação de proteínas de famílias distintas, mediante análise de uma base de "metadados" de diversas proteínas.

2 Algoritmos de extração de informações

Para a geração da base de "metadados" a ser utilizada para o treinamento e teste da RNA a ser utilizada, desenvolveu-se seis algoritmos capazes de extrair informações sobre as seqüências de AAs que representam as proteínas analisadas. Uma idéia semelhante foi desenvolvida por Ding e Dubchak [Ding (2001)].

2.1 Algoritmo 1 - Aminoácido mais freqüente na proteína.

Este algoritmo analisa uma dada seqüência de aminoácidos no formato FASTA, que represente uma dada proteína, retornando o código numérico referente ao aminoácido mais abundante na seqüência analisada (Tabela 1).

Tabela 1. Aminoácidos com os seus respectivos códigos textuais e numéricos.

CÓDIGO	AA	CÓDIGO NUMÉRICO
A	Alanina	0.1
B	Aspartato ou Asparagina	0.2
C	Cisteína	0.3
D	Aspartato	0.4
E	Glutamato	0.5
F	Fenilalanina	0.6
G	Glicina	0.7
H	Histidina	0.8
I	Isoleucina	0.9
K	Lisina	1.0
L	Leucina	1.1
M	Metionina	1.2
N	Asparagina	1.3
P	Prolina	1.4
Q	Glutamina	1.5
R	Arginina	1.6
S	Serina	1.7
T	Treonina	1.8
U	Selenocisteína	1.9
V	Valina	2.0
W	Triptofano	2.1
Y	Tirosina	2.2
Z	Glutamato ou Glutamina	2.3
X	Qualquer AA	2.4

Esta informação é de grande importância para uma avaliação prévia sobre as interações entre as proteínas e o ambiente em que esta se situe em um dado sistema biológico. Os aminoácidos mais abundantes provavelmente determinarão estes tipos de interações [Rost (1996)].

2.2 Algoritmo 2 - Porcentagem do aminoácido mais abundante.

Para o aminoácido identificado como mais abundante em uma proteína (Algoritmo 1), calcula-se a sua porcentagem em relação a todos os aminoácidos presentes em uma dada proteína, segundo a Equação 1:

$$p_{AA} = \left(\frac{n_{AA}}{t_{aa}} \right) * 100 \qquad (1)$$

Onde n_{AA} é o número que corresponde a quantas vezes o aminoácido mais freqüente aparece na seqüência analisada, t_{aa} é o número total de aminoácidos na seqüência e p_{AA} é a porcentagem do aminoácido mais freqüente.

Tabela 2. Grupos de polaridades e seus aminoácidos representativos.

GRUPO DE POLARIDADE	***AAs***
Polar (1)	Asparagina e Glutamina
Polar Positivo (2)	Arginina e Lisina
Polar Negativo (3)	Aspartato e Glutamato
Hidrofóbico (4)	Isoleucina, Leucina, Fenilalanina, Metionina, Valina, Triptofano e Cisteína(formando ligação S-S).

2.3 Algoritmo 3 - Caráter de polaridade da proteína.

Este algoritmo calcula a porcentagem dos grupos de aminoácidos com características de polaridade distintas. Adaptando-se o conceito de polaridade de AA descrito por Taylor (1986), dividimos alguns AAs mais representativos em quatro grupos de polaridades: polar carregado positivo, polar carregado negativo, polar e hidrofóbico, como descrito na Tabela 2.

2.4 Algoritmo 4 - Número de cisteínas presentes na proteína.

Cisteína é um aminoácido capaz de fazer ligações covalentes do tipo pontes dissulfeto com outras cisteínas [Baldi (1998), Enright (2002)]. Estas ligações são de grande importância para a manutenção das conformações tridimensionais das proteínas, assim como para a ancoragem das proteínas em outras estruturas celulares. Este algoritmo calcula a quantidade de cisteínas presentes na seqüência analisada, representando o número de ligações do tipo ponte dissulfeto que a proteína pode produzir.

2.5 Algoritmo 5 - Distância entre cisteínas existentes.

Este algoritmo calcula a distância média entre as cisteínas existentes em uma proteína. Esta informação é importante pois nos dá uma noção sobre o tamanho de possíveis alças (Figura 1) existentes em proteínas, formadas por pontes dissulfeto intraprotéicas.

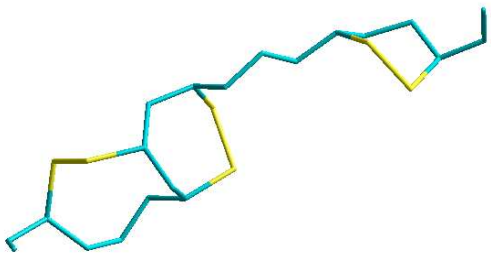

Figura 1. Esquema de uma proteína, com as pontes dissulfeto representadas em amarelo.

2.6 Algoritmo 6 - Distância média entre "ilhas hidrofóbicas".

Definimos "ilhas hidrofóbicas" como sendo subgrupos consecutivos de quatro ou mais aminoácidos hidrofóbicos (ver Tabela 2) dentro de uma seqüência de AAs. Este algoritmo calcula o valor "médio reduzido" da distância entre as ilhas hidrofóbicas existentes em uma proteína, segundo a Equação 2:

$$\mu_{IH} = \frac{\sum d_{IH}}{\log_{10}\left(\prod d_{IH}\right)} \quad (2)$$

Onde, $\sum d_{IH}$ é o somatório das distâncias consecutivas, medidas em número de aminoácidos, entre as ilhas hidrofóbicas existentes na seqüência analisada, $\prod d_{IH}$ é o produtório destas distâncias e μ_{IH} é o valor médio reduzido.

Cada proteína da base de dados utilizada foi analisada pelos seis algoritmos, cujos resultados foram armazenados em um arquivo que foi utilizado para os treinamentos e testes das redes neurais. Todos os algoritmos foram implementados em JAVA.

3 Base de dados utilizada

Para a geração dos "metadados" utilizou-se uma base de dados contendo proteínas de dez famílias diferentes, perfazendo um total de 1413 proteínas. As proteínas foram pesquisadas e suas respectivas seqüências de AAs foram extraídas da base de dados do National Center for Biotechnology Information (NCBI) [NCBI (2006)]. As seqüências foram extraídas no formato FASTA, para facilitar o processamento das seqüências pelos algoritmos de extração.

4 Rede Neural Artificial

Para a resolução do problema de classificação proposto utilizou-se uma rede neural artificial do tipo *feed-foward Multi-layer Perceptron* (MLP), cujo algoritmo de aprendizagem escolhido foi o *back propagation* [Braga (2000), Haykin (1998), Wu (1992)].

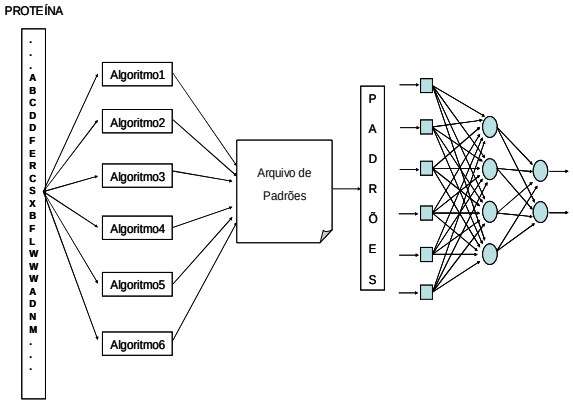

Figura 2. Fluxo de processamento do experimento.

Os parâmetros escolhidos para a construção da rede foram: (i) uma camada escondida com quatro neurônios; (ii) taxa de aprendizado $\eta = 0{,}05$; (iii) épocas de treinamento = 2000.

Para uma melhor avaliação do desempenho da rede construída utilizou-se o recurso de validação cruzada 10-fold, que é um método onde todas as instâncias da base de dados têm oportunidade de participar tanto do treinamento quanto do teste da rede. Como parâmetros para avaliar o desempenho e confiabilidade do sistema de classificação proposto, avaliou-se a taxa de acerto da rede neural, em relação às instâncias analisadas, e o valor da área sob a curva ROC [Fawcett (2006)]. Para a simulação dos experimentos com a rede neural, utilizou-se o *software* WEKA [WEKA Tool (2006)]. A Figura 2 ilustra o fluxo geral do processamento de informações realizado neste experimento.

5 Resultados e Discussão

A base de dados escolhida para os treinamentos e testes das redes neurais construídas é formada pelas famílias de proteínas exibidas na Tabela 3.

Dentre as famílias selecionadas, optou-se aleatoriamente classificar três famílias: DNA polimerase, Citocromo B e Citocromo C subunidade I. Para cada família escolhida para classificação gerou-se um arquivo de metadados específico, onde os padrões para os padrões pertencentes a família atribuiu-se o valor 1 (padrões positivos) como saída correta das RNAs e os padrões das demais famílias atribuiu-se o valor 0 (padrões negativos) como resposta.

Tabela 3. Grupos de polaridades e seus aminoácidos representativos.

Família de proteína	Quantidade de padrões
Citocromo C subunidade I	429
Citocromo B	558
Ligases	15
Miosina	2
Actina	97
Fosfotransferase	42
Capsídeo	10
Fosforilase	55
Tripsina	88
DNA Polimerase	117
Total de Proteínas	**1413**

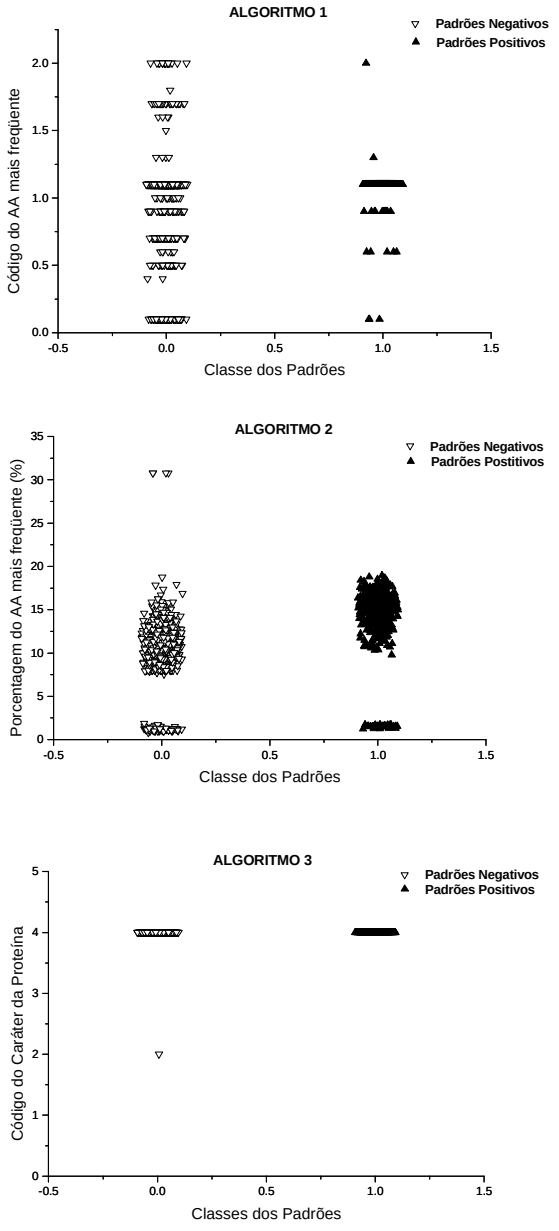

Figura 3. Respostas do processamento dos padrões, realizados pelos algoritmos 1, 2 e 3, tendo a família do Citocromo B como padrões positivos.

Em seguida, processaram-se todas as seqüências de proteínas com os algoritmos implementados, resultando em um arquivo final no formato ARFF, que é capaz de ser interpretado pela ferramenta de simulação WEKA.

As Figuras 3 e 4 exibem as respostas dos processamentos dos algoritmos desenvolvidos sobre os padrões de proteínas, para o caso da família do Citocromo B.

Como as respostas de vários padrões ficaram superpostas, dificultando a visualização dos respectivos pontos nos gráficos, subtraiu-se aleatoriamente, valores entre 0 e 0,5 aos valores dos padrões aos quais cada proteína pertencia. Isto ajuda a melhor visualização de todos os pontos em torno de 0 (padrões negativos) e 1 (padrões positivos).

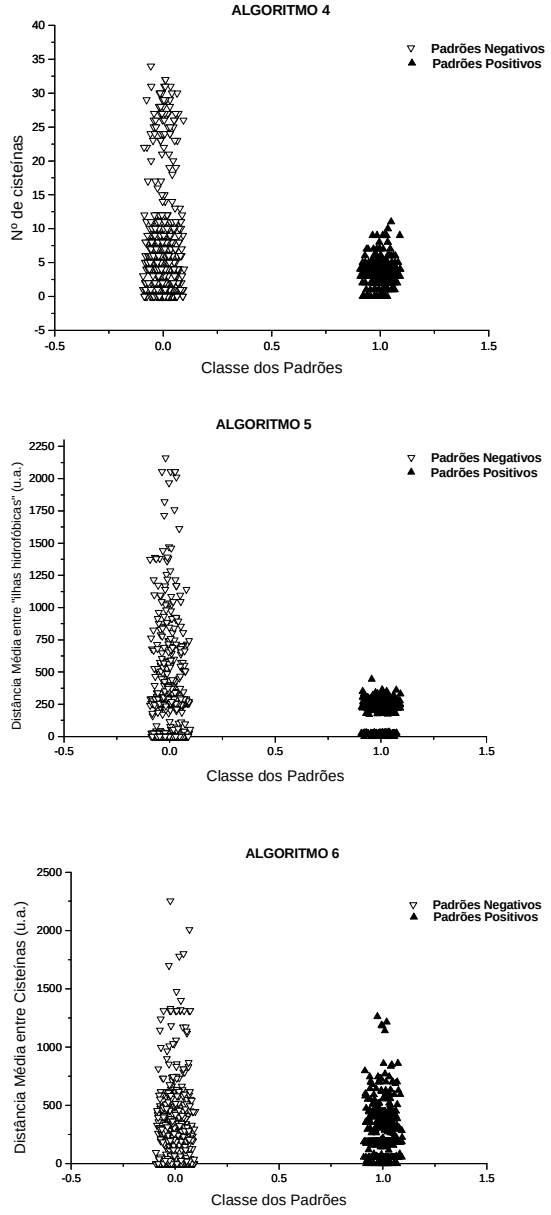

Figura 4. Respostas do processamento dos padrões, realizados pelos algoritmos 4, 5 e 6, tendo a família do Citocromo B como padrões positivos.

Como pode ser observado nas Figuras acima, os algoritmos não possuem, individualmente, capacidade de classificar os padrões processados, visto que as respostas dos padrões positivos e negativos encontram-se distribuídos nos mesmos intervalos, para cada tipo de algoritmo. Isto justifica a utilização de técnicas computacionais de classificação, como é o caso das MLPs.

Os treinamentos e testes realizados para cada MLP foram feitos seguidos os parâmetros descritos na Seção 4. Os resultados dos experimentos realizados encontram-se exibidos na Tabela 4.

Tabela 4. Resultados das classificações realizadas com MLPs.

Bases de dados testadas (Padrões Positivos)	**MLP**	
	Taxa de Acerto (%)	**Área sob Curva ROC (AUC)**
Actina	94.21	0.8488
Citocromo B	91.71	0.9517
Citocromo C subunidade I	92.65	0.8868
DNA polimerase	95.62	0.9294

As taxas de acertos obtidas ns experimentos realizados foram bastante satisfatórias. Este resultado reforça a hipótese de que a combinação dos algoritmos implementados é capaz de produzir um conjunto de metadados representativos da base de dados original e que é capaz de produzir uma boa classificação. Porém, o quão confiável são as taxas de acertos encontradas?

Os valores obtidos de áreas sob a curva ROC para cada experimento (do inglês *Area Under Curve* (AUC)) representam a precisão da classificação de cada padrão, em cada experimento realizado. Quanto mais próximo do valor de 1.0, mais precisa é a resposta de classificação, diminuindo a ocorrência de classificações falso-positivas ou falso-negativas. Este é um recurso bastante útil para validar a metodologia de classificação proposta, tendo em vista a dificuldade de se comparar os resultados obtidos com resultados de outras técnicas de classificação existentes, como aquelas baseadas em análises utilizando o BLAST [Murvai (2001)].

6 Conclusão

Os resultados obtidos nos experimentos utilizando MLPs para a classificação de proteínas apresentaram-se bastante satisfatórios, considerando-se a simplicidade dos algoritmos implementados paraa o pré-processamento das informações. Isto mostra a viabilidade deste tipo de metodologia para a realização de aplicações reais de classificação em bioinformática. A possibilidade de utilização de diferentes

combinações de algoritmos, capazes de extrair diferentes informações químicas e/ou biológicas, com outras ferramentas de classificação, assim como o desenvolvimento de novos algoritmos de pré-processamento é o foco de estudos do nosso grupo de pesquisa.

Como etapas futuras do nosso trabalho iremos aumentar o número de padrões e de famílias de proteínas contidas na base de dados utilizada em nossos experimentos, com o intuito de medirmos o poder de generalização das redes neurais treinadas. Também temos a intenção de estudarmos a aplicação de outras ferramentas de classificação, como é o caso das Máquinas de Suporte Vetorial [Lorena (2007)].

Agradecimentos

Os autores agradecem a CNPq pelo suporte financeiro.

Referências Bibliográficas

Altshul, S.F., Gish, W., Miller, E.W., and Lipman, D. J. (1990). Basic lignment search tool. *J. Mol. Biol.*, **215**: 403-410.

Baldi, P., and Brunak, S. (1998*). Bioinformatics: The machine learning approach*. Cambridge: MIT Press.

Barnes, M. R. and Gray, I. C. (2003). *Bioinformatics for Geneticists*. England: John Wiley & Sons.

Braga, A. P., Ludemir, T. B., Carvalho, A. C. P. L. F. (2000). *Redes neurais artificiais Teoria e aplicações.* Rio de Janeiro: LTC.

Ding, C. H. Q. and Dubchak, I. (2001). Multi-class protein fold recognition using suport vector machines and neural networks. *Bioinformatics*, **17** (4): 349-358.

Enright, A. J., Van Dongen, S. and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, **30**(7): 1575-1584.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Lett*, **27**: 861-874.

Haykin, S. (1998) *Neural Networks: A Comprehensive Foundation*. Prentice-Hall.

Lorena, A. C.,Carvalho, A.C.P.L.F. (2007) Protein cellular localization prediction with Support Vector Machines and Decision Trees. *Computers in Biology and Medicine*, **37**:115-125.

Murvai, J., Vlahovicek, K., et. al. (2001). Prediction of Protein Functional Domains from Sequences Using Artificial Neural Networks. *Genome Research*, **11**: 1410-1417.

NCBI Database. URL: http://www.ncbi.nih.gov/. Acessado em 15/09/2006.

Rost, B., and Sander, C. (1996). Bridging the protein sequence-structure gap by structure predictions. *Annu. Rev. Biophys. Biomol. Struct*, **25**:113-36.

Taylor, W. R. (1986) Classification of amino acid conservation. *J. Theor. Biol.*, **119**: 205-218.Wu, C., Whiston, G., McLarty, J., et. al. (1992). Protein classification artificial neural system. *Protein Sci.*, **1**: 667-677.

Uberbacher, E. C., and Mural, R. J. (1991). Locating protein-coding regions in human DNA sequences by a multiple molecular sequence analysis. *Computers Chem*, **21**(4):237-256.

WEKA Tool. URL: http://www.cs.waikato.ac.nz/ml/weka/. Acessado em 15/09/2006.

Wu, C.H. (1997). Artificial neural networks for sensor-neural network approach. *Proc. Natl. Acad. Sci. USA*, **88**: 11261-11265.