PREDIÇÃO DE ESTRUTURAS SECUNDÁRIAS DE PROTEÍNAS UTILIZANDO REDES NEURAIS ARTIFICIAS

Rilson Machado de Oliveira[1], **Thiago de Souza Rodrigues**[1]

[1]Departamento de Ciência da Computação – Universidade Federal de Lavras (UFLA)
Caixa Postal 37 – 37200-000 – Lavras (MG) – Brasil

rilson@comp.ufla.br, thiago@dcc.ufla.br

Resumo. A pesquisa se encontra na área de bioinformática, objetiva-se a predizer a estrutura secundária de uma proteína a partir de sua seqüência de aminoácidos, ou seja, sua estrutura primária. A predição foi feita utilizando redes neurais artificiais, que é um modelo computacional baseado no funcionamento de neurônios. Um banco de dados de proteínas, PDB (Protein Data Bank) foi utilizado para obter as informações das seqüências. Ao fim da pesquisa obteve uma taxa de exatidão de 78.1 % para a predição

Palavras-chave. bioinformáica, redes neurais artificiais, predição de estruturas de proteínas.

Prediction of secondary structures proteins with neural network

Abstract. *This bioinformatics research aims the prediction of protein secondary structures from its amino acid sequence, in other words, its primary structure. The prediction will be accomplished using artificial neural networks, which are a computational model based on the behavior of neural cells. A protein database PDB (Protein Data Bank) will be used in order to obtain information on the sequences. In the end this research the accuracy was 78.1%.*

Keywords: *bioinformatics, artificial neural networks, protein structure prediction*

(Received October 31, 2008)

1. Introdução

As proteínas estão presentes em todos os organismos vivos, elas desempenham um papel fundamental nestes organismos, sendo uma estrutura básica e fundamental para a vida. Esses componentes básicos desempenham funções variadas, ter o conhecimento da função realizada pelas inúmeras proteínas é de grande utilidade, pois com essas informações podem-se diagnosticar doenças, descobrir curas, desenvolver novos medicamentos, entre outras inúmeras utilidades.

As proteínas possuem uma propriedade muito importante que é a sua estrutura, que por sua vez pode se classificar em diferentes estágios organizacionais, estrutura primária, estrutura secundária e estrutura terciária. Segundo [1] ao estudar a estrutura da proteína pretende-se descobrir quais as propriedades da proteína que levam a cadeia a adotar uma estrutura única e estável e também, investigar como a seqüência de aminoácidos, estrutura primária, de uma proteína está relacionada com essas propriedades.

Segundo [2] prever a estrutura tridimensional de uma proteína a partir de aminoácidos é um dos maiores problemas ainda não resolvido da biologia molecular. Os biologistas estruturais vêm pesquisando a chave deste mistério nas estruturas catalogadas de proteínas, procurando por padrões de aminoácidos que se correlacionem a estruturas específicas. Programas computacionais de previsão da estrutura baseados nestes estudos estatísticos têm sido desenvolvidos e são úteis, mas não perfeitos. Se ao tentar prever a estrutura tridimensional de uma proteína cujo gene foi descoberto recentemente, é mais provável que se tenha mais sucesso se uma proteína similar já foi descoberta e sua estrutura determinada.

Embasando-se nestes problemas este trabalho objetiva prever a estrutura secundária da proteína através de sua seqüência, obtidas de um banco de seqüências de proteínas. Para tal, serão utilizados recursos computacionais, neste caso redes neurais artificiais.

As redes neurais artificiais oferecem um bom suporte para prever as estruturas de proteínas, pois a sua propriedade de aprendizagem e de generalização garante bons resultados de previsões. O tipo de redes neurais utilizada é a rede multilayer perceptron com treinamento de taxa de aprendizado auto ajustável com essa topologia de rede proposta foi psossível conseguir uma taxa de exatidão de 78,1%.

2. Revisão Bibliográfica

A predição de estruturas secundárias de proteínas são importantes pois com elas pode-se identificar quais são as funções desempenhadas pela proteína.

Como [3] disse estudos sobre predição de estruturas secundárias de proteínas por métodos computacionais e métodos estatísticos começaram com Krigbaum e Kuntton, Eles usaram algoritmos de regressão linear pra predizer, em seguida Chou-Fasman usou um método estatístico empírico baseado em freqüências de tipos de estruturas secundárias.

De acordo com [4] desde 1989 procura-se predizer as estruturas secundárias de proteínas por recursos computacionais. Mas até hoje o problema se mantém aberto, pois não se obteve um algoritmo que substitua totalmente os exames de laboratório.

Existem várias formas de se predizer a estrutura secundária a partir da estrutura primária. A mais utilizada deles é a predição por Redes Neurais Artificiais, o recurso computacional utilizado neste trabalho, mas também foram utilizados outras metodologias de predição como, Modelos de cadeia de Markov, Redes Bayesianas, Vector Machinas.

A taxa de aprendizado para predição de estruturas secundárias de proteínas tem sido melhorado constantemente. Existem dois tipos de algoritmos para a predição. A primeira são algoritmos de predição para seqüências simples, que implica em não se ter o conhecimento de proteínas homólogas para realizar a predição. E o segundo tipo é quando se tem informações sobre proteínas homólogas.[5]

A predição por redes neurais artificias normalmente é modelada da seguinte maneira: as proteínas que já se tem o conhecimento de suas estruturas primárias e secundárias são utilizadas pela rede, a estrutura primária, como já foi dito, é a sua seqüência de aminoácidos e será a entrada da rede, com se tem somente a informação dos aminoácidos que as compõe, alguma codificação dessa seqüência deve ser realizada. Pois redes neurais artificias, por definição só possuem dados numéricos como entrada. A estrutura secundária da proteína servirá como o vetor de valores esperados para a rede. Depois de definido a entrada e valores desejados, é preciso escolher a topologia, o algoritmo de treinamento e os ajustes dos parâmetros da rede, tais com número de épocas, função de ativação, número mínimo do gradiente, tempo máximo, e outros parâmetros que o algoritmo de treinamento tiver.

A codificação dos dados, a topologia da rede, o algoritmo de treinamento e os ajustes de parâmetros que são fatores necessários para se criar uma rede neural artificial para a predição de estruturas de proteínas deverá ficar a critério de escolha do pesquisador .

Em [6] desenharam a primeira rede neural para realizar predições de estruturas secundárias de proteínas, conseguindo uma taxa de generalização de 64.5%. Em seguida [7] conseguiram uma melhora taxa de acerto conseguindo chegar aos 65.5% de generalização. Outras pesquisas foram realizadas nos anos seguinte propondo outros algoritmos, mas nenhum trabalho se destacou, até que [5] introduziram na predição profiles alinhados com múltiplas seqüências alinhadas, o método foi chamado de PHD, e possuía performance bem melhor que as anteriores, pois utiliza alinhamento de profiles como entrada da rede, chegando a taxa 70%. [8] fez grades melhorias por ser pioneiro e usar a posição específica pontuando matrizes, esse método foi denominado PSSM com isso gerou-se o PSI-BLAST profile controlados, e logo em seguida pode-se criar o PSIPRED, a generalização desse método era um pouco superior a 70%.

PSIPRED server é um servidor de predição de estruturas de proteínas que está disponível em [10] ele disponibiliza para usuários submissão uma seqüência de proteína, os resultados da predição são enviados como uma mensagem de texto via e-mail, e graficamente via web. Esse servidor é constantemente atualizado, inserindo novos algoritmos e melhores resultados, hoje a

taxa de exatidão do PSIPRED chega a 80% de exatidão. [9]

Recentemente, novas técnicas de treinamentos e topologias de redes neurais artificiais tem sido freqüentemente usadas com intuito de obter melhores resultados para o problema da predição de estruturas secundárias de proteínas. Como por exemplo pode-se citar redes neurais recorrentes, redes neurais holpfiel, redes neurais qprop, nprop, redes neurais com momentun. Com isso esses trabalhos conseguem cada vez mais, uma melhor eficiência nas predições.

Outro método para se prever estruturas secundárias de proteínas é por modelos de Markov. Um modelo escondido de Markov é uma máquina probabilísticas de estados finitos para modelos estocásticos de seqüências. O modelo de Markov é definido pelo conjunto de estados, emissão probabilística associado com cada estado conectado. Um pode associar uma probabilidade com uma seqüência de acordo como o modelo de Markov que gera aquela seqüência.[11]

Para usar um modelo de Markov para rotular uma seqüência é preciso associar uma etiqueta com cada estado, ou mais genericamente uma probabilidade de um marcador específico em virtude do estado. O rótulo atribuído a cada elemento na seqüência depende de qual estado que provavelmente tenha emitido o elemento. Esses modelos têm sido amplamente utilizados em bioinformática porque o conhecimento pode ser codificado para esses modelos e ainda permitindo que outras informações a serem aprendidas através de treinamentos das emissões e da transição das probabilidades dos dados. [11]

[12] fez o primeiro modelo de cadeia de Markov para predição de estruturas secundárias de proteínas, com taxa de exatidão de 70%. Modelos de Markov com algoritmos genéticos foram desenvolvidos a fim obter melhores resultados.

O método de supoort vector machines proposto por [13] é um método muito eficiente para reconhecimento de padrões , aprendizagem por support vector machines é a fronteira entre exemplos pertencentes a duas classes mapeando exemplos de entrada com um grande espaço dimensional, procurando um hiperplano de separação neste espaço. O hiperplano de separação é escolhido de forma a maximizar sua distância com relação aos exemplos de treinamento mais próximos. O hiperplano é chamado de separação hiperplana ótima.

Support vector machines tem um número interessante de propriedades, incluindo a evitação efetiva de overfitting, possui a habilidade de lidar com grandes espaços, informação condensada dos dados, entre outras.

Para [14][15] construir uma support vector machine para predizer estruturas secundárias de proteínas pode ser mais fácil que construir uma rede neural artificial. A estrutura apropriada da rede neural dependerá do nível do desenvolvedor, já no caso de support vector machines é necessário somente selecionar a função e regularizar um parâmetro para se poder começar a treinar. Logo após é preciso determinar a janela de largura ótima para cada binário classificador. As taxas de exatidão para support vector machine, chegando a 77% obtida por [15] demonstra a boa performance para a solução do problema.

3. A rede neural artificial

Pode-se observar que apesar de ser um problema muito visado na biologia computacional, a predição de estruturas secundárias de proteínas é um problema que precisa ser trabalhado. Os métodos aqui implementados visão dar um suporte à mais para os estudos no assunto. Para que algum dia o problema possa ser totalmente resolvido.

Para o treinamento da RNA foi utilizado o banco de dados público de proteínas Protein Data Bank, PDB, esse banco contém informações como o nome da proteína, sua seqüência (estrutura primária), possui informações sobre o tamanho a que estrutura secundária pertence essa seqüência. E outras muitas informações não relevantes para este trabalho. A Figura 1 mostra uma parte do PDB, pode-se observar que possui os campos referentes a quantidade de aminoácidos (tamanho da seqüência), a seqüência e o campo DSSP que representa a estrutura secundária equivalente a cada aminoácido da seqüência. Para exemplificar, pode-se observar que a subseqüencia de aminoácidos FEMLRIDE no início da seqüência, representa uma estrutura secundária. Toda seqüência contínua de letras no campo DSSP representa uma única estrutura secundária. As estruturas alfa-helices são representadas pela letra H, as estruturas folha-beta são representas pela letra E e as estruturas Coils são representadas pela letra

C, as demais letras são outras tipos de seqüências que não são tratadas neste trabalho.

```
Amino-Acids : 162
Sequence    : MNIFEMLRIDEGLR
DSSP        : CCHHHHHHHHHCCE
```

Figura 1: PDB - protein data bank Fonte: dados do trabalho

Para entrada de dados na rede, uma filtragem de dados foi realizada. Como a rede neural aceita somente como entrada seqüências de mesmo tamanho, uma seleção de seqüências de mesmo tamanho foi feita. Os dados foram separados por seqüências de tamanho dez, onde esses dez aminoácidos representam um tipo de estrutura secundária. A escolha do tamanho dez se deve ao fato que em média o tamanho das estruturas são de oito a doze aminoácidos, testes realizados obteve uma maior número de seqüências com tamanho dez.

Com a filtragem realizada pode-se observar que a quantidade de estruturas do tipo alfa-helice, folha-beta e coil são predominantes nas seqüências, foram catalogadas para a rede 90563 subseqüências, de tamanho dez, desses três tipos básicos, e 1047 subseqüências, de tamanho dez, de outros tipos.

Como o tipo de estruturas secundárias mais encontradas são as alfa-helices as folhas-beta e as coil, a rede neural reconhecerá somente estes três tipos mais básicos de estruturas, o reconhecimento de outros tipos acarretaria em uma queda de performance da rede. Pois a quantidade de subseqüências que não são estes três tipos é considerada insignificante, a rede simplesmente não os reconheceriam, e com isso reduziria a performance.

Outro fator importante para a rede neural é que as entradas além de serem do mesmo tamanho necessitam que os dados estejam em valores numéricos, e assim as seqüências filtradas passaram por uma codificação. Uma classificação por hidrofobicidade foi realizada, os aminoácidos recebem valores reais dependendo de seu grau hidrofóbico essa codificação também é denominada escala KD.A Figura 2 expõe as codificações para cada um dos 20 aminoácidos. E também possui informações sobre que categoria de hidrofobicidade ele se encontra, que podem ser hidrofóbico, neutro ou hidrofílico. Pode-se observar que os valores começam em 0.05 e incrementam de acordo com os níveis de hidrofobicidade.

Aminoácido	Escala KD	Valor Real	Categoria
I	+4,5	0,05	Hidrofóbico
V	+4,2	0,10	Hidrofóbico
L	+3,8	0,15	Hidrofóbico
F	+2,8	0,20	Hidrofóbico
C	+2,5	0,25	Hidrofóbico
M	+1,9	0,30	Hidrofóbico
A	+1,8	0,35	Hidrofóbico
G	-0,4	0,40	Neutro
T	-0,7	0,45	Neutro
S	-0,8	0,50	Neutro
W	-0,9	0,55	Neutro
Y	-1,3	0,60	Neutro
P	-1,6	0,65	Neutro
H	-3,2	0,70	Hidrofílico
Q	-3,5	0,75	Hidrofílico
N	-3,5	0,80	Hidrofílico
E	-3,5	0,85	Hidrofílico
D	-3,5	0,90	Hidrofílico
K	-3,9	0,95	Hidrofílico
R	-4,0	1,00	Hidrofílico

Figura 2: valores reais atribuídos a cada aminoácido conforme escala de hidrofobicidade Fonte: dados do trabalho

Foram executados várias configurações de redes neurais artificiais, a fim de obter o melhor resultado, a configuração que obtive melhor resultado será aqui detalhada.

A melhor rede treinada, foi uma rede multi layer perceptron com treinamento backpropagation modificado com taxa de aprendizado adaptativa (treinamento Batch Training traingda do MatLab), feedforward, com taxa de momentum e funções de ativação tangente hiperbólica sigmoidal.

Em relação as camadas da rede, a primeira camada é a camada de entrada, esta é composta por dez neurônios cada neurônio representa um aminoácido e a junção destes dez aminoácidos (já codificados) é a representação de uma subseqüência que representam uma estrutura, cada vetor de entrada (os dez aminoácidos) pode representar uma estrutura diferente, podendo ser alfa-helice, folha-beta ou coil. Como foi escolhida uma topologia de rede multi layer perceptron, MLP, a rede possui 55 neurônios na camada intermediária, a escolha desse número foi randômica, e apresentou melhores resultados que configurações com menos neurônios nesta camada. E por fim tem-se três

neurônios na camada de saída, cada um representando uma estrutura, entrando com uma subseqüência a rede diz se está é um dos três tipos de estruturas. Assim a rede neural artificial em relação a camadas tem a seguinte configuração: camada de entrada com 10 neurônios, camada intermediária com 55 neurônios e camada de saída com três neurônios, que pode ser visto na Figura 3.

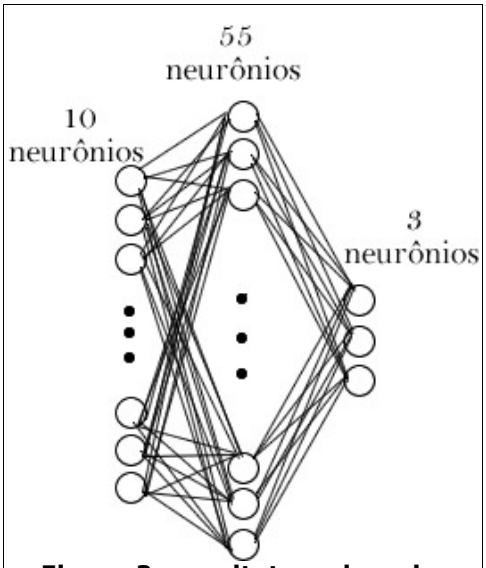

Figura 3: arquitetura da rede. Fonte: dados do trabalho

O algoritmo de treinamento foi o traingda, ou seja algoritmo bacpropagation com taxa de aprendizado de modo adaptativo. O número de épocas executadas foram 6000, a taxa de momentum 0.5 e taxa de aprendizado 0.05. Esses foram os parâmetros da rede.

Dos dados obtidos, cerca de 70 % deles foram separados para o treinamento, e os outros 30% foram separados para validação da rede, quer dizer, 70% dos dados serão carregados como entrada e a rede neural artificial treinará com eles, os outros 30% servirão para a simulação, a verificação da performance da rede. Como pode-se observar a separação dos dados pela Tabela 1.

Estrutura	Para 70% dos dados	Para 30% dos dados
Alfa-Helice	29794	12770
Folha-Beta	20286	8694
Coil	13313	5706
Total	63393	27170

Tabela 1: Quantidade de subsequências para treinamento e dados para validação Fonte: dados do trabalho

3.1. Ambiente de desenvolvimento

O trabalho foi realizado em um computador core 2 duo 1.86 Ghz, 1Gb de memória RAM, com sistema operacional microsoft windows XP service pack 3, o dados foram filtrados e codificados por um programa feito na linguagem Java, e a rede foi feita, treinada e simulada pelo MatLab com a toolbox de redes neurais artificias.

4. Resultados e Discussão

4.1. Os resultados

O problema consiste na predição de estruturas secundárias de proteínas. Isso é, prever qual será a configuração de estruturas secundárias de uma dada proteína, através de sua estrutura primária, ou seja, através de sua seqüência de aminoácidos.

Para isto como já foi dito, existem vários métodos de se prever essas estruturas secundárias de proteínas, dentre elas se destacam a predição por redes neurais artificiais, a predição utilizando métodos estatísticos, outros utilizando support vector machine, entre outros. Bons resultados já foram obtidos utilizando esses métodos, mas esse trabalho focalizou a obtenção das estruturas secundárias utilizando redes neurais artificiais.

Primeiramente ocorreu um tratamento dos dados, foi necessário realizar filtragens e codificações das seqüências de aminoácidos obtidos no banco de dados de proteínas, para que eles ficassem no formato permitido da rede neural artificial. Logo em seguida foi preciso identificar qual a topologia, a arquitetura e os parâmetros da rede. Com isso foi possível treinar a rede

e depois de treinada, realizar simulações para obter os resultados.

Com o treinamento a rede obteve um erro de 0.106599. E com esse resultado obteve uma taxa de acertos totais de 78.1%, sendo que para Alfa-Helices a taxa foi de 89%, para folha-Beta a taxa foi de 77 % e de Coil a taxa foi de 68.3 %, para os 30% dos dados reservados à validação. Como pode ser visto na Tabela 2.

Estrutura	Performance (%)
Alfa-Helice	89
Folha-Beta	77
Coil	68.3
Média	78.1

Tabela 2: Performance da rede. Fonte: dados do trabalho

O resultado para Alfa-Helice foi o melhor resultado pois essas estruturas são as que apresentam maior volume de subseqüências. E a baixa performance para as coils se deve ao fato de uma quantidade reduzidas dessa estrutura nas subseqüências.

4.2. Comparativo dos resultados

Como a taxa de generalização da rede foi de 78.1% ela obteve uma performance relativamente menor que aos que se encontram na literatura

A Tabela 3 mostras alguns dos melhores resultados obtidos na literatura. Os resultados podem ser obtido em[14]-[25].

Como o resultado deste trabalho foi uma taxa de generalização de 78,1%, um comparativo com resultados melhores do que o proposto aqui será detalhado. O algoritmo 8 [21] utilizou uma rede neural bidirecional recorrente com pequenas podas, assim obtendo um taxa de generalização de 79%. O algoritmo 9 [22] utilizou Posição específica de pontos em matrizes como entrada da rede, enquanto a saída com três estados consecutivos, a predição ocorreu com treinamnto por cross validation e foi testado em 1032 proteínas seqüênciadas, conseguindo assim uma taxa de generalização de 80%.O algoritmo 10 [23] usou uma rede com larga escala de treinamento, em um cluster de alto desempenho com 22 processadores, a rede foi implementada com treinamento com cross validation e obterve uma taxa de generalização de 80%. O algoritmo 11 [24] utilizou os resultados de um rede neural, feita por ele, que possuia uma um taxa de generalização de 79% com cross-validation, e aplicou métodos estatísticos e aplicou o método Ψ dihedral angles, assim obtendo 80.7 % a 81.7% de exatidão. E o algoritmo 12 [25] usou para cada aminoácido na proteína alvo, foi combinado os resultados do PROSP e PSIPRED usando uma função hibrida. Foram utilizadas duas bases de dados para o treinamento e a validação, o PDB e DSSPdataset, conseguindo assim uma taxa de generalização de 81.8%.

ID	método	ano	performance (%)
1	markov	2006	70.3
2	estatístico	1998	72.9
3	rede neural	2005	73.5
4	logica fuzzy	2005	75.75
5	SVM	2001	76.2
6	estatístico	2002	76.5
7	SVM	2007	77
8	rede neural	2004	79
9	rede neural	2000	80
10	rede neural	2006	80
11	estatístico	2005	80.7-81.7
12	rede neural	2005	81.8

Tabela 3: resultados para predições

Mesmo não sendo a melhor solução para predição de estruturas secundárias de proteínas, o método aqui apresentado obteve um bom desempenho ao considerar que está entre os melhores resultados já obtidos.

Os resultados desse trabalho são bem melhores que os primeiros resultados, resultados que podem ser vistos no capítulo dois, isso de deve ao fato que o número de proteínas que se tem conhecimento e que servem de base para realizar o treinamento é bem maior do que tinha naquela época, e métodos de treinamentos mais sofisticados, com heurísticas, também proporcionam melhores resultados.

5. Conclusão

Com esse trabalho pode-se concluir que a falta de informações sobre como foram realizados os processos

de obtenção dos dados e tratamento , para os resultados obtidos na literatura não pode-se chegar a uma conclusão na diferença dos resultados. Para uma análise comparativa seria necessário ter todas essas informações, pois não se pode comprar processos onde os dados estão em formato diferente.

As limitações para este trabalho se tem com a falta de detalhes dos algoritmos disponíveis na literatura, fazendo com que o processo de reprodução ou de comparação aos resultados existentes se tornem difíceis.

A complexidade do problema o torna difícil de se tratar, ficando evidente pelo baixo nível dos resultados, onde a melhor predição encontrada leva a taxas de somente 81.8% de exatidão, somente com resultados melhores a predição poderia ser usada a fim de não mais precisar a utilização de métodos caros de laboratório para descobrir as estruturas de uma nova proteína descoberta ou catalogada.

E como trabalhos futuros poderá ser treinada redes separadas para os três tipos estruturas, uma rede treinada para alfa-helices, uma para folha-beta e outra para Coil, a fim de tentar melhorar a taxa de generalização. Assim construir preditores exclusivos para cada tipo de estrutura.

Também poderá realizar o treinamento da rede por outros algoritmos de treinamento com o algoritmo de treinameno Multi-Obetivo. A fim de tentar melhorar a performance dos resultados.

6. Referências Bibliográficas

[1] DILL, K A. Dominate Forces in protein Folding. In: . : Biochemistry, 1990. .

[2] Kreuzer, H., Massey, A.. Engenharia Genética e Biotecnologia. In: . : , 2002. .

[3] BRANDEN, C. & TOOZE, J.. Introduction to Protein Structure. Garland Publishing. In: . : , 1991. .

[4] Kaya, I. E.. Accurate Prediction of ProteinSecondary Structure By Non-Parametric Models. , India, v. , n. , p. , 2008.

[5] Rost B, Sander C. Predictions fo protein secondary structure at better than 70% accuracy. , , v. , n. , p. , 1993.

[6] Qian N, Sejnowski T. J.. Predicting the secondary structure of globular proteins using neural network models., 1988.

[7] Taylor, W. R. & Orengo, C. A. . Prediction of super-secondary structure in proteins. , London , 1989

[8] Jones D. T.. Protein secondary structure prediction structure based on position-specific scoring matrices., 1999.

[9] McGuffin L.J., Bryson K., Jones D. T.. The PSIPRED protein structure prediction server., p. 404-405, 1999.

[10] PSIPRED protein structure prediction server. Bioinformatics. Desenvolvido por: Bioinformatics, Disponível em: <http://bioinf.cs.ucl.ac.uk/psipred/psiform.html>. Acesso em: 25/10/2008.

[11] Won, K. J., Hamelryck, T., Prugel-Bennett, A. and Krogh, A. . Evolving Hidden Markov Models for Protein Secondary Structure Prediction., 2005.

[12] Asai k.. Prediction of protein secondary structure by the hidden Markov model., 1999.

[13] Cortes C., Vapnik V.. Support vector networks, machines learning., 1995.

[14] Hua S. Sun Zhirong. A Novel Method of Protein Secondary Structure Prediction with Segment Overlap Measure: Suppor Vector Machine Approach. , 2001.

[15] Nguyen M. N., Rajapakse J. C.. Prediction of Protein Secondary Structure with two-stage multi-class SVMs., 2007.

[16] Aydin, Z. Altunbasak, Y. Borodovsky, M.. Protein secondary prediction for a single-sequence using hidden semi-Markov models. , USA, 2006.

[17] Cuff J. A., Clamp M. E., Siddiqui A. S., Finlay M., Barton G. J.. Protein secondary structure prediction based on position-specific scoring matrices. , , v. , n. , p. , 1998.

[18] Sen T. Z., Jernigan R. L., Garnier J., Kloczkowski A.. GOR V server for protein secondary structure prediction . , , v. , n. , p. , 2005.

[19] Bondugula R., DuzlevskiO., Xu D.. Profiles and fuzzy k-nearest neighbor algorithm for protein secondary structure prediction. , , v. , n. , p. , 2005.

[20] Jones D.. Protein secondary structure prediction based on position-specific scoring matrices. , , v. , n. , p. , 2002.

[21] Pollastri G., McLysaght A.. Porter: a new, accurate server for protein secondary structure prediction. , , v. , n. , p. , 2004.

[22] Peresen T. N., Lundegaard G., Nielsen M.

Prediction of protein secondary structure at 80% accuracy. , , v. , n. , p. , 2000.

[23] Zhou Y.. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. , , v. , n. , p. , 2006.

[24] Wood M. J.. **Protein secondary structure prediction with dihedral angles**. . : , 2005. p.

[25] Lin H. N., Chang J. M., Wu K. P., Sung T. Y., Hsu W. L.. HYPROSP II-A knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence . , , v. , n. , p. , 2005.