Biomatemática ***17*** *(2007), 87–100* ***ISSN 1679-365X***
Uma Publicação do Grupo de Biomatemática IMECC – UNICAMP

Aplicação de Redes MLP na Predição de Estrutura Secundária de Proteínas - PREDCASA

Luis Paulo B. Scott[1]
CMCC – UFABC, 09.090-400 – Santo André/SP.
Jorge Chahine[2], José R. Ruggiero[3],
Depto. de Física, IBILCE, UNESP, 15.054-000 – S. J. do Rio Preto/SP.

Abstract. The prediction of secondary structure of proteins can contribute to elucidate the protein folding problem. In order to predict these structures we used methods of Artificial Neural Networks (ANN) starting form the primary sequences of amino acids. In this present work we use ANNs in the prediction of the secondary structures of proteins, taking as patterns the structures in helix form (H), beta sheet (E) and coil (C). The ANNs were trained with the Simulator of MATLAB. The obtained data are compared with predictors described: PSA, PSIPRED and PHD in order to have an idea of the quality of the prediction. The present work is composed of 3 networks level. The output form all levels 1 ANNs are then fed a single second level ANNs. The third level is composed of jury decision.

Palavras-chave*: Predição; Proteinas; Redes MLP.*

1 Introdução

Entre as várias classes de moléculas biológicas de grande importância para os seres vivos, encontram-se as proteínas. O termo proteína provém do grego (proteios) que significa "de primeira magnitude". As proteínas são moléculas complexas que possuem uma estrutura terciária (tridimensional) específica. Estas macromoléculas realizam tarefas de extrema importância para o organismo, como a catálise de reações químicas, transporte, reconhecimento e transmissão de sinal. A função das proteínas conhecidas está determinada pela sua estrutura espacial. Portanto é importante conhecermos a estrutura 3D dessas moléculas. O número de seqüências de proteínas

[1]Luis.scott@ufabc.edu.br
[2]chahine@ibilce.unesp.br
[3]zerug@ibilce.unesp.br

(estruturas primárias) conhecidas e depositadas em bancos de dados (swissprot) está crescendo muito mais rápido do que a nossa habilidade de resolver as estruturas terciárias experimentalmente Rost (1998); Rost et al. (1993). Portanto, técnicas eficazes para predição de estruturas são importantes para diminuir a diferença entre o número de seqüências depositadas e de estruturas 3D determinadas (Protein Data Bank – PDB).

Um método tradicionalmente, utilizado na predição de estruturas, é a modelagem por homologia. Porém, pode-se predizer a estrutura 3D de, aproximadamente, apenas 25% a 30% das seqüências de proteínas depositadas nos bancos de dados primários através desta técnica. A investigação e o desenvolvimento de softwares de predição de estruturas protéicas é importante: **para os estudos conformacionais; para auxiliar no estudo do enovelamento protéico e para experimentos tanto *in silico* como *in vitro*.** Esse artigo descreve os resultados da utilização de redes neurais do tipo Multi Layer Perceptron como processos de otimização na predição de estruturas secundárias de peptídeos e proteínas. O trabalho possui como objetivo principal desenvolver um software para **predição 1D Web.**

O trabalho, descrito nesse artigo, consiste na investigação de diferentes arquiteturas de redes neurais do tipo Multi Layer Perceptron para realizar a predição 1D fazendo uso de diversas informações físico-químicas da proteína. Já foram testadas diferentes arquiteturas de redes neurais, inclusive a combinação de duas redes neurais e de diferentes propriedades físico-químicas. As redes foram criadas, treinadas e testadas utilizando o MATLAB. O preditor de estrutura secundária desenvolvido atinge uma média de 70 a 78% de acerto e para algumas proteínas, em particulares, a taxa de acerto chega a 98% dos aminoácidos (posições da estrutura primária).

Nesse momento, pretende-se migrar o preditor da plataforma Windows para plataforma Linux. Além disso, pretende-se mudar de simulador de redes neurais e permitir que a comunidade acesse o preditor através de uma interface para Web. Dessa forma, pretende-se fornecer um serviço de predição de estrutura secundária para a comunidade científica. Foram implementadas e testadas 18 arquiteturas de redes neurais MLP distintas, cada uma contendo uma codificação de entrada (camada de entrada). Os resultados obtidos forma comparados com três preditores descritos na literatura: PHD, PSA e PSIPRED.

2 Contextualização

As proteínas são moléculas complexas que possuem uma estrutura terciária (tridimensional) específica e que realizam tarefas de extrema importância para o organismo, como a catálise de reações químicas, transporte, reconhecimento e transmissão de sinal. A função das proteínas conhecidas está determinada pela sua estrutura espacial. Dessa forma, obter a estrutura tridimensional da proteína no seu estado nativo, e compreender as forças que a estabilizam são um dos problemas fundamentais da Biofísica de proteínas. Numerosos trabalhos utilizando redes neurais artificiais para estudar estruturas protéicas (primária, secundária e terciária) têm sido realizados e descritos na literatura. A predição de estruturas secundárias é um passo útil e importante para compreender como a seqüência de aminoácidos de umas proteínas determina o seu estado nativo.

Os algoritmos de predição de estruturas secundárias com melhores resultados, no momento, são baseados em redes neurais. A maioria dos métodos, aplicando RNAs, correntemente disponíveis possuem três estados de predição e alcançam uma performance de 72% a 78% de forma geral, podendo ser melhor dependendo da base de dados e das informações de entrada da rede neural. Entre as possíveis aplicações de RNAs no estudo de estruturas de moléculas como proteínas estão:

- predição de estruturas secundárias através classificação/reconhecimento de padrões;
- predição de estruturas terciárias de proteínas através de otimização de uma função potencial de energia;
- predição de possíveis seqüências de aminoácidos para uma dada proteína de forma a obter as conformações de mais baixa energia.

Outros aspectos de estruturas de proteínas, tal como a classe estrutural, também pode ser predita utilizando redes neurais. Podem-se utilizar redes neurais para associar as proteínas a uma das grande quatro classes (toda α, toda β, α/β e outras) com uma precisão que pode chegar a 78% em alguns trabalhos.

Um dos trabalhos pioneiros de aplicação de RNA na predição de estrutura é devido a Holley and Karplus Holley e Karplus (1991). Eles usaram uma rede

MLP, não recorrente para predizer elementos de estrutura secundária de proteínas a partir da seqüência de resíduo. Eles codificam os dados de entrada em janelas de resíduos adjacentes e para cada resíduo, usaram 21 entradas binárias (ou seja, assumiram apenas zero e um), sendo que apenas uma das 20 entradas estava ativada (cada entrada representa um tipo de aminoácido) e uma codificada quando a janela sobrepunha o fim da cadeia.

A rede utilizada possuía uma camada intermediária com duas unidades e uma camada de saída também com duas unidades, que representava/codificava uma estrutura secundária associada ao resíduo central na janela. Para facilitar a comparação com outros métodos de predição, Holley & Karplus adotaram três tipos de estrutura secundária: hélice, folha e *coil* Holley e Karplus (1991). O conjunto de dados utilizados para treinar e testar a rede era composto de 62 proteínas, sendo que o conjunto de treinamento formado por 48 proteínas e o conjunto de teste (para predição) por 14 proteínas. Holley & Karplus testaram vários tamanhos de janela e, nesse caso, a janela de tamanho 17 mostrou os melhores resultados. Também foram testadas redes neurais com diferentes tamanhos de cadeia intermediária (variando de 2 a 20). Apesar da rede com 20 unidades (neurônios) na camada intermediária ter apresentado o melhor resultado para o conjunto de treinamento; a rede com duas unidades na camada intermediária mostrou o melhor resultado para o conjunto de teste. A precisão obtida para janelas de tamanho 17 e redes com dois neurônios em sua camada escondida foi de 63.2% para o conjunto de teste e 68.5% para o conjunto de treinamento.

Qian and Sejnowski também utilizaram redes neurais não lineares para predizer estrutura secundária de proteína globulares e avaliaram o efeito de ruído nos dados de treinamento em relação à curva de aprendizado da rede e sua performance nos teste Qian e Sejnowski (1996).

Chandonia and Karplus aplicaram duas redes neurais denominadas primária e secundária para predizer estruturas secundárias e classes estruturais ou classes de estrutura Chandonia e Karplus (1996). Nesse estudo foi utilizado um conjunto de 681 proteínas com estruturas disponíveis no *Protein Data Bank*(*PDB*). A rede neural primária utilizada para predizer a estrutura secundária era similar a várias descritas por Karplus em trabalhos anteriores Chandonia e Karplus (1999).

Kono et. al Kono e Doi (1993) descreveram o uso de uma rede de autômatos

para a predição de seqüência e a conformação de cadeias laterais a partir da geometria da cadeia principal. A Biblioteca de rotâmeros utilizada por Kono et al. foi definida por Ponder e Richards para reduzir o grau de liberdade dos rotâmeros Ponder e Richards (1999). Nesse método, um autômato é associado para cada posição do rotâmero, semelhante a associar um rotâmero para cada neurônio em uma rede de Hopfield e minimizar a função de energia da rede neural. As cadeias laterais possuem um papel fundamental na estrutura da proteína, o que torna importante investigar o estudo de técnicas como redes neurais utilizando bibliotecas de rotâmeros na predição de estruturas de proteínas.

Recentemente Cuff e Barton avaliaram de forma mais detalhada como o uso de tipos diferentes de perfis (*profiles)* de alinhamentos múltiplos, a partir das mesmas seqüências, pode melhorar a performance das redes neurais [9]. Nesse trabalho, Cuff e Barton exploraram e avaliaram como o uso, mais sofisticado, do alinhamento múltiplo e das informações podem ser importantes para melhorar o rendimento das redes neurais na predição de estrutura secundárias e propuseram um novo método baseado em redes neurais. Nesse novo método, as informações de alinhamento são mais exploradas do que na maioria dos métodos de 3a geração Cuff e Barton (1999, 2000).

Cuff and Barton projetaram uma base de dados de 369 proteínas para avaliar o rendimento dos algoritmos de predição propostos na literatura: DSC, PHD, NNSSP e PREDATOR, avaliando a taxa de acerto desses métodos Rabow e Scheraga (1993). Rabow and Scheraga descreveram uma aplicação de redes neurais para predição de estruturas dentro de uma rede cúbica. Define-se uma função potencial e a rede neural faz uma busca por conformações de mais baixa energia dentro da rede cúbica. Rabow e Scheraga compararam os resultados obtidos com os resultados obtidos pelo método de Monte Carlo Muskal e Kim (1992). Nesse trabalho, Scherega e Rabow obtiveram resultados melhores com as redes neurais comparado com o método de Monte Carlo. Muskal e Kim também investigaram o uso de redes neurais do tipo MLP na predição de estruturas secundárias Bohr e Bohr (1998). Bohr utilizou redes do tipo MLP para predizer o estado conformacional (predição3D) de pequenos peptídeos a partir de informações sobre a estrutura eletrônica da molécula Rost e Sander (1994).

As Redes Neurais Artificiais são ótimas ferramentas para classificação e reconhecimento de padrões Haykin (1994). Portanto são boas ferramentas para predição

1D. Nesse projeto de pesquisa, são explorados os efeitos do projeto da arquitetura e o uso de várias propriedades físico-químicas (isoladas e combinadas) no desempenho das redes neurais na predição 1D. Procurando aperfeiçoar o desempenho do preditor já existente. Também será desenvolvido um software para disponibilizar o preditor via Web.

3 Materiais e Métodos

3.1 Base de Dados

O primeiro passo, no desenvolvimento do trabalho, foi a coleta e seleção de três conjuntos de dados (seqüências primárias de proteínas). O primeiro conjunto foi selecionado a partir do Protein Data Bank através de alinhamento múltiplo, obtendo 389 proteínas com baixa identidade na seqüência primária. Esse conjunto de treinamento foi dividido em 4 subconjuntos: **Todas** (389 proteínas), o qual possui todas as proteínas coletadas; **Hélice**, que contém proteínas cujo número de resíduos em estrutura hélice é maior que a soma de resíduos em estrutura folha e resíduos em estrutura *coil* (173 proteínas); **Folha**, que possui proteínas cujo número de resíduos em estrutura folha é maior que a soma de resíduos em estrutura hélice e resíduos em estrutura *coil* (56 proteínas); e Hélice-Folha, ou seja, proteínas cuja porcentagem de resíduos em estrutura hélice é superior a 30% e inferior a 50% e a porcentagem de resíduos em estrutura folha é superior a 30% e inferior a 50% (115 proteínas).

As diferentes redes neurais projetadas foram treinadas com as quatro bases de dados diferentes e os testes mostraram que as RNA treinadas com a **base hélice-folha**, constituída de **115** proteínas não homólogas, obteve o melhor desempenho na predição de estruturas secundárias. Portanto, os resultados apresentados nesse artigo são das redes neurais treinadas com o subconjunto hélice-folha. Para validação foi utilizado um conjunto de 75 proteínas descritas em Holley e Karplus (1991).

O conjunto de dados (base de dados) utilizado para testar as redes foi extraído do CASP. Essas proteínas são usadas, como padrão, pelo *Critical Assessment of Structure Prediction* (CASP) na avalição dos métodos de predição de estrutura secundária decsritos na literatura mundial. Para verificar se as proteínas não possuíam um grau alto de identidade na seqüência foi realizado alinhamento múltiplo através do software Clustal-X.

3.2 Arquiteturas Projetadas

Até o momento, não possuímos habilidade suficiente para predizer a estrutura terciária de uma proteína, a partir de sua seqüência. Mas podemos predizer aspectos mais simplificados da estrutura Bohr e Bohr (1998); Rost e Sander (1994). Uma simplificação para o problema é a predição de estruturas de proteínas em uma dimensão (1D), ou seja, a estrutura secundária e/ou a área acessível ao solvente. O nome 1D vem do fato de associarmos a cada aminoácido uma estrutura secundária.

O objetivo usual dos métodos de predição 1D é associar a cada resíduo um padrão estrutural H (α-hélice), E (β-folha) ou L (volta, isto é, uma estrutura não regular), dependente do conjunto de resíduos que lhe são adjacentes, ou seja, de uma janela da seqüência. A principal idéia por trás da maioria dos métodos de predição de estrutura secundária é o fato de que segmentos de resíduos consecutivos possuem uma preferência para certos estados de estrutura secundária. Dessa forma, o problema de predição de estrutura torna-se um problema clássico de classificações de padrões tratável por algoritmos de reconhecimento de padrões. Foram projetadas e implementadas duas arquiteturas diferentes : uma contendo apenas uma rede neural com uma camada intermediária e outra contendo duas redes neurais cada uma com uma camada intermediária. Sendo que, nesse caso, a saída da primeira rede alimenta a entrada da segunda rede neural (ver Figuras 1 e 2). Para o caso da arquitetura com duas redes neurais, a informação de saída da primeira rede neural é adicionada na janela de dados de entrada.

Para cada arquitetura foram implementadas 9 redes MLP com a camada de entrada variando. Todas as redes foram projetadas de maneira a predizer a estrutura secundária do aminoácido que se encontra no meio da janela de entrada. Foram testadas redes com os seguintes tamanhos de janela de entrada (7,9,11,13,15,17,19,21 e 23). Pelos resultados obtidos pode-se perceber que tamanhos diferenciados de janelas de entrada estão diretamente relacionados com a performance da rede na predição da estrutura . Dessa forma, para uma rede neural com uma janela de tamanho 7, temos 154 neurônios na camada de entrada (22 X 7). Foram desenvolvidos dois softwares em C++ para realizar o pré-processamento e o pós processamento da rede neural.

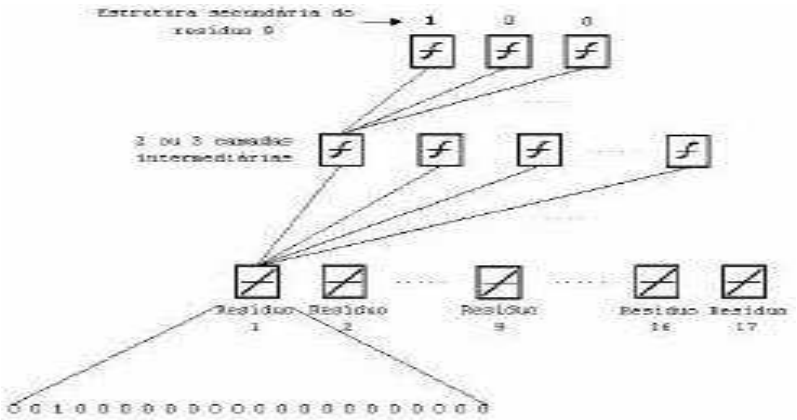

Figura 1: Arquitetura com uma rede neural artificial

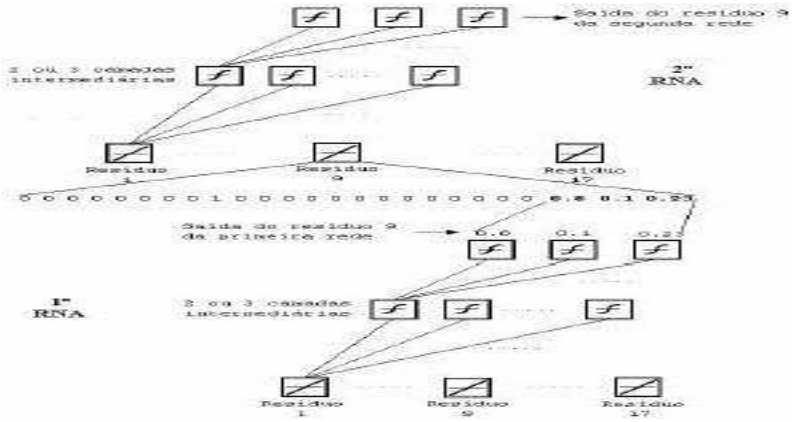

Figura 2: Arquitetura com duas redes neurais artificiais em cascata.

3.3 Coeficientes

Existem alguns coeficientes para verificar o desempenho de uma RNA na predição de estruturas secundárias. Dentre estes, o mais utilizado pelos pesquisadores é conhecido como Q3, que fornece a porcentagem dos resíduos preditos corretamente considerando os três tipos de estruturas secundárias: hélice, folha e coil [5]. O Coeficiente Q3 é dado por :

Onde i = {hélice, folha, coil}, TOTi é o número de "i" existentes nas proteínas de teste, PREDi é número total de "i" que a RNA predisse e CORRi = número de "i" que a RNA predito corretamente. Outros coeficientes importantes e também utilizados no projeto para ajudar na avaliação dos testes são o Q[obs] e Q[pred]. O primeiro dá a porcentagem do número de resíduos preditos corretamente em relação ao número real observado, em um estado particular. Já o Q[prd] dá a porcentagem do número de resíduos preditos corretamente em relação ao número que a RNA predisse em um estado particular. Onde

Para a implementação destes coeficientes que avaliam o desempenho da rede, foi desenvolvido um software, em C^{++} Builder 5, chamado de Comparar.

3.4 Implementação do júri

A inclusão do júri de decisão destina-se a fazer uma leitura da predição final. Na literatura, encontra-se o trabalho de Rost et al. (1993) que utilizam o júri como um 8 filtro, ou seja, que executa uma média aritmética sobre resultados gerados por 12 redes neurais distintas. A partir dessa motivação, desenvolveu-se um *software* chamado 'júri', com a função de realizar a média aritmética sobre os resultados da predição das 18 redes.

4 Resultados

Para efeito de avaliar a qualidade do preditor desenvolvido, batizado com o nome de PREDCASA Foram selecionadas 15 proteínas do CASP para testar o preditor desenvolvido e comparar seu desempenho com 3 preditores disponíveis na literatura e bastante utilizados por pesquisadores da área. Na Tabela 1 é apresentada a comparação do preditor PREDCASA com o PSIPRED, o PHD e o PSA. Percebe-se

uma certa regularidade de percentual de acerto de 47% até 84% para o PREDCASA em relação aos outros. Os resultados apresentados são apenas da arquitetura com duas redes neurais em cascata e utilizando-se o coeficiente Q3.

Tabela 1: Análise das médias de acerto

Proteína	PREDCASA (% de acerto)	PHD (% de acerto)	PSA (% de acerto)	PSIPRED (% de acerto)
1QLQ	84	91	51	87
1EIG	73	86	75	91
1C56	47	67	37	50
1DAQ	85	70	66	78
1EHD	52	55	58	59
1E5B	60	65	63	72
1EJG	80	50	58	67
1ES1	70	74	56	78
1DT4	64	71	63	78
1EDS	61	29	41	38
1G6X	84	91	53	91
1DO1	61	60	54	66
1FD8	79	79	27	84
1FE5	63	66	67	86
MÉDIA	69.13	67,13	55,66	73,4

5 Conclusões

Os resultados comprovam a eficiência do preditor com janelas distintas. A média do acerto do PREDCASA é de 69,13%, esse valor é superior ao do PHD que obteve uma média de acerto de 67,13%, perdendo apenas para o PSIPRED com 73,4%. A regularidade de acerto do PREDCASA é de 50 a 85% enquanto o PSIPRED obteve uma regularidade de 38 a 91%. Portanto a média atingida pelo preditor PREDCASA comprova a importância de redes treinadas com janelas diferentes e a implementação de um júri é extremamente importante para a performance das redes neurais.

Pode-se notar portanto que o projeto bem elaborado de uma base de dados para treinamento das redes e das arquiteturas das redes neurais é de extrema importância para o problema de predição de estrutura secundária de proteínas. É importante ressaltar que se deve estudar, de maneira mais profunda, a influência dos diferentes tamanhos de janela de entrada para esse tipo de problema. Deve-se enfatizar que o PREDCASA é o primeiro preditor de estrutura secundária de proteínas desenvolvido no Brasil.

6 Contribuições e Trabalhos Futuros

Esse trabalho teve como objetivo utilizar redes neurais como processos de otimização na predição de estruturas secundárias de peptídeos e proteínas. Como contribuições desse trabalho pode-se mencionar:

- estudo de como diferentes bases de dados (toda-α, toda -β, $\alpha/\beta\mathfrak{g}$ mistas) podem auxiliar e melhorar a performance de redes neurais artificiais.
- observação que a arquitetura da rede e principalmente a codificação problema são fatores limitantes na performance da predição 1D através de redes neurais do tipo MLP. E a observação que o fato de incluir mais informações para a rede MLP não implica em um acréscimo na performance da mesma.
- confirmação de que a performance das redes neurais, nesse tipo de problema, depende do tamanho da janela de entrada.

Como futuros trabalhos podemos citar: O aperfeiçoamento do sistema de júri, informações evolutivas, projeto e teste de novas bases de dados, testes de novas arquiteturas de redes neurais e o uso de algoritmos genéticos em conjunto com as redes neurais, investigar a influência do tamanho da janela de entrada e o tipo de proteína.

Agradecimentos

A Fapesp pelo apoio financeiro.

Referências

Bohr, H. e Bohr, J. (1998). Protein secondary structure and hology by neural networks. *FEBS Letters*, 241:223–228.

Chandonia, J. M. e Karplus, M. (1996). The importance of larger data sets for protein secondary structure prediction with neural networks. *Protein Science*, 5:768–774.

Chandonia, J. M. e Karplus, M. (1999). New methods for accurate prediction of protein secondary structure. *PROTEINS:Structure, Function and Genetics*, 35:293–306.

Cuff, J. A. e Barton, J. G. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *PROTEINS: Structure, Function and Genetics*, 34:508–519.

Cuff, J. A. e Barton, J. G. (2000). Application of multiple sequence alignment profiles to improve protein secondary structure preditor. *PROTEINS: Structure, Function and Genetics*, 40:502–511.

Haykin, S. (1994). *Neural Networks: a comprehensive foundation*. IEEE Press, New York.

Holley, L. H. e Karplus, M. (1991). Neural networks for protein structure prediction. *Methods in Enzymology*, 202:204–224.

Kono, H. e Doi, J. (1993). Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Protein Science*, 19:244–255.

Muskal, M. S. e Kim, H. S. (1992). Predicting protein secondary structure content a tandem neural network approach. *Journal of Molecular Biology*, 225:713–727.

Ponder, J. W. e Richards, F. M. (1999). Tertiary templates for protein use packing criteria in the enumeration of allowed sequences for different structural classes. *Journal of Molecular Biology*, 193:775–791.

Qian, H. e Sejnowski, M. (1996). Prediction of helix in proteins based on thermodynamic parameters from solution chemistr. *Journal of Molecular Biology*, 256:663–666.

Rabow, A. A. e Scheraga, A. H. (1993). Lattice neural network minimization:application of neural network optimization for locating the global-minimum conformations of proteins. *Journal of Molecular Biology*, 232:1157–1168.

Rost, B. (1998). Protein structure prediction in 1d, 2d and 3d. *The Encyclopedia of Computational Chemistry*, 3:2242–2255.

Rost, B. e Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *PROTEINS: Structure, Functions and Genetics*, 19:55–72.

Rost, B., Schneider, R., e Sander, C. (1993). Progress in protein structure prediction? *TIBS*, 18:120–123.