ER³MAC
U·F·P·E

Rede Neural Recorrente Bidirecional Enriquecida por Grafos de Interações para Predição de Estruturas Secundárias de Proteínas

Ryan Ribeiro de Azevedo, Francisco do Nascimento Júnior, Fred Freitas, Tsang Ing Ren, Edson C. de Barros. C. Filho

Centro de Informática – Universidade Federal de Pernambuco (CIN-UFPE)

Av. Prof. Luiz Freire, s/n Cidade Universitária 50740-540 - Recife, PE - Brasil

{ rra2, fnj, fred, tir, ecdbcf}@cin.ufpe.br

Resumo: O problema de predição de estruturas secundárias de proteínas tem sido um grande desafio para a Biologia Computacional e inúmeras abordagens de soluções já foram propostas. Redes Neurais Recorrentes são apresentadas como uma destas soluções por se tratar de dispositivos que conseguem aprender curtas e longas seqüências de proteínas. O presente artigo apresenta uma extensão das Redes Neurais Recorrentes envolvendo informações do contexto passado, futuro e grafos de interações representando dependências entre posições da seqüência, formando a Rede Neural Recorrente Bidirecional Enriquecida por Grafos de Interações.

Palavras-chave: Redes Neurais Recorrentes, Biologia Computacional, Predição de Estruturas Secundárias de Proteínas.

1. Introdução

Nos últimos anos, o projeto de seqüências de proteínas artificiais para alcançar estruturas terciárias e secundárias tem recebido considerável atenção. Esse processo objetiva não apenas elucidar o problema do enovelamento de proteínas, mas também produzir proteínas que possuam uma estrutura desejada. Para isso métodos como Algoritmos Genéticos, Redes Neurais Artificiais e Simulações com Monte Carlo vêm sendo utilizados a fim de predizer estruturas primárias de aminoácidos que levem às estruturas secundárias ou terciárias desejadas.

A determinação da estrutura secundária de proteínas a partir da sua seqüência de aminoácidos é importante para a engenharia de proteínas e o desenvolvimento de novos fármacos. Uma alternativa para este problema tem sido a aplicação de técnicas de RNAs - Redes Neurais Artificiais. As abordagens utilizando RNAs tem obtido resultados relevantes, porém estão restritas a pequenas proteínas, com dezenas de aminoácidos e a algumas classes de proteínas. Este trabalho propõe a investigação de uma abordagem utilizando RNAs para a predição da estrutura secundária de proteínas independentemente do seu tamanho e classe. Os resultados obtidos demonstram que apesar das dificuldades encontradas a abordagem investigada constitue-se em uma alternativa em relação aos métodos clássicos de determinação da estrutura secundária das proteínas.

A proposta apresentada e implementada neste artigo é um método para aprendizagem seqüencial de dados relacionais, realçando o conhecimento que longas e curtas dependências interferem na saída obtida. O método foi utilizado no problema de predição de estruturas secundárias de proteínas, a partir de uma seqüência de entrada formada por aminoácidos, inicialmente, e estendida acrescentando uma informação de interação entre aminoácidos não necessariamente adjacentes, representada por mapas de contatos. A solução final apresentada é uma implementação de uma rede neural recorrente bidirecional enriquecida de um grafo de interações, utilizando o *Backpropagation* para a aprendizagem.

O artigo está estruturado da seguinte forma: a Seção 2 trata do modelo de Rede Neural Artificial utilizado para a solução, de seu processo de aprendizagem e sua arquitetura, a Seção 3 foca o problema de predição de estruturas secundárias de proteínas e sua motivação, na Seção 4 são apresentados os dados utilizados nos experimentos e o modo como foram obtidos, na Seção 5, os resultados parciais com os experimentos são exibidos e, por fim, a Seção 6, conclui e apresenta os trabalhos futuros.

2. Redes Neurais Recorrentes Bidirecionais Enriquecidas por Grafos de Interação

Redes Neurais Recorrentes (RNR) são estruturas de processamento capazes de representar uma grande variedade de comportamentos dinâmicos. A presença de realimentação de informação permite a criação de representações internas e dispositivos de memória capazes de processar e armazenar informações temporais e sinais seqüenciais [12]. A presença de

conexões recorrentes ou realimentação de informação pode conduzir a comportamentos complexos, mesmo com um número reduzido de parâmetros.

Redes Neurais Recorrentes são processos dinâmicos causais, ou seja, são sistemas que utilizam conhecimento obtido no passado para inferir a saída atual. No contexto do problema em questão, há uma necessidade de olhar para as futuras posições da entrada, ou seja, incluir o conhecimento futuro para se obter um melhor resultado na classificação.

De acordo com [4] e [12], na aprendizagem de Redes Neurais Recorrentes, ocorre um problema relativo ao treinamento das redes para produzir uma resposta desejada no tempo corrente que depende dos dados de entrada no passado distante, chamado de problema da extinção dos gradientes.

Uma solução indicada por [6] e [7], apresenta um grafo de interações para adicionar informações de dependências entre posições distantes. Este grafo é obtido a partir dos mapas de contatos gerados da estrutura primária da proteína, representando as possíveis interações entre determinadas posições da seqüência.

2.1 Aprendizagem Supervisionada Seqüencial

Para a entrada da RNA – Rede Neural Artificial proposta utilizou-se um par de seqüências de entrada $(\boldsymbol{x}, \boldsymbol{y}) = (\{x[1], x[2], x[3],... x[N]\}, \{y[1], y[2], y[3],... y[M]\})$ onde $x[t] \in \boldsymbol{x}$, t=1, 2, 3..., N, $y[s] \in \boldsymbol{y}$, s =1, 2, 3..., M, por uma amostragem fixa e desconhecida $p(\boldsymbol{x}, \boldsymbol{y})$. No padrão de aprendizado supervisionado seqüencial, os dados são apresentados em uma série de pares de entrada e saída $\boldsymbol{Dm} = \{(\boldsymbol{x_i}, \boldsymbol{y_i})\}^{\boldsymbol{m}}_{\boldsymbol{i}=1}$, onde cada $\boldsymbol{x_i}$ é uma seqüência de entrada de tamanho $\boldsymbol{N_i}$ e $\boldsymbol{y_i}$ é a seqüência de saída correspondente de tamanho $\boldsymbol{M_i}$, ou seja, os pares de seqüências de entrada e saída possuem o mesmo tamanho $\boldsymbol{N_i} = \boldsymbol{M_i}$ denotados por $f(\boldsymbol{x})[t]$ o t-th elemento de $f(\boldsymbol{x_i})$ para $t = 1, 2, 3, ..., N_i$.

2.2 Aprendizagem Supervisionada Seqüencial Enriquecida com Grafos de Interação

Grafos de Interação são utilizados agora junto com as seqüências de entrada $(\boldsymbol{x}, \boldsymbol{y})$ citadas na Seção 2.1, mapeadas como pares da seguinte forma: $(\boldsymbol{x}, \boldsymbol{G})$ onde $\boldsymbol{x}$ permanece sendo uma seqüência de entrada e $\boldsymbol{G} = (V, E)$ um grafo não-dirigido cujo conjunto de vértices é $V = \{1, 2, 3,..., N\}$ e as arestas E representam as interações, isto é, $\{t, s\} \in E$ se e somente se as posições das arestas t e s interagirem.

2.3 Arquitetura da Rede Proposta

A arquitetura de uma RNR pode ser projetada a partir do desdobramento dos estados em redes *feedforwards*. No caso da RNA proposta, temos uma RNRB – Rede Neural Recorrente Bidirecional, a qual é composta por estados *forward* e *backward*, responsáveis, respectivamente por manter os contextos do passado e do futuro. Uma possível implementação destes contextos é através de redes *feedfowards*, compartilhando os pesos em cada passo de tempo. E ainda, compondo a RNRB, encontra-se uma outra rede *feedfoward* para unir as saídas dos contextos e a entrada da rede e, assim, inferir na predição da classe da respectiva entrada. É apresentada na Figura 1 a arquitetura da RNA proposta.

Esta arquitetura foi estendida, a fim de receber o grafo de interações citados na seção anterior, juntamente com a seqüência de entrada para melhorar a qualidade da informação obtida pelas funções de transição de estado, especificamente para conseguir resolver o problema da perda da aprendizagem a longas dependências. Formando, então, a RNRBEI – Rede Neural Recorrente Bidirecional Enriquecida por Grafos de Interações.

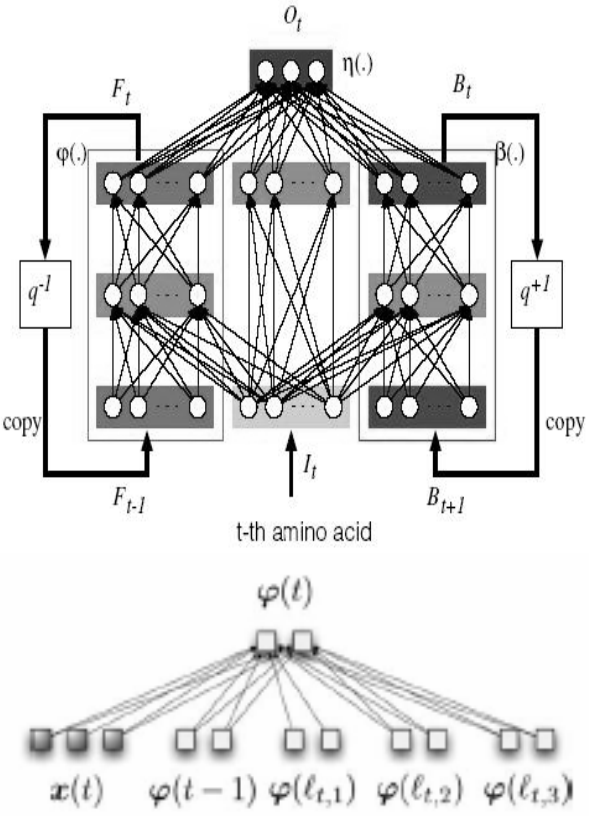

Figura 1. Arquitetura da RNA proposta.

A RNRBEI é baseada nas seguintes dinâmicas não-casuais:

$$\varphi[t] = \mathcal{T}_{\mathrm{F}}(x[t], \varphi[t-1], \varphi[\ell_{t,1}], \varphi[\ell_{t,2}], \ldots, \varphi[\ell_{t,K}]; \boldsymbol{\theta}_{\mathrm{F}})$$

$$\boldsymbol{\beta}[t] = \mathcal{T}_{\mathrm{B}}(x[t], \boldsymbol{\beta}[t+1], \boldsymbol{\beta}[r_{t,1}], \boldsymbol{\beta}[r_{t,2}], \ldots, \boldsymbol{\beta}[r_{t,K}]; \boldsymbol{\theta}_{\mathrm{B}})$$

A qual recebe como parâmetro a entrada, o valor da função do estado anterior (ou posterior) e o valor da função aplicada a vértices adjacentes. Utilizando as seguintes condições limitantes:

$$\boldsymbol{\varphi}[\mathbf{O}] = \boldsymbol{\beta}[\mathcal{N} + \mathbf{1}] = \mathbf{O}.$$

Estas funções recursivas paramétricas dos contextos de passado e futuro possuem pesos ajustáveis através de treinamento das redes MLP – *Multilayer Perceptron* que as implementam. Após o computo dos valores das funções de contexto, pode-se obter o valor da função de saída, que também é uma função paramétrica:

$$f(\boldsymbol{x})[t] = \eta(\boldsymbol{\varphi}(\boldsymbol{x})[t], \boldsymbol{\beta}(\boldsymbol{x})[t]; \boldsymbol{\theta}_{Y})$$

Para classificar, foi utilizada a função *softmax*, para Q classes, a fim de se obter uma distribuição de probabilidade para a saída [6] [7]. Para a aprendizagem da RNA proposta foi utilizado o algoritmo do *Backpropagation* a fim de propagar o erro, calculado a partir da diferença entre a saída obtida e a desejada, por todas as camadas até a entrada.

2.4 Implementação

O sistema foi implementado na linguagem de programação *Java*. É apresentada na Figura 2 uma representação em blocos do fluxo dos dados, destacando as camadas da arquitetura e as conexões. Alguns parâmetros devem ser configurados no momento da execução, são eles:

- NB – Número de neurônios da camada de saída do contexto *backward*.
- NBH – Número de neurônios da camada oculta do contexto *backward*.
- NF – Número de neurônios da camada de saída do contexto *forward*.
- NFH – Número de neurônios da camada oculta do contexto *forward*.
- Y – Número de neurônios da camada de saída da rede central.
- NYH – Número de neurônios da camada oculta da rede central.
- K – Número de neurônios da camada de entrada da RNA principal.
- Taxa – Taxa de aprendizagem
- Usar grafo – Define se os grafos de interação serão ou não utilizados.

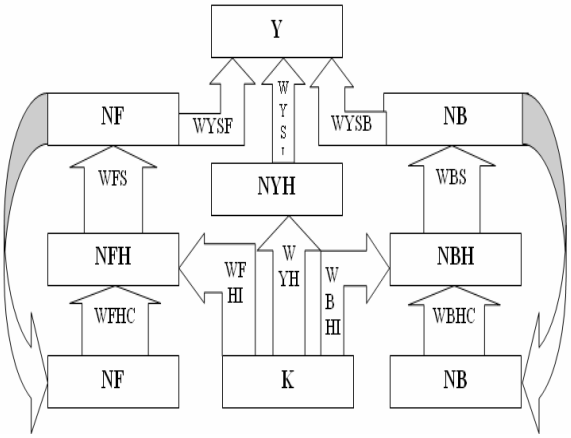

Figura 2. Representação do fluxo de dados da RNA proposta.

3. Predição de Estruturas Secundárias das Proteínas

As proteínas são polímeros de unidades monoméricas ligadas por ligações peptídicas, contendo um grupo amina (NH_2), um grupo ácido carboxílico (COOH) e um radical R ligado a um átomo de carbono C. Os monômeros constituídos desta forma são denominados aminoácidos ou resíduos [8] [19]. As proteínas são compostas por 20 aminoácidos distintos formando uma estrutura primária de aminoácidos. Segundo [6], a função biológica de uma proteína depende do seu *fold,* ou estrutura terciária, e a determinação experimental destas estruturas é um processo custoso e realizado atualmente por métodos de Cristalografia por Difração de Raios-X ou por Ressonância Magnética Nuclear. Snow [27] define uma proteína como sendo: uma cadeia de aminoácidos ligados covalentemente, entre o grupo amina de um determinado aminoácido e o grupo carboxila do aminoácido seguinte, formando, assim, ligações peptídicas que definem seu esqueleto conforme a sua liberdade rotacional.

As proteínas apresentam diferentes níveis de estruturas, estrutura primária que consiste de uma seqüência de aminoácidos que formam a cadeia polipeptídica, estrutura secundária que refere-se à organização local de partes da cadeia polipeptídica formando diferentes estruturas estáveis conhecidas como *α-hélices*, *β-strands* e *γ-coils*.

Atualmente são conhecidas aproximadamente 2 milhões de seqüências de proteínas não redundantes e

apenas 30.000 destas seqüências armazenadas no PDB – *Protein Data Bank* possuem suas estruturas terciárias conhecidas, resultando numa taxa de 60:1, e que continua a crescer [5] [7]. Métodos de predição de estruturas secundárias são propostas bastante relevantes, obtidas às estruturas secundárias pode-se assim obter as estruturas terciárias nas quais poderão auxiliar os pesquisadores em pesquisas que empregam Algoritmos Genéticos, fornecer subsídios para a compreensão dos processos biológicos, como os mecanismos de reações enzimáticas, da ação de anticorpos, transporte de oxigênio, transporte através de membranas e no planejamento de novas drogas e vacinas.

Segundo [2] os métodos de predição são divididos basicamente em três classes principais: modelagem comparativa ou por homologia (consiste em utilizar uma ou mais proteínas homólogas e de estrutura terciárias já conhecidas), *ab-initio* que consiste em obter a estrutura terciária a partir de sua estrutura primária necessitando de um maior poder computacional e por métodos experimentais citados anteriormente (Ressonância Magnética Nuclear e Cristalografia por Difração de Raios-X).

3.1 Estado da Arte em Predição de Estruturas Secundárias

O desenvolvimento de preditores baseados em Redes Neurais Artificiais foi iniciado por [24] e refinado durante os anos 90 por [25] incorporando informações evolucionárias na forma de perfis de alinhamentos múltiplos, considerada uma técnica que melhorou significativamente a precisão da predição de estruturas secundárias [6] [7]. Pode-se citar também os trabalhos de [13] que utilizaram uma RNA do tipo MLP e codificaram os dados de entrada da rede em janelas de resíduos adjacentes e [8] que aplicaram duas Redes Neurais Artificiais, denominadas primária e secundária utilizando um conjunto de 681 proteínas com estruturas disponíveis no PDB.

Também em 1996 [25] introduziu um número significante de melhorias arquiteturais, tal como a utilização de um pós-processador para filtrar erros de predição local. [15] sugeriu o uso de matrizes específicas de posição para informações evolucionárias que foram incorporadas por [26] e obteve resultados satisfatórios em termos de predição de estruturas. De acordo com [6] e [7] preditores baseados em HMMs – *Hidden Markov Models* [16] e Redes Neurais Artificiais Recorrentes Bidirecionais [2] [23] também foram introduzidas levando a resultados, porém, nenhum avanço radical no estado da arte.

Segundo [6] e [7] uma medida de performance utilizada para avaliar a precisão dos preditores na predição de estrutura secundária é o ***Q3*****,** definido pela formula abaixo:

$$Q3 = \sum Ni / NT$$

Onde *Ni* corresponde ao número total de resíduos identificados corretamente para cada classe e *NT* é o número total de resíduos existentes na proteína, o melhor preditor disponível atualmente apresenta um nível ***Q3*** entre 70 e 79%.

3.2 Predição de Estruturas Secundárias e Mapas de Contato

A utilização de mapas de contatos no auxilio da predição de estruturas secundárias é bastante relevante devido à quantidade de informações de longas escalas entre as seqüências. Os mapas de contato representam graficamente a relação de vizinhança espacial entre os aminoácidos, onde, um lado **{*t, s*}** em um mapa de contato indica que a distância entre um átomo **C-α** dos resíduos na posição ***t*** e ***s*** é menor que o *threshold* predefinido. É apresentado na Figura 3 um exemplo de mapa de contatos.

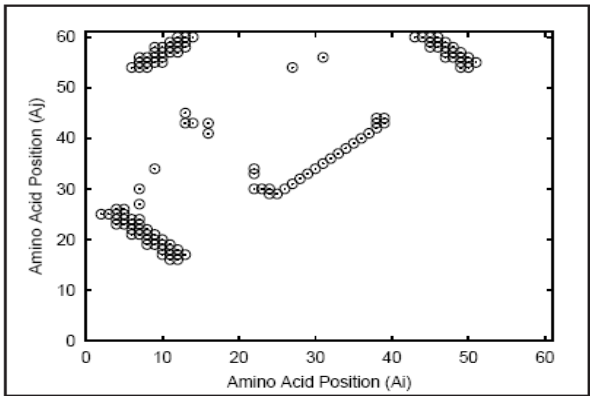

Figura 3. Representação dos mapas de contatos gerados pelo CMAPpro

De acordo com [7] há várias razões de se utilizar informações obtidas dos mapas de contato para predição de estruturas secundárias, tais como: mapas de contato podem ser obtidos a partir de seqüências primárias ou obtidos de estruturas preditas pelo método *ab-initio*; predições podem ser melhoradas pela utilização de informações a respeito das distâncias entre os resíduos de uma seqüência primária de aminoácidos, informações essas obtidas dos mapas de contato; mapas preditos podem conter informações úteis para melhorar a precisão das propriedades de ordem mais baixa das estruturas secundárias.

4. Dados Utilizados nos Experimentos

Os dados utilizados nos experimentos são cadeias não-homólogas extraídas do PDB – *Protein Data Bank*. As seqüências extraídas fazem parte da versão de 18 de Agosto de 2006 [17]. Foram obtidas 195 seqüências de entrada, suas saídas desejadas e seus mapas de contato podendo ser obtidas através do endereço:ftp://ftp.cmbi.kun.nl/pub/molbio/data/pdbfinder2/.

Deste conjunto, separou-se 32 seqüências de entrada para validação, 30 para teste e o restando ficando para o treinamento. O conjunto de validação foi utilizado para critério de parada, sendo validado a cada 5 iterações de treinamento, verificando o valor do erro, e caso fosse identificado um crescimento durante 3 observações consecutivas, seria considerado uma possível perda de generalização, e então, o treinamento seria finalizado.

As 8 classes do DSSP foram reduzidas para 3 classes principais da seguinte forma: H para as *α-hélices***,** E para as *β-strands* e B, C, I, S, T e G para *γ-coils* mapeadas na saída da RNA aqui proposta por H, E e Y respectivamente.

Foi realizado um pré-processamento nas seqüências primárias obtendo assim apenas as seqüências com alta qualidade, utilizou-se o programa DSSP para obter-se a saída desejada das seqüências primárias retidas no pré-processamento, e os mapas de contatos foram definidos usando um *threshold* de 8 *Angstrom* e obtidos a partir de http://www.ics.uci.edu/~baldig/index.html [14] e de http://gpcr.biocomp.unibo.it.

5. Resultados

Nesta seção são apresentados os resultados iniciais obtidos durante a predição das estruturas secundárias das proteínas relacionadas na fase de pré-processamento para a classe α-*hélices* considerada a mais relevante das três classes principais. Os resultados obtidos pela RNRBEI são resultados pré-liminares que avançarão com o andamento da pesquisa realizada.

5.1 Predição de Estruturas Secundárias a partir dos Mapas de Contatos e das Seqüências

A arquitetura das RNAs utilizadas para o treinamento, teste e validação possuem as seguintes configurações apresentadas na Tabela 1. A taxa de aprendizagem foi fixada em 0,01 e 0,05 e os pesos iniciais escolhidos de forma aleatória. Os resultados obtidos nos experimentos em percentagem de nível de precisão para a classe *α-hélice* utilizando à configuração onde a taxa de aprendizado foi fixada em 0,05 obteve uma melhor performance chegando a classificar corretamente em 70,93%, também em percentagem de erro, a RNRBEI classificou 7,45% como *β-strands e* 21,62% como *γ-coils* quando deveria classificar como *α-hélice*

Tabela 1 – Parâmetros configurados para cada rede testada.

Configuração das RNRBEI Utilizadas nos experimentos							
Estado *backward*		Estado *forward*		Rede Principal		Saída e Taxa de Aprendizado	
NB	**NBH**	**NF**	**NFH**	**K**	**NYH**	**Y**	**TA**
10	15	10	15	20	15	3	0,01
NB	**NBH**	**NF**	**NFH**	**K**	**NYH**	**Y**	**TA**
5	10	*5*	10	20	15	3	0,05

6. Conclusão e Trabalhos Futuros

Os resultados obtidos indicam que a RNRBEI melhora a predição quando exploram as informações contidas nos mapas de contatos associados a sua respectiva estrutura primária, porém, mesmo com a exploração dos mapas de contato associados a sua estrutura primária, permanece uma percentagem de erro.

O primeiro protótipo do modelo de RNA aqui proposto encontra-se implementado e com os resultados considerados dentro do que os preditores atuais apresentam atualmente. O segundo protótipo esta em andamento e poderá incorporar algumas funcionalidades a fim de se obter melhorias na performance do modelo, tais funcionalidades são:

- Utilização de alinhamentos múltiplos nas seqüências de entrada;
- Homologia abaixo dos 25%;
- Analise de outros algoritmos de aprendizagem (BPTT – *Backpropagation Through Time* e filtros de *Kalman*);
- Método de Máxima Verosemelhança;

Espera-se com a finalização do segundo protótipo chegar a resultados de classificação corretos para as três classes principais: ***α-hélice***, ***β-strands***, ***γ-coils*** acima dos 80% de acertos e o nível de performance ***Q3*** também acima de 80%. Além das novas funcionalidades que encontra-se em fase de implementação será utilizado uma amostragem de dados maior com aproximadamente 1200 seqüências primárias divididas em três conjuntos (Treinamento, Teste e Validação) para que seja considerado um bom treinamento e que seja possível obter os resultados esperados.

Agradecimentos

Ao CIN-UFPE – Centro de Informática da Universidade Federal de Pernambuco pelo apoio, incentivo e oportunidade da pesquisa realizada e ao Professor Aluízio Fausto Ribeiro Araújo.

Referências

[1] Baldi, P., & Brunak, S. 2001. *Bioinformatics: The machine learning approach (2nd ed.)*. Cambridge, MA: MIT Press.

[2] Baldi, P., Brunak, S., Frasconi, P., Soda, G., & Pollastri, G. 1999. *Exploiting the past and the future in protein secondary structure predicton*. Bioinformatics, 15, 937–946.

[3] Baldi, P., & Pollastri, G. 2003. *The principled design of large-scale recursive neural network architectures-dag-rnns and the protein structure prediction problem*. Journal of Machine Learning Research, 4(Sep), 575–602.

[4] Bengio, Y., & Frasconi, P. 1996. *Input–output HMM's for sequence processing*. IEEE Transactions on Neural Networks, 7(5), 1231–1249.

[5] Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., et al. 2000. *The protein data bank*. Nucleic Acids Research, 28, 235–242.

[6] Ceroni, Alessio, Frasconi, Paolo, Pollastri, Gianluca, 2005. *Learning protein secondary structure from sequential and relational data*. Neural Networks, 18.

[7] Ceroni, 2005. *Prediction of Structure and Function of Proteins and Ligands By Means of Neural and Kernel Methods for Strutured Data*. Alessio Ceroni. Dissertation submitted in partial fulfillment of the requirements for the degree of doctor of Philosophy in computer Science and Control Engineering. Università Degli Studi di Firenze. 2004-2005.

[8] Chandonia, J. M & Karplus, M, 1996. *The Importance of Larger Data Sets for Protein Secondary Structure Prediction with Neural Networks*. Protein Science.

[9] DILL, K, A et all, 1995. *Principles of Protein Fold. A perspective Form Simple Exact Models*. Protein Science.

[10] Frasconi, P., Gori, M., & Sperduti, A, 1998. *A general framework for adaptive processing of data structures*. IEEE Transactions on Neural Networks, 9(5), 768–786.

[11] Gianluca Pollastri, Pierre Baldi, Pietro Fariselli e Rita Casadio, 2001. *Improved prediction of the number of residue contacts in proteins by recurrent neural networks*. Bionformatics. March 21,.

[12] Haykin, Simon, 2004. Redes Neurais. Princípios e prática.. 2ª Edição..

[13] Holley, L. H & Karplus M, 1991. *Neural Networks for Protein Structure Prediction*. Methods in Enzymology.

[14] J. Cheng, A. Randall, M. Sweredoski, P. Baldi, 2005. *SCRATCH: a Protein Structure and Structural Feature Prediction Server, Nucleic Acids Research*, Web Server Issue, vol. 33, w72-76.

[15] Jones, D, 1999. *Protein secondary structure prediction based on positionspecific scoring matrices*. Journal of Molecular Biology, 292, 195–202.

[16] Karplus, K., Barrett, C., & Hughey, R, 1998. *Hidden markov models for detecting remote protein homologies*. Bioinformatics, 14, 846–856.

[17] Krieger E. RWW Hooft, S Nabuurs, G Vriend, 2004. *A database for protein structure analysis and prediction* – Submitted.

[18] Lombardo, V., & Frasconi, P, 2003. *Learning firstpass structural attachment preferences with dynamic grammars and recursive neural networks*. Cognition, 88(2), 133–169.

[19] Mathews, C, K, Van Hold, K, E, 1990. *Biochemistry*. Redwood City: Benjamin/Cummings Publishing Company Inc.

[20] Mike Schuster and Kuldip K. Paliwal, 1997. *Bidirectional Recurrent Neural Networks*. Member, IEEE. IEEE Transactions on Signal Processing, Vol 45, no. 11. November.

[21] Pierre aldi, Soren Brunak, Paolo Frasconi, Giovanni Soda and Gianluca Pollastri, 1999. *Exploiting the past and the future in protein secondary structure prediction*.. Bioinformatics, vol 15, Nov.

[22] Pollastri, G., & Baldi, P, 2002. *Prediction of contact maps by giohmms and recurrent neural networks using lateral propagation from all four cardinal corners*. Bioinformatics, 18(Suppl. 1), S62–S70.

[23] Pollastri, G., Przybylski, D., Rost, B., & Baldi, P, 2002c. *Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles*. Proteins, 47(2), 228–235.

[24] Qian, N., & Sejnowski, T. J, 1988. *Predicting the secondary structure of globular proteins using neural network models*. Journal of Molecular Biology, 202, 865–884.

[25] Riis, S. K., & Krogh, A, 1996. *Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments*. Journal of Computational Biology, 3, 163–183.

[26] Rost, B., & Sander, C, 1993. *Improved prediction of protein secondary structure by use of sequence profiles and neural networks*. Proceedings of the National Academy of Sciences of the United States of America, 90(16), 7558–7562. Sturt, P., Costa, F.

[27] Snow, M. E, 1992. Powerfull Simulated-Annealing Algorithm Locates Global Minimum of Protein-Fold. Potentials from Multiple Starting Conformation. J Comp. Chem.