Uma Arquitetura de Redes Neurais Artificiais *MLP* para Predição de Estruturas Secundárias de Proteínas

Emerson Cordeiro Morais, Sandro Pereira Vilela, Rubem Mondaini
Universidade Federal do Rio de Janeiro, UFRJ, Centro de Tecnologia, COPPE,
Ilha do Fundão, 21.945-972, P. O. Box 68511, Rio de Janeiro, Brasil.
E-mail: emersonc, sandro, mondaini @cos.ufrj.br

Resumo*: Neste trabalho propõe-se a utilização de reconhecimento de padrões e redes neurais artificiais MLP na predição de estruturas secundárias de proteínas. Parte-se das últimas melhorias de trabalhos anteriores de forma estanque, a saber: variação do tamanho de janelas, utilização de informações evolucionárias, utilização de júri de decisão e criação de redes neurais em cadeia, e propõe reunir todas estas últimas melhorias em uma ferramenta única; realizar testes com um conjunto de proteínas já testado por trabalhos anteriores; e realizar a comparação da ferramenta do trabalho com outros classificadores já disponíveis na web, todos originados de trabalhos acadêmicos, para posteriormente, a partir da arquitetura proposta, realizar um processo de otimização dos parâmetros da rede neural.*

1. Introdução

O método de predição de estruturas secundárias de proteínas utilizado neste trabalho classifica resíduos adjacentes em padrões do tipo *H* (α-hélices), *F* (folhas-β) e *C* (*coil* ou fita aleatória). Um ponto importante dos métodos de predição de estrutura secundária é o fato de que segmentos de resíduos consecutivos possuem uma preferência por certos estados de estrutura secundária. Então o problema de predição de estrutura torna-se um problema clássico de classificações de padrões, que é tratável por algoritmos de reconhecimento de padrões, por exemplo, redes neurais artificiais.

O objetivo do trabalho foi construir a *PSSPA* (*Protein Secondary Structure Prediction Architecture*), uma rede neural artificial *MLP* convencional, isto é, com inicializações aleatórias, utilizando todas as melhorias apresentadas em trabalhos mais recentes, como: variação do tamanho de janelas, utilização de informações evolucionárias, utilização de júri de decisão e criação de redes neurais em cadeia, onde a saída da rede anterior é utilizada como entrada na arquitetura de rede posterior. A rede *PSSPA* foi treinada e testada com um conjunto específico de proteínas já utilizadas em trabalhos anteriores. Estes primeiros resultados serão comparados com os principais preditores apresentados em outros trabalhos e disponíveis na *web*.

2. Materiais e Métodos

Este trabalho concentra-se em métodos estatísticos de reconhecimento de padrões, que é uma das abordagens mais populares e bem conhecidas. Basicamente, um sistema de reconhecimento estatístico de padrões pode ser composto pelas seguintes partes [3][7]: um sistema de *aquisição de dados*; um sistema de *pré-processamento*, para eliminar ruídos ou distorções; um *extrator de características* (ou atributos), que cria um vetor de características com dados extraídos dos objetos adquiridos, reduzindo os dados a atributos, propriedades ou características; um *seletor de características*, que analisa o conjunto de características e elimina as mais redundantes; e um *classificador*, que analisa um padrão obtido e toma uma certa decisão.

O classificador toma decisões baseando-se no aprendizado realizado a partir de um conjunto de treinamento, o qual contém exemplos de padrões de todas as classes existentes no sistema. Em reconhecimento estatístico de padrões, a classificação é realizada utilizando estimativas de distribuições probabilísticas, por isso o nome dessa abordagem. O reconhecedor de padrões é avaliado através de um conjunto de testes, preferencialmente composto por padrões de todas as classes, mas que não estejam no conjunto de treinamento. Além do classificador, o

pré-processamento, o extrator e o seletor de características podem ser dependentes dos dados de treinamento. Para este trabalho foi escolhido um classificador não-linear, a rede neural *MLP* (*MultiLayer Perceptron*), com treinamento realizado através do algoritmo *Backpropagation* que ajusta os pesos através de:

$$w_{ji}(t+1) = w_{ji}(t) + \eta\delta_j(t)x_i(t) \quad \textbf{(1)}$$

Neste trabalho, o primeiro passo para a implementação do projeto foi a coleta e a seleção de um conjunto de seqüências primárias de proteínas para utilizá-lo no treinamento da rede neural. Este conjunto de proteínas foi selecionado a partir do *PDB – Protein Data Bank* através de alinhamento múltiplo. Foram selecionadas ao todo 106 proteínas, destas 102 provenientes de [10] e outras 14 descritas nos testes de [6], com 10 destas fazendo interseção entre as duas bases e totalizando assim 106 proteínas diferentes que são responsáveis pelo treinamento das diferentes arquiteturas de *RNA*.

A maioria das proteínas apresenta mais do que uma cadeia e neste trabalho, foram utilizadas todas as cadeias de todas as proteínas, totalizando 170 seqüencias de aminoácidos, 32.049 resíduos, sendo classificados pelo software *DSSP* (http://www.sander.ebi.ac.uk/DSSP) como: 9.657 resíduos alpha e 6798 resíduos beta. Estas proteínas foram divididas em quatro subconjuntos para o treinamento das *RNA*'s:

a) ***Todas***, o qual possui todas as 106 proteínas coletadas;

b) ***Hélice***, que contém proteínas cujo número de resíduos em estrutura α-hélices é maior que a soma de resíduos em estrutura folhas- β e resíduos em estruturas *coil*. Fazem parte deste subconjunto as proteínas classificadas como estruturas *All Alpha* no *SCOP* e *Mainly Alpha* no *CATH*;

c) ***Folha***, que contém proteínas cujo número de resíduos em estrutura folhas- β é maior que a soma de resíduos em estrutura α-hélices e resíduos em estruturas *coil*. Fazem parte deste subconjunto as proteínas classificadas como estruturas *All Beta* no *SCOP* e no *Mainly Beta* no *CATH*; e

d) ***Hélice-Folha***, ou seja, proteínas cuja classificação seja *Alpha and Beta Proteins* no *SCOP* e *Alpha Beta* no *CATH*.

O conjunto de proteínas utilizado para testes foi obtido através do *CASP - Critical Assessment of Structure Prediction* [1]. Estas 15 proteínas são usadas, como padrão, pelo *CASP* na avaliação dos métodos de predição de estrutura secundária descritos na literatura. Este mesmo conjunto de proteínas é encontrado em [13] e [4] o que facilita a comparação entre os preditores.

Neste trabalho, foram utilizadas informações evolucionárias (informações de segmentos da estrutura primária que se mantém conservada) obtidas por alinhamentos múltiplos de seqüências que foram utilizadas como entrada ao invés de seqüências simples, aproveitando a idéia central do trabalho [11]. Para verificar se as proteínas não possuíam um alto grau de identidade na seqüência, foi realizado alinhamento múltiplo através do software Clustal-X (http://genome.jouy.inra.fr/doc/clustal/clustalx.html). Este software baseia-se no conceito de alinhamento progressivo, o qual determina os alinhamentos para cada par de seqüências e constrói uma matriz de distâncias que reflete estes alinhamentos. O alinhamento múltiplo é classificado como eficiente quando sua porcentagem de homologia for baixa (30% de homologia caracterizada bem alinhada).

Para comparação de resultados serão utilizados preditores amplamente difundidos na *Web*, resultados de trabalhos acadêmicos bem sucedidos na literatura de Biologia Computacional. Seguem os preditores que serão utilizados para comparação:

a) *PredictProtein* [12];
b) *PSIPRED* [8];
c) *JPred* [2];
d) *PREDATOR* [5];
e) *PSA* [14];
f) PREDCASA [13]

Foram projetadas e implementadas duas arquiteturas diferentes: uma contendo apenas uma rede neural com três camadas intermediárias, pelo motivo de ter alcançado o melhor resultado em [10] e outra contendo duas redes neurais cada uma com três camadas intermediárias, pelo mesmo motivo. Sendo que, nesse caso, a saída da primeira rede alimenta a entrada da segunda rede neural. 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. O objetivo da criação desta arquitetura em cadeia é melhorar os resultados dos testes, acreditando em uma técnica já utilizada em [15].

Os aminoácidos da seqüência primária de cada proteína são codificados em dados numéricos binários, para cada resíduo de aminoácido foram atribuídas 22 unidades, sendo 20 reservadas para os aminoácidos, uma para o heteroátomo (*X*) e a última reservada para indicar o fim da janela ou o aminoácido não identificado pelo *DSSP*. Uma dessas 22 unidades é sinalizada com valor 1, identificando um determinado resíduo e as outras unidades são atribuídos valor 0.

A saída da rede é a estrutura referente ao resíduo central. Para codificar os dados simbólicos (estrutura α-hélices, folhas-β e *coils*) em dados numéricos, foram definidos para cada saída da rede três unidades. Os resíduos com formação de α-hélices recebem o valor (1 0 0), folhas-β (0 0 1) e *coils* (0 1 0).

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 centro 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. Vale ressaltar que em [10] os melhores resultados foram alcançados com tamanho de janela 13. Pelos resultados obtidos pode-se perceber que tamanhos diferenciados de janelas de entrada estão diretamente relacionados com o desempenho da rede na predição da estrutura.

Com o objetivo de melhorar os resultados dos testes realizados com a arquitetura composta por uma *RNA*, foi implementada uma nova arquitetura utilizando duas redes, sendo que a saída da primeira rede é a entrada da segunda rede. É necessária a tarefa de conversão dos dados de entrada par a segunda rede. Essa tarefa é desenvolvida na inclusão de três colunas, cujos valores são o resultado da predição da estrutura secundária codificado na linha da matriz. Dessa forma, a nova matriz construída possuirá tanto o resíduo como também a sua estrutura secundária identificada pela rede (por exemplo, 0.80, 0.10, 0.23). Na inclusão dessas informações, a matriz original, de 22 unidades, passa a ter 25.

A inclusão do júri de decisão destina-se a fazer uma leitura de predição final [11]. Este júri tem a função de realizar a média aritmética sobre os resultados da predição das 18 redes. Logo após o cálculo da média, executa-se o critério de classificação. Essa classificação se baseia na análise do maior valor em cada coluna. Se a primeira coluna for o maior valor, classifica-se como α-hélice, se for a segunda, classifica-se como *coil* e se for a terceira, como folha-β, gerando, assim, o resultado da predição da estrutura secundária do resíduo em questão. Na prática, é como se o júri recebesse a predição de diferentes redes e realizasse uma média para decidir qual a estrutura está associada ao resíduo central.

3. Resultados

A rede neural *PSSPA* (arquitetura com uma e duas *RNA*'s) apresentada neste trabalho foi implementada, treinada e testada utilizando a Linguagem de Programação C, através do software Bloodshed Dev-C++ 4.0, em ambiente *MS-Windows*. A preferência pelo desenvolvimento completo da Rede Neural, preterindo softwares comerciais tais como MATLAB, objetiva o melhor entendimento da configuração da rede para posterior otimização de seus parâmetros.

Nesta arquitetura de rede neural única foi utilizada a configuração de rede que obteve melhores resultados em [10]. A rede é composta de 5 camadas (entrada, 3 intermediárias e saída). A camada de entrada possui 286 neurônios (janelamento de 13 aminoácidos x 22), as camadas intermediárias possuem 15 neurônios e a saída 3 neurônios, cada um representando uma classe de estrutura secundária. A função de ativação da camada de entrada é linear e das

demais camadas é tangente sigmoidal. Todas as simulações foram realizadas com um ciclo de treinamento igual a 1000.

A melhor porcentagem de acerto verificada, tanto com uma, quanto com duas redes, foi com a base de treinamento hélice-folha, alcançando 55,4% para a arquitetura simples e 56,2% para a arquitetura em cascata. Portanto, os resultados apresentados neste teste são das redes neurais treinadas com o subconjunto hélice-folha, cuja distribuição de estruturas secundárias e: 26% resíduos hélice e 22% resíduos folha.

Com o objetivo de verificar a porcentagem de acerto de cada janela foram realizados testes com nove redes com janelas variando de 7 a 23. Primeiramente as projetadas com uma rede e depois as projetadas com duas redes neurais, totalizando 18 redes neurais. As Tabelas 1 e 2 apresentam tais resultados.

Tabela 1: Porcentagem de acerto dos testes da Arquitetura de uma *RNA*

Proteína	7 Jan. (%)	9 Jan. (%)	11 Jan. (%)	13 Jan. (%)	15 Jan. (%)	17 Jan. (%)	19 Jan. (%)	21 Jan. (%)	23 Jan. (%)	Média dos acertos por Proteína
1QLQ	68	59	58	47	21	58	47	60	58	**52,89**
1EIG	51	58	53	49	37	54	61	56	58	**53,00**
1C56	46	57	52	51	59	57	62	49	58	**54,56**
1DAQ	45	64	44	55	58	49	38	59	60	**52,44**
1EHD	58	49	55	65	55	55	57	49	48	**54,56**
1E5B	59	52	52	47	58	57	49	53	58	**53,89**
1EJG	47	55	44	35	58	61	49	55	57	**51,22**
1ES1	55	60	55	61	51	54	59	57	33	**53,89**
1DT4	61	57	39	37	43	57	57	51	56	**50,89**
1EDS	50	49	42	60	54	56	52	64	51	**53,11**
1G6X	66	53	44	55	57	58	59	59	59	**56,67**
1DOI	64	64	55	54	56	65	67	61	60	**60,67**
1FD8	62	52	49	50	56	55	58	49	58	**54,33**
1FE5	61	62	61	69	55	61	50	60	57	**59,56**
1EHJ	49	61	50	44	52	59	53	57	53	**53,11**
Média dos acertos por Janelamento	**56,13**	**56,80**	**50,20**	**51,93**	**51,33**	**57,07**	**54,53**	**55,93**	**54,93**	

Tabela 2: Porcentagem de acerto dos testes da Arquitetura de duas *RNA*'s

Proteína	7 Jan. (%)	9 Jan. (%)	11 Jan. (%)	13 Jan. (%)	15 Jan. (%)	17 Jan. (%)	19 Jan. (%)	21 Jan. (%)	23 Jan. (%)	Média dos acertos por Proteína
1QLQ	61	60	65	43	58	67	43	65	44	**56,22**
1EIG	57	50	60	52	33	49	54	55	56	**51,78**
1C56	61	62	23	67	67	67	34	50	58	**54,33**
1DAQ	55	73	45	65	68	61	42	67	55	**59,00**
1EHD	59	57	56	65	67	65	50	56	55	**58,89**
1E5B	58	59	57	59	51	68	50	65	68	**59,44**
1EJG	47	61	49	60	59	61	44	65	59	**56,11**
1ES1	57	43	61	58	61	51	59	69	58	**57,44**
1DT4	57	61	69	59	43	62	55	51	59	**57,33**
1EDS	60	48	40	61	51	66	50	62	70	**56,44**
1G6X	65	57	54	56	57	62	53	71	68	**60,33**
1DOI	61	66	65	51	56	63	64	66	63	**61,67**
1FD8	65	55	34	53	44	59	59	55	67	**54,56**
1FE5	60	63	62	70	70	51	61	61	67	**62,78**
1EHJ	59	60	50	41	34	63	50	58	60	**52,78**
Média dos acertos por Janelamento	**58,80**	**58,33**	**52,67**	**57,33**	**54,60**	**61,00**	**51,20**	**61,07**	**60,47**	

Também houve utilização de júri de decisão, a idéia que possibilitou a melhoria de predição das redes neurais de [11]. Foram utilizados três júris: o primeiro avaliou os resultados alcançados pelas nove arquiteturas de uma rede neural (JÚRI_SIMPLES); o segundo avaliou os resultados das outras nove arquiteturas de duas redes neurais (JÚRI_DUPLO); e um terceiro que avaliava os resultados de todas as dezoito arquiteturas de rede neurais (JÚRI_FINAL). Os resultados alcançados pelos júris de decisão podem ser visualizados na Tabela 3.

Os resultados apresentados na Tabela 3 mostram que a aplicação do júri de decisão nas nove redes projetadas de uma única rede comprovou um acerto significativo entre 57 e 78%. Na implementação do júri para a arquitetura com duas redes a acurácia alcançou o intervalo entre 59 e 71%, porém com uma média de acerto para o conjunto de proteínas de 65,60%, isto é, cerca de 1,5% melhor do que o JÚRI_SIMPLES. Na implementação do júri para as 18 arquiteturas de redes a acurácia aumentou para o intervalo de 60 a 79%, tendo uma média de 69,33% para as 15 proteínas do conjunto de teste.

Tabela 3: Resultados de predição alcançados pelos três Júris de Decisão

Proteína	JÚRI_SIMPLES (%)	JÚRI_DUPLO (%)	JÚRI_FINAL (%)	Média dos acertos por Proteína
1QLQ	62	59	65	**62,00**
1EIG	57	60	60	**59,00**
1C56	63	62	64	**63,00**
1DAQ	58	70	71	**66,33**
1EHD	66	67	70	**67,67**
1E5B	60	65	67	**64,00**
1EJG	57	71	71	**66,33**
1ES1	61	67	69	**65,67**
1DT4	66	61	71	**66,00**
1EDS	62	62	69	**64,33**
1G6X	66	65	69	**66,67**
1DOI	73	70	74	**72,33**
1FD8	60	67	69	**65,33**
1FE5	70	68	72	**70,00**
1EHJ	78	70	79	**75,67**
Média dos acertos por Tipo de Júri	**63,93**	**65,60**	**69,33**	

Através destes resultados, considera-se que a rede *PSSPA* (rede neural perceptron de múltiplas camadas convencional), que será utilizada para comparações com outros preditores terá a seguinte configuração: uma arquitetura composta por nove redes neurais simples e nove redes neurais em cascata (representando os janelamentos entre 7 e 23), com uma camada de júri de decisão composto pela votação das dezoito redes neurais em questão e com treinamento realizado pela subconjunto hélice-folha de proteínas. A arquitetura da ferramenta *PSSPA* pode ser visualizada através da Figura 1.

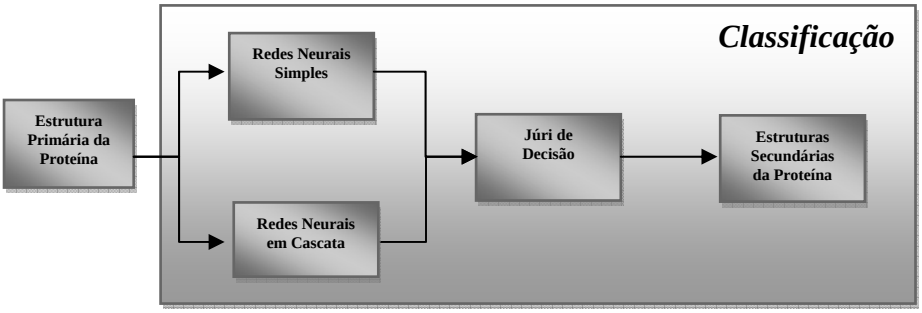

Figura 1: Arquitetura da Ferramenta *PSSPA*

Os resultados obtidos com a rede *PSSPA*, para as 15 proteínas de teste, foram comparados com preditores disponíveis na *Web*, para avaliar a qualidade do preditor desenvolvido. Todos os preditores também foram testados com o mesmo conjunto de proteínas. Os resultados da predição realizada pelas ferramentas para cada proteína seguem na Tabela 4.

Tabela 4: Comparação dos resultados da *PSSPA* com os principais preditores

Proteína	PSSPA (%)	PredictProtein (%)	PSIPRED (%)	JPred (%)	PREDATOR (%)	PSA (%)	PREDCASA (%)
1QLQ	65	91	87	90	69	51	84
1EIG	60	86	91	91	40	75	73
1C56	64	67	50	57	47	37	47
1DAQ	71	70	78	62	61	66	85
1EHD	70	55	59	82	53	58	52
1E5B	67	65	72	65	42	63	60
1EJG	71	50	67	72	56	58	80
1ES1	69	74	78	65	49	56	70
1DT4	71	71	78	49	48	63	64
1EDS	69	29	38	70	57	41	61
1G6X	69	91	91	80	53	53	84
1DOI	74	60	66	34	55	54	61
1FD8	69	79	84	90	49	27	79
1FE5	72	66	86	67	60	67	63
1EHJ	79	53	76	79	67	66	74
Média dos acertos por Ferramenta	**69,33**	**67,13**	**73,40**	**70,20**	**53,73**	**55,67**	**69,13**

A rede *PSSPA* teve um desempenho médio melhor em relação à quatro preditores e pior em relação à dois. Para verificar que a diferença não chega a ser significativa, quando a rede *PSSPA* é comparada somente com os preditores *PSIPRED* e *Jpred*, isto é, somente com os preditores que apresentaram melhor desempenho, observa-se que apesar da média inferior, a rede *PSSPA* obteve uma função mais próxima de uma constante, não apresentando grandes diferenças de porcentagem de predição na variação das proteínas.

4. Conclusões e Trabalhos Futuros

Atualmente, existe uma diversidade de programas para predição de estruturas secundárias e a idéia central é desenvolver um preditor e comparar seus resultados com os resultados obtidos por estes preditores. A partir da construção de uma ferramenta que seja equivalente o estado da arte na área de predição de estruturas secundárias, pretende-se partir para o processo de otimização da rede com a esperança de aumentar a acurácia da predição, como foi proposto em [9].

Após este trabalho inicial, pretende-se otimizar a rede *PSSPA*, para torná-la uma rede neural *MLP* otimizada (*OPSSPA*) [9], e assim, realizar novos treinamentos e testes com as mesmas proteínas. A realização desta otimização acontece na utilização de parâmetros do projeto da rede neural baseada em predições por estimativa usando ***análise multiclasse discriminante linear*** (*multi-class Linear Discriminant Analysis – LDA*) e extração de subespaços com o **Critério de Fisher de Pesos** (*weighted Fisher Criterion - wFC*) [16].

Referências

1. Casp, Community wide experiment on the critical assessment of techniques for protein structure prediction. http://predictioncenter.llnl.gov. Acessado em Dezembro de 2008.

2. Cuff, J. A.; Clamp, M. E.; Siddiqui, A. S. *et al.*, Jpred: A Consensus Secondary Structure Prediction Server, Bioinformatics, vol. 14, pp. 892-893, (1998).
3. Duda, R. O.; Hart, P. E. & Stork, D. G., "Pattern classification - Second edition" New York, John Wiley & Sons, Inc, 2000.
4. Ferreira, F. R., "O uso de rede neural artificial MLP na predição de estruturas secundárias de proteínas", Dissertação de Mestrado, Departamento de Biofísica da UNESP, São José do Rio Preto, 2004.
5. Frishman, D. & Argos, P., Seventy-five percent accuracy in protein secondary structure prediction, Proteins, vol. 27, pp. 329-335, (1997).
6. Holley, L. H. & Karplus, M., Neural Networks for Protein Structure Prediction, Physical Review, vol. 2002, pp. 204-224, (1991).
7. Jain, A. K.; Murty, M. N. & Flynn, P. J., Data clustering: a review, ACM Computing Surveys, vol. 31, pp. 264-323, (1999).
8. Jones, D. T., Protein secondary structure prediction based on position-specific scoring matrices, Journal of Molecular Biology, vol. 292, pp. 195-202, (1999).
9. Morais, E. C. & Mondaini, R. P., Optimized Multilayer Perceptrons by Weighted Fisher Criteria for Protein Secondary Structure Prediction, em: "Resumo das Comunicações do CNMAC 2008", Belém-PA, 2008.
10. Qian, N. & Sejnowski, T. J., Predicting the secondary structure of globular proteins using neural network models, Journal of Molecular Biology, vol. 202, pp. 865-884, (1988).
11. Rost, B. & Sander, C., Third generation prediction of secondary structure, Protein Structure Prediction: Methods and Protocols, vol. 143, pp. 71-95, (2000).
12. Rost, B., Yachdav, G. & Liu, J., The PredictProtein Server, Nucleic Acids Research, vol. 32 (Web Server Issue), pp. W321-W326, (2004).
13. Scott, L. P. B.; Chahine, J. & Ruggiero, J., Aplicação de redes MLP na predição de estruturas secundárias de proteínas, Biomatemática, vol.17, pp. 87-100, (2007).
14. Stultz, C. M.; Nambudripad, R.; Lathrop, R. H. & White, J. V., Predicting Protein Structure with Probabilistic Models, capítulo em: "Protein Structural Biology in Biomedical Research, vol. 22B", (E.E. Bittar, eds), pp. 447-506, JAI Press, Greenwich, 1997.
15. Chandonia, J.M. & Karplus, M., The importance of larger data sets for protein secondary structure prediction with neural networks, Protein Science, vol. 5, pp. 768-774 (1996).
16. Wang, Z.; Wang, Y; Xuan, J. *et al.*, Optimized multilayer perceptrons for molecular classification and diagnosis using genomic data, Bioinformatics, vol. 22, pp. 755-761, (2006).