http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 4 (2): 189-196, 2005
ISSN 1812-5638

Predicting Protein Secondary Structure Using Artificial Neural Networks: Current Status and Future Directions

[1]Saad Osman Abdalla and [2]Safaai Deris
[1]Faculty of Applied Sciences, Sohar University, Sohar, Sultanate of Oman
[2]Faculty of Computer Science and Information Systems,
University of Technology Malaysia, UTM Skudai, Johor, Malaysia

Abstract: Novel researchers in the area of protein secondary structure prediction using artificial neural networks take a considerable time to get the most important knowledge in this area. This study was conducted to make the most foundational and directive knowledge in protein biological aspects, protein secondary structure prediction and protein neural network predictors get elucidated. Several neural network methods have contributed and influenced significantly the field of bioinformatics in general and the area of secondary structure prediction from protein sequences in specific. Present research suggest that there is a lot of work to be done to fully exploit artificial neural networks in this area.

Key words: Protein secondary structure prediction, artificial neural networks, intelligent systems, biological sequence analysis

INTRODUCTION

Protein is considered as series of amino acids linked together into contiguous chains. The production of proteins in a cell is governed by codes and information transferred to the DNA and RNA in the organism's cell. The DNA of an organism encodes its proteins in a sequence of nucleotides, namely: adenine, cytosine, guamine and thymine. These nucleotides considered as information that governs the process of protein synthesis[1].

The amino acids consist of a carbon as a central atom linked to hydrogen or oxygen which forms molecules that connect with each other.

There are 64 different amino acids correspond from four nucleotides that make the universal genetic code (Table 1) but only twenty different types of amino acids work as basic building units of a protein (Table 2)[2,3].

The amino acid sequence is the primary structure of a protein. It is usually represented by the one letter notation of the amino acids. Amino acids combine to form a protein through polypeptide bonds and here the protein could be considered as:

Table 1: The universal genetic code

	Second position											
	T			C			A			G			
First position												Third position
T	TTT	Phe	(F)	TCT	Ser	(S)	TAT	Tyr	(Y)	TGT	Cys	(C)	T
	TTC	Phe	(F)	TCC	Ser	(S)	TAC	Tyr	(Y)	TGC	Cys	(C)	C
	TTA	Leu	(L)	TCA	Ser	(S)	TAA	Ter	(end)	TGA	Ter	(end)	A
	TTG	Leu	(L)	TCG	Ser	(S)	TAG	Ter	(end)	TGG	Trp	(W)	G
C	CTT	Leu	(L)	CCT	Pro	(P)	CAT	His	(H)	CGT	Arg	(R)	T
	CTC	Leu	(L)	CCC	Pro	(P)	CAC	His	(H)	CGC	Arg	(R)	C
	CTA	Leu	(L)	CCA	Pro	(P)	CAA	Gln	(Q)	CGA	Arg	(R)	A
	CTG	Leu	(L)	CCG	Pro	(P)	CAG	Gln	(Q)	CGG	Arg	(R)	G
A	ATT	Ile	(I)	ACT	Thr	(T)	AAT	Asn	(N)	AGT	Ser	(S)	T
	ATC	Ile	(I)	ACC	Thr	(T)	AAC	Asn	(N)	AGC	Ser	(S)	C
	ATA	Ile	(I)	ACA	Thr	(T)	AAA	Lys	(K)	AGA	Arg	(R)	A
	ATG	Met	(M)	ACG	Thr	(T)	AAG	Lys	(K)	AGG	Arg	(R)	G
G	GTT	Val	(V)	GCT	Ala	(A)	GAT	Asp	(D)	GGT	Gly	(G)	T
	GTC	Val	(V)	GCC	Ala	(A)	GAC	Asp	(D)	GGC	Gly	(G)	C
	GTA	Val	(V)	GCA	Ala	(A)	GAA	Glu	(E)	GGA	Gly	(G)	A
	GTG	Val	(V)	GCG	Ala	(A)	GAG	Glu	(E)	GGG	Gly	(G)	G

Corresponding Author: Saad Osman Abdalla Subair, Faculty of Applied Sciences, Sohar University, P.O. Box 44, Sohar, Sultanate of Oman Tel: +968 26720101 E-mail: subair@soharuni.edu.om

polypeptide chain and the amino acids as residues. Anyhow the sequence direction is very important and usually represented from the amino (N) terminus to the carboxyl (C) terminus[1-3].

PROTEIN SEQUENCES AND STRUCTURES

The sequence of amino acids in a protein chain forms the protein structure. Protein structures could be classified into four levels or classes: primary, secondary, tertiary and quaternary structure[1-3]. When the sequences of primary structures tend to arrange themselves into regular formations, these units are referred to as secondary structure.

The angles and hydrogen bond patterns between the backbone atoms are determinant factors in protein secondary structure. Moreover, secondary structure is subdivided into three parts: α-helix, β-sheet and loop[4]. α-helices and β-sheets are the most common form of secondary structure, in proteins. Loops usually serve as connection points between α-helices and β-sheets and they do not have even patterns like α-helices and β-sheets. However, in most cases any patterns which are not α-helices or β-sheets are considered as loops[1,3].

Different folds that often possess similar arrangements of a two to four consecutive recurring units of secondary structures are called super-secondary structures or motifs[5-7].

The three-dimensional structure of the protein, which is formed from the secondary structures as subunits elements, is known as the protein's tertiary structure. Forces like hydrophobic side-chains in the core of proteins, hydrogen bonds, van der Waals forces and oppositely charged amino acid side-chains are considered as the driving force of tertiary structure formation[8,9]. An individual protein that its independent fold or substructures form a three dimensional structure of the protein is known as quaternary structure. This is true for some proteins because they can not work in isolation; haemoglobin and RNA polymerase are examples of such proteins[1,3]. However, the most important obstacle in protein folding problem is that although there is considerable amount of information about the forming a three dimensional structure of a protein from simple sequences, protein folding fails *in vitro* as it does not do *in vivo*[10,11].

In the mid of nineties, bacterial genome was the first to be sequenced and at the year 2001, the first draft of human genome has been announced. However, until now there are over 65 organisms have been entirely sequenced[12-15].

Table 2: The twenty types of amino acids that forms the proteins

Protein name	Abbreviation
Alanine	A
Arginine	R
Asparagine	N
Aspartic acid	D
Cysteine	C
Glutamic acid	E
Glutamine	Q
Glycine	G
Histidine	H
Isoleucine	I
Leucine	L
Lysine	K
Methionine	M
Phenylalanine	FV
Proline	P
Serine	S
Threonine	T
Tryptophan	W
Tyrosine	Y
Valine	V

METHODS OF DETERMINING PROTEIN STRUCTURE

Three-dimensional structures of a protein can be determined in very details that describe the relative position of a single atom within the protein using two laboratory methods: x-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy. X-ray crystallography is lengthy and complicated process. It requires high level of technical ability in the laboratory to reach to an inference of the x-ray diffraction patterns. In NMR spectroscopy the protein is put in a strong magnetic field and subjected to Radio Frequency (RF) pulses. This will force the protein to emit RF radiation. Then information of protein structure can be inferred from the frequencies and intensities of the emitted radiation. NMR process is not easy and there are many biochemical constraints in this process[16]. However, these methods require several months or even years of laboratory work and they are not viable for some proteins. This led to the fact that introducing procedures or processes of protein sequence prediction could save a considerable time and efforts[16-20].

Proteins of the same family are known as homologous proteins or homologs. Proteins change conservatively through evolution and similar proteins express similar functions[21]. Comparing two different proteins homologs is known as protein sequence alignment. One of three states occurs in the alignment process: substitution which is the replacement of one or more residues, deletion which the removal of one or more residues, insertion which is the addition of one or more residues[22]. Sequence alignment is performed when different protein sequences are put in rows while columns

represent regions of mach or mismatch. When aligning two sequences, regions of mismatch in the other sequence are deleted and represented by dashes. These deleted regions are called gaps.

Alignments that contain two protein sequences are known as pairwise alignment, while those contain many sequences are known as multiple alignments. Researchers showed that similar protein sequences usually reflect similar functions, although there are exceptions of the previous conclusion[23,24].

Since Anfinsen[25] concluded that the amino acid sequence is the only source of information to survive the denaturing process and hence the structured information must be somehow specified by the primary protein sequence, researchers have been trying to predict secondary structure from protein sequence. Anfinsen's hypothesis suggests that an ideal theoretical model of predicting protein secondary structure form its sequence should exist anyhow.

The gap between known structures and known sequences is growing wider. It is known that protein structure is difficult to be predicted from the protein sequence However, considerable works has been done in this area[11,26]. The present research showed that membrane helices can be predicted much more accurately than globular helices and internal helices are predicted less accurately[27-29].

Using small datasets of protein in experimental methods to predict secondary structure adversely affected the accuracy of methods[30-32]. At present there is enough data for experimental methods to boost their accuracy[33]. Many algorithms and methods have been applied to the secondary structure of protein. Most known methods are: Statistical Methods[34,35]. Nearest-neighbour algorithms[36,37] and Neural networks methods[38,39]. Although many workers on these methods claimed accuracy as high as 78 %, using correct data set made the range of accuracy drop to the level of 60%[40,41].

The reliability index provides a good tool to study some key regions predicted at high levels of expected accuracy. Accuracy of prediction is correlated with the reliability index. This means that residues with higher reliability index are predicted with higher accuracy than others[40]. However, alignment is key point here, because bad alignments lead to bad prediction.

However, it is not always true to combine different prediction methods to reach higher accuracy. For some methods like EVA, combining different methods decreased accuracy over the best individual methods, although averaging over the better ones was better than averaging the best ones[42-44].

Protein secondary structure formation is influenced by long-range interactions and the environment[45]. Consequently, stretches of adjacent residues can be found in different secondary structure states this non-local effects are contained in the exchange patterns of protein families[46,47].

PREDICTION USING NEURAL NETWORKS

As an efficient pattern classification tool[48], neural network has been used in protein structure prediction problem by many researchers[28,41,49]. Qian and Sejnowski[55] followed by Bohr *et al.*[57] greatly influenced the approach of predicting protein structure by their work when first introduced neural networks in this area. Artificial neural networks combined with evolutionary information showed the capability and the potential of the system and paved the way of using this powerful tool in the bioinformatics field[56,59]. The high degree of prediction of protein secondary structure showed that the performance of neural networks exceeded the performance of other systems[49,50,58]. Nevertheless, neural networks has been extended to predict the location of active sites and content of protein[51,52].

Using an advanced neural network design, Baldi *et al.*[53] have improved the level of accuracy in predicting β strand pairings. Their bidirectional recurrent neural network outperformed the work of many researchers.

NEURAL NETWORKS ARCHITECTURES

Various architectures of artificial neural networks had been described by many researchers[48,54,55]. In this section we will explore some of the foundational architectures that had been used in protein secondary structure prediction. Inside the neural network (Fig. 1), many types of computational units exist; the most common type sums its inputs (x_i) and passes the result through a nonlinear approximation or activation function (Fig. 2) to yield an output (y_i). Thus the output $(y_i) = f_i (x_i)$, where f_i is the transfer function of each unit. This is summarized in Eq. 1 and 2.

$$x_i = \sum_{j \in N-(i)} w_{ij} y_j + w_i \quad (1)$$

$$y_i = f_i(x_i) = f_i\left(\sum_{j \in N-(i)} w_{ij} y_j + w_i \right) \quad (2)$$

where, w_i is the bias or threshold of the unit i.

A transfer function may be a linear function like the identity function of the regression analysis and hence the unit i is a linear unit. However, in Artificial Neural Networks most of the time the transfer functions are non

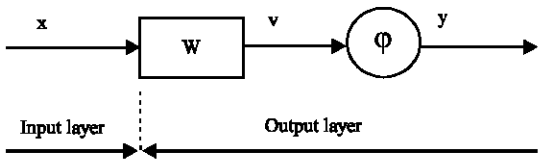

Fig. 1: Basic graphical representation of a single neuron Artificial Neural Networks

linear like sigmoid and threshold logic functions. Bounded activation functions are often known as squashing functions. When f is a threshold or bias function, then:

$$f(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Then Eq. 4 shows a sigmoid transfer function of type logistic transfer function which can estimate the probability of binary event.

$$f(x) = \sigma(x) = \frac{1}{1+e^{-x}} \quad (4)$$

One of the most important properties of Artificial Neural Networks is that they can approximate any reasonable function to any degree of precision[60,61]. If we have a continues function y = f(x) where, both y and x are one dimensional units and if x changes in the interval [0, 1], thus the value of x within a precision ϵ, where f is continuous over the compact interval [0, 1], there exists an integer n such that:

$$|x_2 - x_1| \le \frac{1}{n} \Rightarrow |f(x_2) - f(x_1)| \le \varepsilon \quad (5)$$

Then f can be approximated with a the function g(x) = f(k/n) for any x in the interval [(k-1/n, k/n)] and any unit representing k = 1,…,n.

If the data of our Artificial Neural Networks is assumed to be consisting of a set of independent input-output pairs $D_i = (d_i, t_i)$ where, d_i is the input for unit i and t_i is the output for unit i. The Artificial Neural Networks operation is a deterministic one in as seen in Eq. 6.

$$P((d_i, t_i)|w) = P(d_i|w)P(t_i|d_i, w) = P(d_i)P(t_i|d_i, w) \quad (6)$$

Hence inputs d could be assumed as independent of the parameter w, using the Bayesian inference, Eq. 6 can be transformed into[72]:

$$-\log P(w|D) = -\sum_{i=1}^{K} \log P(t_i|d_i, w) - \sum_{i=1}^{K} \log P(d_i) - \log P(w) + \log P(D) \quad (7)$$

In the case of Gaussian regression, the probabilistic model assuming that the covariance matrix is diagonal and that there are n output units indexed by j, then:

$$P(t|d, w) = \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi\sigma_j}} \exp\left(-\frac{(t_j - y_j)^2}{2\sigma_j^2}\right) \quad (8)$$

The derivative of the log likelihood *E* with respect to an output y_i is shown in Eq. 9 which really the regular Least Mean Square (LMS) error function.

$$\frac{\partial E}{\partial y_j} = \frac{\partial E}{\partial x_j} = -\frac{t_j - y_j}{\sigma_j} = -\frac{t_j - y_j}{\sigma} \quad (9)$$

For a Artificial Neural Networks that classify an input into two classes (a and a^-), the target output can represented as 0 or 1. This model is a binomial model and can be estimated by a sigmoidal transfer function as shown in Eq. 10.

$$y = y(d) = P(d \in A) = P(t|d, w) = y^t(1-y)^{(1-t)} \quad (10)$$

The relative entropy between the output distribution and the observed distribution is expressed by:

$$E = -\log P(t|d, w) = -t\log y - (1-t)\log(1-y) \quad (11)$$

If the output transfer function is the logistic function, then:

$$\frac{\partial E}{\partial y} = -\frac{t-y}{y(1-y)} \quad (12)$$

$$\frac{\partial E}{\partial x} = -(t-y) \quad (13)$$

Consequencely, in binomial classification, the output transfer function is logistic and the likelihood error function is the relative entropy between the predict distribution and the target distribution.

If the classification task of our Artificial Neural Networks has n possible classes $(a_1,\ldots, a_n)$ for a given input d, then target out put t is a vector with a single 1 and n-1 zeros. However, Eq. 14-17 summarises the multi classes classification of Artificial Neural Networks.

$$P(t|d, w) = \prod_{j=1}^{n} y_j^{t_j} \quad (14)$$

$$E = -\log P(t|d, w) = -\sum_{j=1}^{n} t_j \log y_j \quad (15)$$

$$\frac{\partial E}{\partial y_j} = -\frac{t_j}{y_j} \quad (16)$$

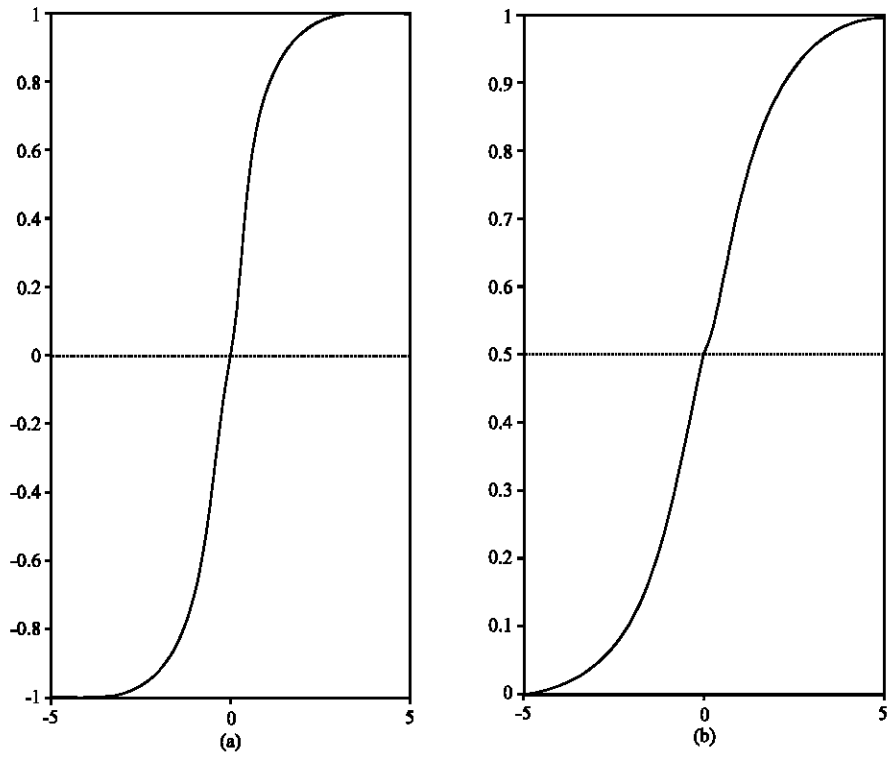

Fig. 2: The sigmoid function conventionally used in feedforward Artificial Neural Networks a. sigmoid unipolar and its derivative function. b. sigmoid bipolar and its derivative function

$$\frac{\partial E}{\partial x_j} = -(t_j - y_j) \tag{17}$$

The main differences among these Artificial Neural Networks exist in topology where connectivity of nodes, methods of training and applications of the network differ.

One of the frequently used Artificial Neural Networks is the feedforward Artificial Neural Networks trained with backpropagation for rule extraction purposes. It is termed feedforward because information is provided as input and propagated in a forward manner, with each computational unit (perceptron) integrating its inputs and firing according to its specific non-linearity.

The simplest way to reduce the network error is by changing the connections according to the derivative of the error with respect to the connections, in a process known as the gradient descent. This is often referred to as back-propagating the error through the neural network[62,63]. To avoid being trapped in local minima, in practise, the actual training is typically performed by a variant of this algorithm that permits up-hill of the curve moves[59,64].

With enough hidden units neural networks can learn to separate any set of patterns. Typical applications require to extract particular features (underlying rules) present in the patterns rather than to learn the known examples. A successful extraction of such features permits the network to generalise, i.e., to also correctly classify patterns that have not been learned explicitly. Generalisation requires a balance between the number of training examples (enough to enable feature extraction) and the number of connections (enough to separate patterns). As a rule-of-thumb the number of connections should be an order of magnitude lower than the number of patterns to avoid over-fitting the training data (this learning exactly of the training set is also referred to as over-training)[59,64].

CONCLUSIONS

Artificial Neural Networks proved that they have the ability to making complex decisions based on the unbiased selection of the most important factors from a large number of competing variables. This is particularly important in the area of protein structure determination, where the principles governing protein folding are complex and not yet well understood.

Since the neural network can be trained to map specific input signals or patterns to a desired output, information from the central amino acid of each input

value is modified by a weighting factor, grouped together then sent to a second level of a hidden layer where the signal is clustered into an appropriate class, the have great opportunities in the prediction of proteins secondary structures.

However, perdition of protein secondary structure can not be completely accurate due to the facts that the assignment of secondary structure may vary up to 12% between different crystals of the same protein. In addition, â-strand formation is more dependent on long-range interactions than β-helices and there should be a general tendency towards a lower prediction accuracy of α-strands than β-helices. However, this suggests that there still room of improvement since the accuracy of 88% has not reached yet.

ACKNOWLEDGEMENTS

We would like to thank the late Dr. Nassrudin Zenon, Faculty of Computer Science and Information Systems, Universiti Teknolgi Malaysia, for valuable recommendations in this topic. We would also like to thank Professor Joachim Diederich, Professor of Computer Science, University of Queensland for seminars and discussion in this topic and Support Vector Machines topics.

REFERENCES

1. Branden, C. and J. Tooze, 1991. Introduction to Protein Structure. Garland Publ., New York, London.
2. Brian, H., 1998. Computing science: The invention of the genetic code. American Scientist, 86: 9-14.
3. Lattman, E.E. and G.D. Rose, 1993. Protein folding-what's the question? Proc. Natl. Acad. Sci. USA., 90: 439-441.
4. Kendrew, J.C. *et al.*, 1960. Structure of myoglobin. Nature, 185: 422-427.
5. Rao, S.T. and R.M. G., 1973. Comparison of super-secondary structures in proteins. J. Mol. Biol., 286: 241-256.
6. Richardson, J.S., 1981. The Anatomy and Taxonomy of Protein Structure. Adv. Prot. Chem., 34: 168-339.
7. Sternberg, M.J.E. and J.M. Thornton, 1976. On the conformation of proteins: The handedness of the β-strand-α-helix-β-strand unit. J. Mol. Biol., 105: 367-382.
8. Dill, K.A., 1990. Dominant forces in protein folding. Biochemistry, 39: 31.
9. Hubbard, T.J. and J. Park, 1995. Fold recognition and abinitio structure predictions using hidden markov-models and β-strand pair potentials. Proteins: Struct., Funct., Genet., 23: 398-402.
9. Levitt, M. and C. Chothia, 1976. Structural patterns in globular proteins. Nature, 261: 552-557.
10. Ellis, R.J., C. Dobson and U. Hartl, 1998. Sequence does specify protein conformation. TIBS, 23: 468.
11. Wright, P.E. and H.J. Dyson, 1999. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. J. Mol. Biol., 293: 321-331.
12. Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton and E.F. Kirkness *et al.*, 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science, 269: 496-512.
13. The genome international sequencing consortium, 2001. Initial sequencing and analysis of the human genome. Nature, 409: 860-921.
14. Venter, J.C., M.D. Adams, E.W. Myers, P.W. Li and R.J. Mural *et al.*, 2001. The Human genome. Science, 291: 1304-1351.
15. O'Donovan, C., R. Apweiler and A. Bairoch, 2001. The Human Proteomics Initiative (HPI). TIBTECH, 19: 178-181.
16. Taylor, W.R. and J.M. Thornton, 1984. Recognition of super-secondary structure in proteins. J. Mol. Biol., 173: 487-514.
17. Fischer, D. and D. Eisenberg, 1996. Protein fold recognition using sequence derived predictions. Prot. Sci., 5: 947-955.
18. Defay, T.R. and F.E. Cohen, 1996. Multiple sequence information for threading algorithms. J. Mol. Biol., 262: 314-323.
20. Rice, D.W. and D. Eisenberg, 1997. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence J. Mol. Biol., 267: 1026-1038.
21. Jacob, F., 1977. Evolution and tinkering. Science, 196: 1161-6.
22. Cuff, J.A. and G.J. Barton, 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins, 40: 502-511.
23. Daniel, F., B. Christian, B. Kevin, E. Arne, G. Adam, J. David, K. Kevin, A. Lawrence, Kelley, M. Robert, P. Krzysztof, R. Burkhard, R. Leszek and S. Michael, 1999. CAFASP1: Critical assessment of fully automated structure prediction methods. Proteins: Structure, Function and Genetics, Supplement, 3: 209-217.
24. Gilbrat, J., T. Madej and S. Bryant, 1996. Surprising similarities in structure comparison. Curr. Opin. Struct. Bioi., 6: 377-85.
25. Anfinsen, C.B., 1973. Principles that govern the folding of protein chains. Science, 181: 223-230.

26. Netzer, W.J. and F.U. Hartl, 1997. Recombination of protein domains facilitated by co-translational folding in eukaryotes. Nature, 388: 343-349.
27. Kneller, D.G., F.E. Cohen and R. Langridge, 1990. Improvements in protein secondary structure prediction byan enhanced neural network. J. Mol. Biol., 214: 171-182.
28. Presnell, S.R. and F.E. Cohen, 1993. Artificial neural networks for pattern recognition in biochemical sequences. Annu. Rev. Biophys. Biomol. Struct., 22: 283-298.
29. Kernytsky, A. and B. Rost, 2003. Static benchmarking of membrane helix predictions. Nucleic Acids Res., 31: 3642-3644.
30. Rost, B. and C. Sander, 1992. Exercising Multi-layered Networks on Protein Secondary Structure. In: Neural Networks: From Biology to High Energy Physics (Benhar, O., S. Brunak, P. DelGiudice and M. Grandolfo, Eds.). Intl. J. Neural Sys., Elba, Italy, pp: 209-220.
31. Zhang, X., J.P. Mesirov and D.L. Waltz, 1992. Hybrid system for protein secondary structure prediction. J. Mol. Biol., 225: 1049-63.
32. Frishman, D. and P. Argos, 1992. Recognition of distantly related protein sequences using conserved motifs and neural networks. J. Mol. Biol., 228: 951-962.
33. Rost, B. and L. Jinfeng, 2003. The Predict Protein server. Nucleic Acids Research, 31: 3300-3304.
34. Ferran, E.A. and B. Pflugfelder, 1993. A hybrid method to cluster protein sequences based on statistics and artificial neural networks. CABIOS, 9: 671-680.
35. Bengio, Y. and Y. Pouliot, 1990. Efficient recognition of immunoglobulin domains from amino acid sequences using a neural network. CABIOS, 6: 319-324.
36. Lebeda, F.J. and M.A. M.A. Olson, 1997. Predicting differential antigen-antibody contact regions based on solvent accessibility. J. Prot. Chem., 16: 607-618.
37. Gulukota, K. and C. DeLisi, 2001. Neural network method for predicting peptides that bind major histocompatibility complex molecules. Meth. Mol. Biol., 156: 201-209.
38. Blom, N., J. Hansen, D. Blaas and S. Brunak, 1996. Cleavage site analysis in picornaviral polyproteins: Discovering cellular targets by neural networks. Prot. Sci., 5: 2203-2216.
39. Petersen, T.N., C. Lundegaard, M. Nielsen, H. Bohr and J. Bohr *et al.*, 2000. Prediction of protein secondary structure at 80% accuracy. Proteins, 41: 17-20.
40. Liu, J. and B. Rost, 2002. Target space for structural genomics revisited. Bioinformatics, 18: 922-933.
41. Arbib, M., 1995. The Hand Book Ofbrain Theory and Neural Networks. Bradford Books/The MIT Press, Cambridge, MA.
42. Rost, B., C. Sander and R. Schneider, 1994. PHD-an automatic server for protein secondary structure prediction. CABIOS, 10: 53-60.
43. Sasagawa, F. and K. Tajima, 1993. Prediction of protein secondary structures by a neural network. CABIOS, 9: 147-152.
44. Koh, I., V.A. Eyrich, M.A. Marti-Renom, D. Przybylski, M.S. Madhusudhan, E. Narayanan, O. Gran˜a, A. Valencia, A. Sali and B. Rost, 2003. EVA: evaluation of protein structure prediction servers. Nucleic Acids Res., 31: 3311-3315.
45. Liu, J. and B. Rost, 2003. NORSp: Predictions of long regions without regular secondary structure. Nucleic Acids Res., 31: 3833-3835.
46. Zhou, X., F. Alber, G. Folkers, G.H. Gonnet and G. Chelvanayagam, 2000. An analysis of the helix-to-strand transition between peptides with identical sequence. Proteins, 41: 248-256.
47. Compiani, M., P. Fariselli, P.L. Martelli and R. Casadio, 1999. Neural networks to study invariant features of protein folding. Theoretical Chemistry Accounts, 101: 21-26.
48. Haykin, S., 1999. Neural Networks: A Compernhensive Foundation. Prentice-Hall, Inc.
49. Wu, C.H., 1997. Artificial neural networks for molecular sequence analysis. Comp. Chem., 21: 237-256.
50. Rost, B. and C. Sander, 1992. Jury returns on structure prediction. Nature, 360: 540.
51. Gutteridge, A., G. Bartlett and J. Thornton, 2003. Using a neural network and spatial clustering to predict the location of active sites in enzymes. J. Mol. Biol., 330: 719-734.
52. Cai, Y.D., X. Liu, X. Xu and K. Chou, 2002. Artificial neural network method for predicting protein secondary structure content. Comp. Chem., 26: 347-350.
53. Baldi, P., S. Brunak, P. Frasconi, G. Soda and G. Pollastri, 1999. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics, 15: 937-946.
54. Baldi, P. and S. Brunak, 2001. Bioinformatics: The Machine Learning Approach, MIT Press.
55. Qian, N. and T.J. Sejnowski, 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol., 202: 865-884.

56. Pollastri, G., D. Przybylski, B. Rost and P. Baldi, 2002. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins, 47: 228-235.
57. Bohr, H., J. Bohr, S. Brunak, R.M.J. Cotterill and B. Lautrup *et al*., 1988. Protein secondary structure and homology by neural networks. FEBS Lett., 241: 223-228.
58. Holley, H.L. and M. Karplus, 1989. Protein secondary structure prediction with a neural network. Proc. Natl. Acad. Sci. USA., 86: 152-156.
59. Rost, B., 1996. NN Which Predicts Protein Secondary Structure. In: Hand Book of Neural Computation (Fiesler, E.and R. Beale, Eds.), Oxford Univ. Press, New York, pp: G4.1.
60. Hornik, K., M. Stinchocombe and H. White, 1990. Universal approximation of unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks, 3: 535-549.
61. Hornik, K., M. Stinchocombe and H. White and P. Auer, 1994. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Computation, 6: 1262-1275.
62. Muller, B. and J. Reinhardt, 1990. Neural Networks. Springer, Berlin, F.R.G.
63. Hertz, J.A., A. Krogh and R.G. Palmer, 1991. Introduction to the theory of neural computation. Addison-Wesley, Redwood City, C.A., USA.
64. Rost, B. and C. Sander, 1993. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 232: 584-599.