Internet Electronic Journal of Molecular Design **2002**, *1*, 332–338 ISSN 1538–6414
BioChem Press http://www.biochempress.com

Inter*net* Electronic Journal of Molecular Design

July 2002, Volume 1, Number 7, Pages 332–338

Editor: Ovidiu Ivanciuc

Special issue dedicated to Professor Milan Randić on the occasion of the 70th birthday
Part 3

Guest Editor: Mircea V. Diudea

Prediction of Protein Structural Classes by a Neural Network Method

Yu–Dong Cai,[1,5] Junda Hu,[2] Yi–Xue Li,[3] and Kuo–Chen Chou[4]

[1] Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai, 200233, China
[2] Department of Mathematics, University of Texas at Austin, Austin, TX, 78712, U.S.A.
[3] Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Yue Yang Rd. 319, 200031, Shanghai, China
[4] Computer-aided drug discovery, Upjohn Laboratories, Kalamazoo, Michigan 49001-4940, U.S.A.
[5] Current address: Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, U.K.

Received: April 23, 2002; Revised: July 1, 2002; Accepted: July 8, 2002; Published: July 31, 2002

Citation of the article:
Y.–D. Cai, J. Hu, Y.–X. Li, and K.–C. Chou, Prediction of Protein Structural Classes by a Neural Network Method, *Internet Electron. J. Mol. Des.* **2002**, *1*, 332–338, http://www.biochempress.com.

Inter*net* Electronic Journal of Molecular Design

BioChem Press
http://www.biochempress.com

Prediction of Protein Structural Classes by a Neural Network Method[#]

Yu–Dong Cai,[1,5,*] Junda Hu,[2] Yi–Xue Li,[3] and Kuo–Chen Chou[4]

[1] Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai, 200233, China
[2] Department of Mathematics, University of Texas at Austin, Austin, TX, 78712, U.S.A.
[3] Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Yue Yang Rd. 319, 200031, Shanghai, China
[4] Computer-aided drug discovery, Upjohn Laboratories, Kalamazoo, Michigan 49001-4940, U.S.A.
[5] Current address: Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, U.K.

Received: April 23, 2002; Revised: July 1, 2002; Accepted: July 8, 2002; Published: July 31, 2002

Internet Electron. J. Mol. Des. **2002**, *1* (7), 332–338

Abstract

Protein structures can be classified as all–α, all–β, α/β, α+β and ζ according to protein chain folding topologies. Previous studies have shown evidence that some correlation between the protein structural class and amino acid composition does exist, and the protein structural class can be predicted to some extent according to amino acid composition alone. In this study we apply Kohonen's self–organization neural network to approach this problem. The results obtained show that the structural class of a protein is considerably correlated with its amino acid composition, and the neural network is a useful tool for predicting the structural classes of proteins.

1 INTRODUCTION

Protein structures can be classified into all–α, all–β, α/β, α+β [1] and ζ classes [2] according to protein chain folding topologies. Prediction of protein structural class is very important to many aspects of molecular biology. Previous studies have shown evidence that some correlation between the protein structural class and amino acid composition does exist, and the protein structural class can be predicted to some extent from the amino acid composition alone [3–16]. This implies that protein structural class is significantly determined by the amino acids composition, although it is well known that three dimensional protein structure is determined by the amino acid interactions over the entire sequence chain. In this paper, we apply Kohonen's self–organization neural network to approach this problem. The neural network method was applied to a protein data set [17] derived

[#] Dedicated to Professor Milan Randić on the occasion of the 70th birthday.
* Correspondence author; E–mail: y.cai@umist.ac.uk.

from the SCOP database (http://scop.mrc–lmb.cam.ac.uk/scop/) [18]. As a result, high rates of self–consistency (calibration) and jackknife (leave–one–out cross–validation) tests were obtained. This shows that the structural class of a protein is considerably correlated with its amino acid composition, and the neural network is a useful tool for predicting the structural classes of proteins.

2 MATERIALS AND METHODS

Kohonen's self–organization neural network [19] is a two–layer network (Figure 1). Output nodes are arranged regularly on a planar mapping grid. Each input node is connected to each output node via a variable connection weight. Weights are adjusted interactively during training by input data and organized gradually such that topologically close nodes are sensitive to inputs that are physically similar. Kohonen's model is well known for its self–organizing and self–adaptability by learning and training with some representative examples to learn the fundamental characteristics of the objects. Therefore, it can be used to predict the structure classes of proteins. The learning algorithm of Kohonen's network model is formulated below.

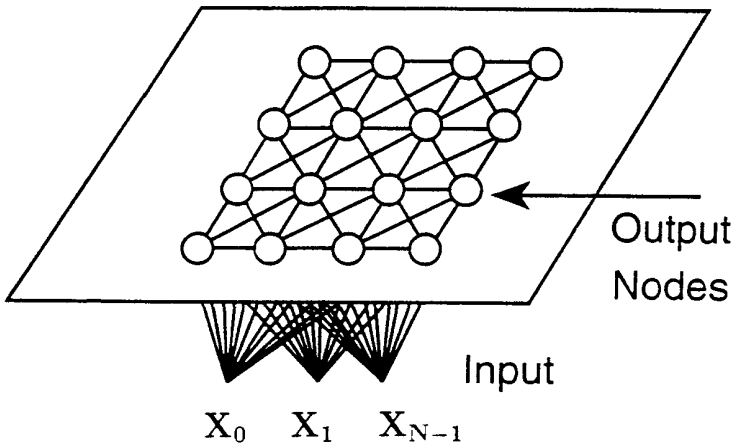

Figure 1. Self–Organization Neural Network.

Denote the feature number of samples with N, and the pre–specified block number as K.

Step 1: Initialize weights to small random values

$$0 < W_{ij} < 1, i = 0,1,\cdots,N-1; j = 0,1,\cdots,K-1$$

Step 2: Present a new sample.

$$X = \left(X_0, X_1, \cdots, X_{n-1}\right),$$

Step 3: Compute distance between X and each output node

$$d_j(t) = \sum_{i=0}^{N-1}\left(X_i - W_{ij}(t)\right)^2, j = 0,1,\cdots,K-1$$

Step 4: Select output node $j*$ which has minimum distance

$$d_{j*}(t) = \min_{0 \le j \le N-1} \left(d_j(t) \right)$$

Step 5: Update weights

$$W_{ij}(t+1) = W_{ij}(t) + \alpha(t) \times \left(X_i - W_{ij}(t) \right), j \in NE_{j*}(t)$$
$$W_{ij}(t+1) = W_{ij}(t), j \notin NE_{j*}(t)$$

where $0 < \alpha(t) = 30/(150+t)$ is a gain function that decreases in time, $NE_{j*}(t)$ is a neighborhood which contains neighboring nodes of $j*$ in a certain range. During this process the range of the neighborhood is changed such that $j*$ initially has many nodes in its neighborhood and at the end of the learning process it possesses only a few or no neighbors.

Step 6: Go to step 7 after all samples are processed. Otherwise go to step 2.

Step 7: Stop if ending criterion

$$\max_{0 \le i \le N-1, 0 \le j \le K-1} \left\{ \left| W_{ij}(t+1) - W_{ij}(t) \right| \right\} < \varepsilon$$

is satisfied or a pre–set computational time is reached. Otherwise go to step 2.

Table 1. The PDB Code of the 204 Protein Chains

(1) 52 α–proteins									
1aep_	1ash_	1bcfA	1cnt1	1gdy_	1hlb_	1ilk_	1maz_	1mls_	1rhgA
1spgB	1sra_	1vls_	2fal_	2hbg_	3sdhA	1allA	1flp_	1ibeA	1ithA
2gdm_	2lhb_	1hdsB	1myt_	1osa_	1sctA	1spgA	1fslA	1hlm_	1lht_
1outA	1outB	1pbxA	1pbxB	1sctB	1babB	2asr_	1babA	1bgc_	1bgeA
1emy_	1hdaB	1hdsA	1ibeB	1mbs_	2mm1_	2pghA	2pghB	1hdaA	1hrm_
1mygA	1vlk_								
(2) 61 β–proteins									
1bbt2	1cfb_	1edhA	1gen_	1sacA	1tcrA	2ayh_	3hhrC	6fabL	8fabB
1pex_	1vcaA	1mfbL	1gnhA	1yna_	8fabA	1flrH	1ggiH	1indH	1jelH
2cgrH	7fabH	1bbdH	1eapA	1gafL	1gbg_	1ggiL	1ghfH	1hilB	1ncbL
1nldH	1opgL	1ospL	1vgeL	2fbjL	2mcg1	7fabL	1acyL	1bafL	1bjmA
1bqlH	1bqlL	1dfbL	1forL	1ghfL	1iaiL	1iaiM	1igcL	1ikfL	1indL
1macA	1mamL	1mreH	1ngqH	1nsnH	1plgH	1plgL	1tetH	1xnd_	1yuhA
3hfmH									
(3) 45 α/β proteins									
1amp_	1ceo_	1cvl_	1dorA	1gca_	1ghr_	1gym_	1lbiA	1lucA	1masA
1nar_	1pbn_	1pfkA	1sbp_	1scuA	1thtA	1vdc_	1vpt_	1xel_	1xyzA
2bgu_	2ctc_	2ebn_	3pga1	8abp_	1enp_	1gdhA	1lucB	1obr_	1cnv_
1exp_	1trb_	1ghsA	1hdgO	1lwiA	1wsaA	2alr_	3ecaA	4pfk_	1agx_
1cerO	1gia_	2lip_	1ula_	2gbp_					
(4) 46 α+β proteins									
1aak_	1afb1	1bplA	1cof_	1cyw_	1def_	1doi_	1epaB	1fil_	1grj_
1gtqA	1hjrA	1htp_	1ino_	1itg_	1lit_	1mkaA	1msc_	1nhkL	1pkp_
1poc_	1rbu_	1seiA	1sfe_	1snc_	1std_	1tfe_	1vhh_	1vhiA	1vsd_
1whtB	1ytbA	2tbd_	8atcB	1apyB	1div_	1pvuA	1npk_	2uce_	1ril_
2prd_	1hup_	1nueA	1cdwA	1pne_	2kmb1				

In this paper, we used our own implementation (C language) of the Kohonen neural network method. The data set in [17] consists of 204 protein chains (Table 1), which fall into one of the following four structural classes: (1) all–α, (2) all–β, (3) α/β, (4) α+β. Owing to its correlation with the amino acid composition, a protein can be represented by a point or a vector in a 20–D space. Suppose there is a set of N proteins. Each of these proteins corresponds to a point in a 20–D space, as can be formulated by

$$X_k = \begin{bmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,20} \end{bmatrix} \qquad (k = 1, 2, \cdots, N)$$

where $x_{k,1}, x_{k,2}, \cdots, x_{k,20}$ are the components of 20 amino acids for the k–th protein X_k. In this research, $1/(1+X_k)$ is taken as the input of the neural network. Therefore, the number of units in the input layer of the neural network is 20. The computations were carried out on a Silicon Graphics IRIS Indigo work station (Elan 4000).

The training process of neural network is to adjust the network parameters (weights) according to the learning algorithm until the error function of the network reaches its minimum. Each sample is calculated in one cycle of the learning algorithm mentioned in this section before. The output nodes form a $X \times Y$ lattice (see Figure 2). After the learning process, the ending value ε of the training set reaches a very small value (*e.g.* 0.001) and these samples can be perfectly recognized by the neural network. The trained network (including the important information at the weights) has the function to identify the protein structure class. A testing sample (a protein) can be classified according to which output node it falls nearest to (the maximum value is corresponding to the similar point with the maximum scale product). In this research, we first test the self–consistency of the method, later the method will be cross–validated (jackknife test).

3 RESULTS AND DISCUSSION

3.1 Success Rate of Self–Consistency and Prediction of Neural Network

In this research, the examination for the self–consistency (using the same data as a test case that was used to perform the original training) of the neural network method was tested for a data set from ref. [17] that contains 204 proteins: 52 all–α, 61 all–β, 45 α/β, 46 α+β. The rates of correct prediction for the four structural classes were 185/204 = 90.1% (all–α proteins: 52/52 = 100%; all–β proteins: 61/61 = 100%; all α/β proteins: 37/45 = 82.2%; α+β proteins: 35/46 = 76.1%). This indicates that after being trained, the neural network has grasped the complicated relationship between the amino acid composition and protein structure classes (Figure 2).

1	1	1	1	1	**14**	1	4	1	1	**27**	3	4	1	1	**40**	4	1	1	1
2	1	4	1	3	**15**	1	1	4	3	**28**	1	4	1	4	**41**	4	1	3	2
3	3	3	2	2	**16**	3	2	2	2	**29**	2	2	2	2	**42**	2	2	2	2
4	2	2	2	2	**17**	2	2	2	2	**30**	2	2	2	3	**43**	2	2	2	4
5	3	1	1	1	**18**	4	1	4	4	**31**	4	4	4	4	**44**	4	1	2	2
6	1	1	1	1	**19**	1	4	4	3	**32**	4	4	4	3	**45**	4	3	4	1
7	3	2	3	3	**20**	3	4	4	1	**33**	1	1	1	1	**46**	1	1	1	1
8	4	4	4	4	**21**	1	4	4	3	**34**	4	4	3	3	**47**	4	3	2	1
9	4	3	1	1	**22**	3	1	1	3	**35**	3	3	3	3	**48**	1	3	1	2
10	1	1	1	4	**23**	4	4	3	3	**36**	3	3	3	4	**49**	3	3	3	1
11	3	3	2	2	**24**	2	2	2	2	**37**	2	2	2	2	**50**	2	2	2	2
12	2	2	2	2	**25**	2	2	4	4	**38**	2	2	2	2	**51**	2	2	2	3
13	3	3	1	4	**26**	3	4	4	3	**39**	4	3	3	2					

Figure 2. The Final SOM Grid (51×4).

3.2 Success Rate of Jackknife Test of Neural Network

Furthermore, we apply the cross–validation test (jackknife test) to the method. During the process of jackknife analysis, both the training and testing data sets are actually open, and a protein will in turn move from each to the other. As a result, the rates of correct prediction for the four structural classes of 204 proteins were 127/204 = 62.3% (all–α: 39/52 = 75%; all–β: 54/61 = 88.5%; α/β proteins: 19/45 = 42.2%; $\alpha+\beta$ proteins: 15/46 = 32.6%.

3.3 Speed of Prediction

Only additions and multiplication are needed in the prediction for the independent data set, so the speed is very high. In this research, it takes several milliseconds to compute the prediction for a sample. If we make use of parallel computers or produce a special neural network hardware, the speed will be even higher.

The above results, together with those obtained by the other prediction algorithms [4–17], indicate that the structural class of a protein is considerably correlated with its amino acid composition. It is anticipated that the Kohonen's self–organization neural network and the covariant discriminant algorithm [11, 14–17], if complemented with each other, will become a very useful tool for predicting the structural classes of proteins.

Finally, it should be pointed out that the very high success rates of self–consistency obtained here by no means represent the general rate of correct prediction in practical application, it only reflects the excellent self–consistency of the current approach for a typical and well–defined data set. The high self–consistency rates reported by many other investigators [4–16] should also be interpreted as such. Although this is common sense in statistics, unfortunately there is some confusion by some investigators who misinterpreted such a high rate as a "paradox" when compared with the best secondary structure prediction rates (about 70% so far achieved). Actually, there is no paradox at all in this regard. First, although the structural class prediction is based on the amino–acid–composition not including the effect of sequence order, it counts all the amino acids of

an entire protein (or domain) chain [17]. In contrast, although the secondary structure prediction includes the effect of sequence order, the sequence is limited within a very small portion of a protein chain without including the "long–range" interaction with the other part of the protein. Hence there is reason whatsoever to be surprised if the secondary structure prediction rate thus obtained is lower than the structural class prediction rate. Second, it has been clearly pointed out in many previous publications [14–17] that the higher than 90% self–consistency rates should not be misunderstood as the general rate of correct prediction in practical application. These rates only reflect the fact that for the same testing data sets, the rate of correct prediction can be significantly improved after taking into account coupling effect among amino–acid components [14–17]. Third, a more appropriate way to compare these two types of prediction (*i.e.*, the secondary structure prediction and the structural class prediction) should be based on their jackknife rates. However, owing to lack of a complete training data set, it is too premature to give a general deduction of such a rate for structural class prediction. Without a complete or approximately complete training data set, any attempt trying to find the upper–limit for the structural class prediction rate is invalid, and the results thus obtained is misleading.

We have to mention that besides the Kohonen self–organization neural network, the prediction of the protein structural classes can be made with other efficient classification algorithm, such as the support vector machines, which were successfully used for predicting the membrane protein types [20].

5 REFERENCES

[1] M. Levitt and C. Chothia, Structural patterns in globular proteins, *Nature* **1976**, *261*, 552–558.
[2] J. J. Chou and C. T. Zhang, A joint prediction of the folding types of 1490 human proteins from their genetic codons, *J. Theor. Biol*. **1993**, *161*, 251–262.
[3] P. Y. Chou, Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent, Las Vegas, Nevada, 1980.
[4] P. Y. Chou, Prediction of Protein Structure and the Principles of Protein Conformation, Ed. G. D. Fasman, Plenum Press, New York, 1989.
[5] H. Nakashima, K. Nishikawa, and T. Ooi, The folding type of a protein is relevant to the amino acid composition, *J. Biochem*. (*Tokyo*) **1986**, *99*, 153–162.
[6] P. Klein and C. Delisi, Prediction of protein structural class from the amino acid sequence, *Biopolymers* **1986**, *25*, 1659–1672.
[7] B. A. Metfessel, P. N. Saurugger, D. P. Connelly, and S. T. Rich, Cross–validation of protein structural class prediction using statistical clustering and neural networks, *Protein Sci*. **1993**, *2*, 1171–1182.
[8] I. Dubchak, S. R. Holbrook, and S.–H. Kim, Prediction of protein folding class from amino acid composition, *Proteins* **1993**, *16*, 79–91.
[9] K. C. Chou, and C. T. Zhang, Predicting protein folding types by distance functions that make allowances for amino acid interactions, *J. Biol. Chem*. **1994**, *269*, 22014–22020.
[10] B. Mao, K. C. Chou, and C. T. Zhang, Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins, *Protein Eng*. **1994**, *7*, 319–330.
[11] K. C. Chou, A novel approach to predicting protein structural classes in a (20–1)–D amino acid composition space, *Proteins* **1995**, *21*, 319–344.
[12] J. M. Chandonia and M. Karplus, Neural networks for secondary structure and structural class predictions, *Protein Sci*. **1995**, *4*, 275–285.
[13] I. Bahar, A. R. Atilgan, R. L. Jernigan, and B. Erman, Understanding the recognition of protein structural classes

by amino acid composition, *Proteins* **1997**, *29*, 172–185.
[14] K. C. Chou, W. M. Liu, G. M. Maggiora, and C. T. Zhang, Prediction and classification of domain structural classes, *Proteins* **1998**, *31*, 97–103.
[15] G. P. Zhou An intriguing controversy over protein structural class prediction, *J. Protein Chem*. **1998**, *17*, 729–738.
[16] K. C. Chou and G. M. Maggiora, Domain structural class prediction, *Protein Eng*. **1998**, *11*, 523–538.
[17] K. C. Chou, A key driving force in determination of protein structural classes, *Biochem. Biophys. Res. Commun*. **1999**, *264*, 216–224.
[18] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J. Mol. Biol*. **1995**, *247*, 536–540.
[19] T. Kohonen, *Self–Organization and Associative Memory*, Berlin, Springer–Verlag, 1988.
[20] Y.–D. Cai, X.–J. Liu, X. Xu, and K.–C. Chou, Support Vector Machines for Predicting Membrane Protein Types by Incorporating Quasi–Sequence–Order Effect, *Internet Electron. J. Mol. Des.* **2002**, *1*, 219–226, http://www.biochempress.com.