ELSEVIER

Polymer 44 (2003) 1751–1756

polymer

www.elsevier.com/locate/polymer

Folding rate prediction based on neural network model

Linxi Zhang[a,*], Jing Li[b], Zhouting Jiang[a], Agen Xia[a]

[a]Department of Physics, Zhejiang University, Hangzhou 310028, People's Republic of China
[b]Department of Psychology, Zhejiang University, Hangzhou 310028, People's Republic of China

Received 13 November 2002; received in revised form 3 January 2003; accepted 3 January 2003

Abstract

One of the most important challenges in biology is to understand the relationship between the folded structure of a protein and its primary amino acid sequence. A related and challenging task is to understand the relationship between sequences and folding rates of proteins. Previous studies found that one of contact order (CO), long-range order (LRO), and total contact distance (TCD) has a significant correlation with folding rate of protein. Although the predicted results from TCD can provide better results, the deviation is also large for some proteins. In this paper, we adopt back-propagation neural network to study the relationship between folding rate and protein structure. In our model, the input nodes are CO, LRO, and TCD, and the output node is folding rate. The number of nodes in the hidden layer is seven. Our results show that the relative errors for the predicted results are even lower than other methods in the literature. We also observe a best excellent correlation between the folding rate and contact parameters (including CO, LRO, and TCD), and find that the folding rate depends on CO, LRO and TCD simultaneously. This means that CO, LRO and TCD are similarly important in folding rate of protein. Some comparisons are made with other methods.

Keywords: Contact order; Long-range order; Total contact distance; Folding rate; Back-propagation neural network; Protein folding

1. Introduction

Protein folding is the process by which a protein progresses from its denatured state to its specific biologically active conformation. It has been proposed that this process has to follow a specific pathway or set of pathways in order to fold in a finite time. Predicting the native structures of proteins from their amino acid sequences has remained an elusive goal for many years. A related important task is to understand the relationship between sequences and folding rates of proteins. The folding rate of proteins that fold with two- or weakly three-state kinetics has a significant correlation with the average sequence separation of all contacting residues in the native state, defined by the parameter contact order (CO) [1]. CO is defined as

$$\mathrm{CO} = \frac{1}{n_c n_r} \sum_{k=1,|j-i|>l_{\mathrm{cut}}}^{n_c} |j-i| \tag{1}$$

where n_r is number of amino acid residues of a protein (excluding disordered regions), and n_c is number of nonlocal residue–residue contacts, i and j are two residues. A nonlocal contact is defined as two heavy atoms within a cutoff distance R_{cut} and separated by at least a residue separation cutoff value l_{cut}. This parameter reflects the relative importance of local and nonlocal contacts in protein structures. The dependence of folding rates on the CO reflects the contribution of chain entropy loss to the folding free energy barrier. Bonneau et al. developed a method for ab inito protein structure prediction, which is based on a picture of protein folding in which local sequence segments flicker between different possible local structures, and produced a dearth of high CO structures and a excess of low CO structures [2–4]. Several kinetic theories to predict folding rates from native structures are developed, and the accuracy can be improved further [5,6]. Later a different parameter is found to correlate better with $\ln k_f$ than CO. The parameter is called long-range order (LRO) for a protein from the knowledge of long-range contacts (contacts between two residues that are close in space and far in

[*] Corresponding author. Tel.: +86-49-571-88273270; fax: +86-49-571-88273276.

E-mail address: lxzhang@hzcnc.com (L. Zhang).

0032-3861/03/$ - see front matter © 2003 Elsevier Science Ltd. All rights reserved.
doi:10.1016/S0032-3861(03)00021-1

the sequence) in protein structure. It is defined as [7]

$$\mathrm{LRO} = \frac{\sum n_{ij}}{n_\mathrm{r}} \qquad n_{ij} = \begin{cases} 1 & |i-j| > 12 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where i and j are two residues for which the $C^\alpha - C^\alpha$-distance is ≤ 8.0 Å and n_r is number of amino acid residues of a protein. The new predicting result suggests that the long-order contact play an important role on folding kinetics. The difference between LRO and CO exists because LRO only considers long-range contacts, and CO discusses all the contacts of proteins.

Their results also show that that either CO or LRO parameter has a significant correlation with the folding rate [5–7]. It is shown that the significance of the balance between long- and short-interactions in determining protein structure [8]. Recently, Zhou and Zhou brought forward a new parameter, total contact distance (TCD), to predict folding rate [9]. The new parameter includes the sequence separation per contact and total number of contacts simultaneously, and is shown to be best in correlation with the logarithms of folding rates. However, the deviation between the experimental $\ln k_\mathrm{f}$ and the predicted $\ln k_\mathrm{f}$ is large for some proteins [9]. The reason may be that the folding rate does not depend only on CO, or LRO, or TCD, and has a correlation with CO, LRO and TCD simultaneously. In fact, CO, or LRO, or TCD depends on the property of contact (or the protein structure). At that point, it is the same for CO, or LRO, TCD. Of course these relationships may be more complex and difficult to express with a certain mathematical formula. In the meantime, those investigations only provide some external relationships between the folding rates and the properties of proteins (such as CO, LRO, or TCD). Analysis of the thermodynamics and kinetics of the folding process provides an understanding as to how interatomic interactions determine the native conformation of a globular. Zwanzig was the first to apply the master equation to describe the kinetics of protein folding [10–12]. Then, Hao and Scheraga solved this equation for several models of the transition rate constants for simple model proteins [13,14]. Recently, Ye and Scheraga reported the general solution of the master equation to describe the folding kinetics of an arbitrary protein model [15]. It is solved by using the Laplace transformation to calculate the rate of electron transfer through protein and DNA molecules. Ye et al. used a master equation to describe the mechanism of folding of Staphylococcal Protein A [16], and Ghosh et al. used a stochastic difference equation for an atomically detailed study of the folding pathways of protein A [17]. Those theoretical works on folding mechanism can help understand mechanisms of protein folding. Because of the complexity of the protein-folding problem, various methods should be adopted, here including molecular dynamics simulation [18–20], and theoretical investigation [13–17]. Of course, it is necessary to find some internal factors and external factors which effect on folding rate, and help us understand the mechanisms of protein folding in more detail. Here a new method of artificial neural networks [21–26] is adopted to study this relationship. Artificial neural networks utilize weight matrices to perform mathematical transformations of the input vector to the output, and have one obvious advantage: there is no need to know the exact form of the analytical function on which the model should be built. Therefore artificial neural networks are suitable for experimental condition optimization since the relationship between the evaluating indices and the experimental input parameters is complex, nonlinear, and often cannot be expressed with a certain mathematical formula.

Neural networks—a pattern recognition technique is widely used in protein science, mostly for the prediction of protein secondary structure from sequence. The accuracy level of prediction of protein secondary structure has been recently substantially increased by using a neural network and information from multiple sequence alignment, thereby surpassing the 70% level of average three-state accuracy [27,28]. An area particularly suited to neural network methods is the identification of protein sorting signals and the prediction of their cleavage sites, as these functional units are encoded by local, linear sequences of amino acids rather than global 3D structures [29–32]. We also use neural networks method to predict the statistical properties of polymers [24–26]. In this paper, we will discuss the folding rate based on artificial neural networks. In our model, the input node is CO, LRO, and TCO, and the output node is folding rate.

2. Artificial neural network model

The back-propagation neural network is one of the most popular neural-network topologies. It has the advantages of being easy to understand and easy to implement [33]. The architecture of the three-layer network is shown in Fig. 1. The input nodes are fully connected to a hidden layer of nodes, which in turn are fully connected to a set of output

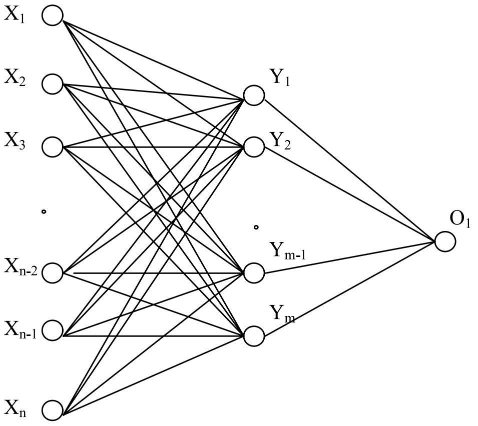


Fig. 1. Architecture of the neural network model in this paper.

nodes. The number of input nodes and output nodes are n and p, respectively, and the number of hidden nodes is m. Here the output node is the folding rate, therefore, the number of output nodes p is 1. This neural network algorithm can evolve a set of weights to produce an arbitrary mapping from input to output by presenting pairs of input nodes and their corresponding output nodes.

For an input layer node, the output (Xo) is equal to its input (X):

$$Xo_i^L = X_i^L \qquad (i = 1, 2, \dots, n;\ L = 1, 2, \dots S) \tag{3}$$

The subscript i indicates the ith input layer node, and L indicates the Lth trained sample. n is the number of input nodes, and S is the number of train samples The outputs from the inputs are weighted and fed to the hidden nodes. The net input of jth hidden node in the Lth trained sample is:

$$\text{net}_j^L = \sum_{i=1}^{n} Xo_i^L W_{ij}^L \qquad (j = 1, 2, \dots, m;\ L = 1, 2, \dots, S) \tag{4}$$

where W_{ij}^L indicates the weight of the connection between the ith input node and the jth hidden node in the Lth trained sample, and m is the number of hidden node. The outputs of the hidden layer nodes (Yo_k^L) are given after transformation by the sigmoidal function:

$$Yo_k^L = f(\text{net}_k^L), \quad (k = 1, 2, \dots, m) \tag{5}$$

Here $f(x)$ is the sigmoidal function:

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{6}$$

The output of the last layer (output layer) nodes are given by the sum after weighted input from the hidden nodes and transformation:

$$O_l^L = f\left(\sum_{j=1}^{m} Yo_j^L V_{jl}^L\right) \qquad (l = 1, 2, \dots, p;\ L = 1, 2, \dots, S) \tag{7}$$

Here V_{jl}^L is the weight connecting the jth hidden node to the lth output layer node. After carrying out the procedure about for all combination of input signals, the root-mean-square error (RMSE) is given by:

$$\text{RMSE} = \sqrt{\sum_{\alpha=1}^{p} \sum_{\beta=1}^{S} \frac{E_{\alpha\beta}^2}{pS}}, \qquad E_{\alpha\beta} = \frac{Y_\alpha^\beta - O_\alpha^\beta}{Y_\alpha^\beta} \tag{8}$$

Here p is the number of output layer nodes ($p = 1$ in this paper), and S is the number of trained samples, and Y_α^β is the target value of βth trained samples.

Every weight, the initial value of which is randomly given by the computer, is changed automatically by Eqs. (9) and (10), and then training is repeated until RMSE falls into an acceptable region

$$\Delta W_{ij}^L(t) = \eta \delta_j^L Xo_i^L + \mu \Delta W_{ij}^L(t-1) \tag{9}$$
$$(i = 1, 2, \dots, n;\ j = 1, 2, \dots, m;\ L = 1, 2, \dots, S)$$

$$\Delta V_{kl}^L(t) = \eta \xi_l^L Yo_k^L + \mu \Delta V_{kl}^L(t-1) \tag{10}$$
$$(k = 1, 2, \dots, m;\ l = 1, 2, \dots, p;\ L = 1, 2, \dots, S)$$

here t is the number of iterations of neural computation, and η is the learning rate and determines how fast the changes $\Delta W_{ij}^L(t)$ and $\Delta V_{kl}^L(t)$ should be implemented in the iteration cycles. The momentum constant μ prevents sudden changes in the direction in which corrections are made. The errors ξ_l^L and δ_j^L can be expressed as

$$\xi_l^L = (Y_l^L - O_l^L) f'(O_l^L) \qquad (l = 1, 2, \dots, p) \tag{11}$$

for the output layer lth node and

$$\delta_j^L = \left(\sum_{\alpha=1}^{p} \xi_\alpha^L V_{\alpha j}^L\right) f'(y_j^L) \qquad (j = 1, 2, \dots, m) \tag{12}$$

for hidden layer jth node. Here $f'(x)$ is the derivative with respect to x of the transformation function, and $f'(x) = (1 - f(x)) \times f(x)$.

3. Results and discussion

Although the values of CO, LRO, and TCD are not the same for a protein, the values depend mainly on the protein structure. Previous investigations only emphasize the correlation of folding rate with one of CO, LRO, and TCD. In fact the folding rate may have a complex relation with CO, LRO, and TCD. Therefore in our calculation the input nodes are chosen as CO, LRO, and TCD, and the number of the input nodes (n) therefore is three. The values of CO, LRO are calculated by Eqs. (1) and (2) and the values of TCD are taken from Ref. [9], and the values of CO, LRO, and TCD of all proteins are given in Table 1. It also contains experimental data of folding rate of 28 proteins. The initial architecture of the neural network is only an initial guess, it must be modified after performing calculations. The learning rate η and the momentum constant μ should be selected carefully. In the training process, if the value of the learning rate and the momentum constant were too small, RMSE would show large fluctuation, and could not reach a global minimum quickly. In our training, $\eta = 0.1$ and $\mu = 0.89$.

The selection of the number of hidden layer nodes is important in the artificial neural network. In order to obtain the best neural network structure, several artificial neural network systems with different number of hidden nodes are tested. Here the testing proteins of output nodes are 2ABD, 1NYF, and 1CSP, and the number of protein samples is 25, and RMSE = 0.038. Here the proteins are randomly chosen. The results are shown in Fig. 2, and the relative error (Δ) is defined as

$$\Delta = \frac{|\ln k_\mathrm{f}(\text{experiment}) - \ln k_\mathrm{f}(\text{predicted})|}{\ln k_\mathrm{f}(\text{experiment})} \tag{13}$$

Table 1
The values of three parameters (CO, LRO, and TCD) and the experimental value of ln k_f used in this study

Protein	CO (%)[a]	LRO[b]	TCD[c]	ln k_f[d]
1LMB	18.4	0.61	0.75	8.50
2ABD	24.4	1.15	1.04	6.55
1IMQ	22.0	0.85	0.93	7.31
2PDD	23.9	0.49	0.75	9.80
1NYF	32.9	1.40	1.22	4.54
1PKS	32.3	1.92	1.38	− 1.05
1SHG	34.8	1.51	1.35	1.41
1SRL	34.5	1.55	1.25	4.04
1FNF_9	33.9	1.99	1.30	− 0.91
1FNF_10	30.1	1.87	1.14	5.48
1HNG	33.3	1.56	1.25	2.89
1TEN	32.8	1.92	1.24	1.06
1TIT	33.2	2.07	1.26	3.47
1WIT	35.1	2.48	1.48	0.41
1CSP	30.9	1.52	1.10	6.98
1MJC	30.8	1.49	1.14	5.24
2AIT	34.3	2.07	1.42	4.20
1APS	34.9	2.09	1.52	− 1.48
1HDN	31.1	1.73	1.35	2.70
1URN	30.4	1.46	1.20	5.73
2HQI	31.1	2.15	1.48	0.18
1PBA	29.8	1.32	1.08	6.80
1UBO	29.1	1.18	1.07	7.33
2PTL	31.1	1.37	1.23	4.10
1FKB	31.6	1.98	1.30	1.46
1COA	31.0	1.42	1.14	3.87
1DIV	24.4	0.84	0.88	6.58
2VIK	21.7	1.67	0.97	6.80

[a] $R_{\mathrm{cut}} = 0.60$ nm (based on the heavy atom distance) and $l_{\mathrm{cut}} = 2$.
[b] $R_{\mathrm{cut}} = 0.80$ nm (based on the $C^{\alpha}-C^{\alpha}$ distance) and $l_{\mathrm{cut}} = 2$.
[c] $R_{\mathrm{cut}} = 0.60$ nm and $l_{\mathrm{cut}} = 2$.
[d] Experimental value of ln k_f are taken from 1IMQ [35], 2PDD [36], 1FNF_9 [37], 1FNF_10, 1HNG, 1TIT, 1WIT [38], 2HQI [39], 1PBA [40], 1DIV [41] and all others are from Ref. [42].

here the average error is averaged over 2ABD, 1NYF, and 1CSP, and m is the number of hidden layer nodes. We find that $m = 7$ may be the best fit for our system. Similar relationships between average error and the number of hidden layer are obtained in our calculations. Therefore, we choose the 3–7–1 structure as the optimal system. The weight matrices W_{ij} and V_{kl} depends on many factors, such as the input node (CO, LRO, TCD), the output node (ln k_f), and the number of samples. They also depend on the RMSE. If the same values of the input nodes and the output node are given, the weight matrices W_{ij} and V_{kl} are different when the RMSE changes in the neural network training. In the meantime, the weight matrices W_{ij} and V_{kl} may be different in each run. As a fortran program to make neural network predictions of folding rates from CO, LRO and TCD can be freely available, the weight matrices W_{ij} and V_{kl} is not shown here.

We randomly choose three proteins in all 28 proteins. We train on 25 and test on three, and the results are given in Table 2, here RMSE = 0.038. In Table 2 six runs are given. Each run represents a different set of proteins. Except for 2AIT, the error is acceptable. In Table 2 the Zhou's results are also given, here Zhou's predicted values are obtained from ln $k_f = -13.2\,\mathrm{TCD} + 19.73$ [9]. In general, our results provide more accurate prediction than the previous results. In order to compare, we adopt the jackknife cross-validation method, and plot the predicted vales of ln k_f vs the observed one, and the results are given in Fig. 3. In Fig. 3(b) TCO make a good prediction, whereas neural network provides the best results, see Fig. 3(a).

In folding rate predicted by back-propagation neural network, one of the obvious outliers is 2AIT Our predicted ln k_f is 2.04, whereas the experimental one is equal to 4.20. The experimental one is greater than the predicted one by 100%. In Zhou's prediction based on TCD, the predicted value of ln k_f is only 0.70, and is about 0.03 times slower than the actual folding rate. Although our predicted result is

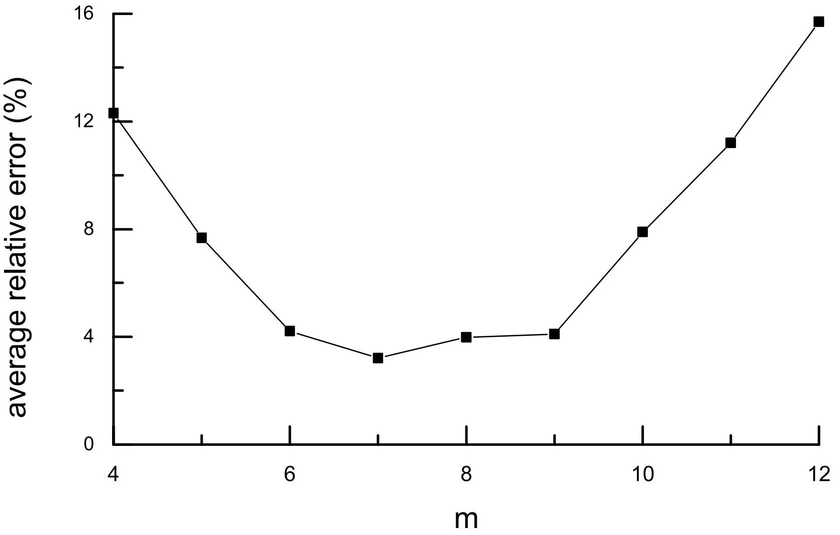


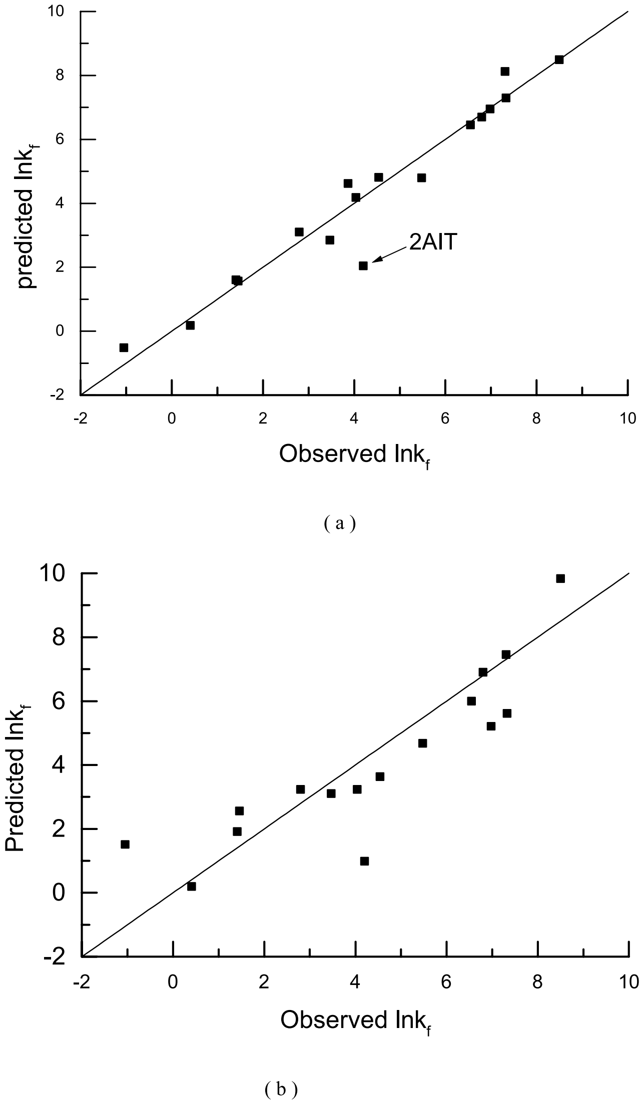
Fig. 2. A plot of average relative error versus m (the number of hidden nodes) for 2ABD, 1NY, and 1CSP with RMSE = 0.038.

Table 2
Results of the predicted $\ln k_f$ for selected proteins

Protein	Predicted	Experimental $\ln k_f$	Zhou's results[a]
Run A			
1SHG	1.60	1.41	1.91
1FNF_10	4.79	5.48	4.68
2VIK	6.69	6.80	6.90
Run B			
1LMB	8.49	8.50	9.83
1TIT	2.85	3.47	3.10
1FKB	1.56	1.46	2.56
Run C			
2ABD	6.45	6.55	6.00
1NYF	4.81	4.54	3.63
1CSP	6.95	6.98	5.21
Run D			
1SRL	4.18	4.04	3.23
2AIT	2.04	4.20	0.986
1UBO	7.29	7.33	5.61
Run E			
1IMQ	8.12	7.31	7.45
1PKS	-0.52	-1.05	1.51
1HNG	3.10	2.80	3.23
Run F			
1IMQ	8.14	7.31	7.45
1WIT	0.18	0.41	0.19
1COA	4.62	3.87	4.68

[a] Obtained from $\ln k_f = -13.2\text{TCD} + 19.73$. [9].

more close to the experimental one than the previous predicted one, the deviation is also large for 2AIT. The reason is that 2AIT has disulfide bonds. Experimental studies [34] have shown that removing one disulfide bond via mutation would reduce the folding rate of 2AIT. Thus, the observed folding rate after a single disulfide bond mutation ($\ln k_f = 2.1$) is a closer to the predicted one (2.04) based on neural network method.



(a)

(b)

Fig. 3. Scatter plot of the experimental and predicted folding rates by jackknife test. (a) Neural network model; (b) TCD. Here RMSE = 0.038.

4. Conclusion

We have made an attempt to predict the folding rate using neural network model. In fact, sequence separation per contact (LRO), total number of contact (CO), and TCD are similarly important in determining the rate of folding. Although TCD is the most accurate among the three parameters (CD, LRO, TCD), a large deviations between the experimental $\ln k_f$ and the predicted $\ln k_f$ for some proteins exist. This means that TCD maybe have a less correlation with folding rate for some proteins. Our neural network results provide the fact that the folding rate has a correlation with CO, LRO, and TCD simultaneously. In fact, CO, LRO, and TCD depend on the contact. In this point, they are the same. Although the values of CO, LRO, and TCD are different for a protein, they all have a tight relation with protein structure. Our results have even more accurate prediction than the previous results.

One of the inadequate factors which effects our predictions is that there are no enough experimental value of $\ln k_f$ of proteins (28 proteins). Another may be that our prediction only suits fast-folding proteins. If there are fast-folding proteins and slow-folding proteins simultaneously in protein samples, the training should be taken dividedly. In the meantime, neural network method will probably fail in prediction of the folding rate of proteins with the chaperon-type kinetics. The reason may be that the process of folding is different for these proteins.

In fact our results agree well with the previous method [1, 5–9]. If we can consider CO, long-range contact, and TCD simultaneously, the accuracy can be improved. However, if we only discuss one of CO, LRO, and TCD, some good results can also obtained for some proteins. This means that CO, LRO and TCD has an interior relationship between them. This discussion can help us know the folding rate from protein structure clearly.

A fortran program (nn-folding.f) that provides neural network predictions of folding rates from CO, LRO and TCD will be freely available upon request (lxzhang@hzcnc.com).

Acknowledgements

This research was financially supported by National Natural Science Foundation of China (Nos. 29874012, 20174036, 20274040), Natural Science Foundation of Zhejiang Province (No. 10102) and the Special Funds for Major State Basic Research Projects (G1999064800). We thank Dr Yaoqi Zhou for providing us a fortran program calculating TCD, CO, and LRO from a pdb structure. We also thank the referees for their critical reading of the manuscript and their good ideas.

References

[1] Plaxco KW, Simons KT, Baker D. J Mol Biol 1998;277:985–94.

[2] Simons KT, Bonneau R, Ruczinshi I, Baker D. Proteins 1999;37: 171–6.

[3] Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Proteins 1999;34:82–95.

[4] Bonneau R, Ruczinski I, Tsai J, Baker D. Protein Sci 2002;11: 1937–44.

[5] Alm E, Baker D. Proc Natl Acad Sci USA 1999;96:11305–10.

[6] Dinner AR, Kaplus M. Nature Struct Biol 2001;8:21–2.

[7] Gromiha MM, Selvaraj S. J Mol Biol 2001;310:27–32.

[8] Wako H, Scheraga HA. Macromolecules 1981;14:961–9.

[9] Zhou H, Zhou Y. Biophys J 2002;82:458–63.

[10] Zwanzig R, Szabo A, Bagchi B. Proc Natl Acad Sci USA 1992;89: 20–2.

[11] Zwanzig R. Proc Natl Acad Sci USA 1995;92:9801–4.

[12] Zwanzig R. Proc Natl Acad Sci USA 1997;94:148–50.

[13] Hao MH, Scheraga HA. J Chem Phys 1997;107:8089–102.

[14] Hao MH, Scheraga HA. J Chem Phys 1997;107:8089–102.

[15] Ye YJ, Scheraga HA. Kintics of protein folding, in slow dynamics in complex systems. In: Tokuyama M, Oppenheim I, editors. Eighth Tohwa University International Symposium. AIP Conference Proceedings, vol. 469. American Institute of Physics; 1999. p. 452–75.

[16] Ye YJ, Ripoll DR, Scheraga HA. Comput Theor Polym Sci 1999;9: 359–70.

[17] Ghosh A, Elber R, Scheraga HA. Proc Natl Acad Sci USA 2002;99: 10394–8.

[18] Boczko EM, Brooks III CL. Science 1995;269:393–6.

[19] Guo Z, Brooks III CL, Boczko EM. Proc Natl Acad Sci USA 1997;94: 10161–6.

[20] Shea JE, Onuchic JN, Brooks III CL. Proc Natl Acad Sci USA 1999; 96:12512–7.

[21] Bishop C. Neural networks for pattern recognition. London: Oxford University Press; 1995.

[22] Zupan J, Gasteiger J. Neural network for chemist. Weinheim: VCH; 1993.

[23] Specht DF. IEEE Trans Neural Network 1991;2:568–76.

[24] Zhang L, Xia A, Zhao D. J Polym Sci Polym Phys Ed 2000;38: 3163–7.

[25] Zhang L, Zhao D, Huang Y. Chin J Polym Sci 2002;20:25–30.

[26] Huang Y, Xia A. J Zhejiang Univ Sci Ed 2001;28:617–20. in Chinese.

[27] Rost B, Sander C. Proteins: Struct Funct Genet 1994;19:55–72.

[28] Salamov AA, Solovyev VV. J Mol Biol 1995;247:11–15.

[29] Claros MG, Brunak S, Heijne GV. Curr Opin Struct Biol 1997;7: 394–8.

[30] Presnell SR, Cohen FE. Annu Rev Biophys Biomol Struct 1993;22: 283–98.

[31] Hirst JD. Biochemistry 1992;31:7211–8.

[32] Chadonia JM, Karplus M. Protein Sci 1996;5:768–74.

[33] Rumelhart D, Hinton G, Williams R. Nature 1986;323:533–6.

[34] Schonbrunner N, Koller KP, Scharf M, Engels J, Kiefhaber T. Biochemistry 1997;30:9051–6.

[35] Ferguson N, Capaldi AP, James R, Kleanthous C, Radford SE. J Mol Biol 1999;286:1597–608.

[36] Spector S, Kuhlman B, Fairman R, Wong E, Boice J, Raleigh DP. J Mol Biol 1998;276:479–89.

[37] Plaxco KW, Spitzfaden C, Campbell ID, Dobson CM. J Mol Biol 1997;270:763–70.

[38] Clarke J, Cota E, Fowler SB, Hamill SJ. Struct Fold Des 1999;7: 1145–53.

[39] Aronsson G, Broesson AC, Sahlman L, Jonsson BH. FEBS Lett 1997; 41:359–64.

[40] Villegas V, Azuaga A, Catasus L, Reverter D, Mateo PL, Aviles FX, Serrano L. Biochemistry 1995;34:15105–10.

[41] Kuhlman B, Luisi DL, Evans PA, Raleigh DP. J Mol Biol 1998;284: 1661–70.

[42] Jackson SE. Fold Des 1998;3:R81–R91.