Journal of Electronics and Information Science (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/jeis.2024.090218
ISSN 2371-9524 Vol. 9 Num. 2

Protein folding rate prediction integrating multi-level structural information

Mingxiao Xu, Zhouting Jiang*, Zhenan Wu

College of Science, China Jiliang University, Hangzhou, Zhejiang, China

Keywords: BP neural network; activation function; protein folding rate prediction; contact ordert

Abstract: Studying protein folding can not only drive the great development of life sciences, but also provide tremendous help for human disease prevention and treatment, and has great application value in the fields of medicine and bioengineering. This article uses the BP neural network model to predict the rate of protein folding, and provides an effective way to find the key factors of protein folding kinetics. The main research contents are as follows: (1) Optimization of the neural network model. This article selected 4 types of optimizers and 36 types of activation function combinations to assess the performance of the neural network model in predicting the rate of protein folding. From the results, it is more accurate and fast to predict the rate of protein folding when using the Adam optimizer and Sigmoid and Tanh function combinations as the parameters of the neural network model. (2) The influence of chain length in primary structure information on prediction accuracy is studied and compared. From the prediction results, it was found that the prediction accuracy was higher when using the effective chain length of the protein than using the protein chain length. (3) Study the effect of different separation cutoffs on protein folding rates. The results show that when the separation cutoff value lcut of the contact order is 3, the most accurate prediction value can be obtained.

1. Introduction

The mechanism of protein folding, that is, how to develop from the denatured state of the protein to a specific biologically active conformation, is crucial for structural and functional biology. Understanding the basic factors that regulate the folding process can provide answers to some issues in functional genomics and biotechnology. It has been proposed that the folding process must follow one or a set of specific paths in order to complete the folding in a limited time. If this pathway is proven to be narrow, then it is only necessary to sample a small part of the conformational space to avoid Levinthal's paradox[1].Predicting the folding rate of proteins from amino acid sequences is one of the greatest challenges in computational and molecular biology[2].Interactions between residues are very important for controlling the folding process and maintaining the stability of the protein structure. In 1998, Plaxco and Baker investigated the relationship between the topological complexity of protein tertiary structures and protein folding rates. They found that the folding rate of proteins (ln(kf)) is inversely related to the contact order

(CO) [3] and proposed a prediction method based on the parameter CO. The formula is as follows:

$$CO = \frac{1}{n_r n_r} \sum_{|i-j|>l_{cut}}^{n_c} |i-j| \quad (1)$$

The formula includes n_r, which is the number of amino acid residues in a protein (excluding disordered regions), and n_c, which is the number of non-local residue-residue contact pairs. Here, i and j represent the sequence numbers of the two residues. Non-local residue-residue contact pairs are defined as two heavy atoms that are within a cutoff distance R_{cut} and separated by at least the residue separation cutoff value l_{cut}. The l_{cut} describes the number of intervals between the two residues that form a contact pair along the protein chain. It analyzes the composition of the surrounding residues from the level of sequence position; contributions from $<\pm3$ residues are considered short-range contact distances, ±3 or ±4 residues as medium-range contact distances, and $>\pm4$ residues are considered long-range contact distances[10].In CO, the l_{cut} is between 2 and 6 [4-6]. However, this parameter only considers two-state proteins and does not involve non-two-state proteins, and the prediction accuracy is highly dependent on the dataset (with a correlation coefficient as high as 0.81 for 12 proteins, but only 0.64 for 18 proteins)[8]. Although CO, as an empirical parameter based on 3D structure, has a good correlation with folding rates, its physical meaning is not clear. Gong and others believe that CO may just be a proxy for some other underlying physical variables, which are the real factors determining folding rates. They suggest that CO is a composite variable whose true meaning represents the content of the protein's secondary structure (SSC) [7].Based on this idea, they proposed a method for predicting protein folding rates based on SSC, with the following formula:

$$\ln k_f = aT + bH + cB + dL^{-1} - e \quad (2)$$

The formula includes T for turn content, H for helix content, B for beta-hairpin content, and L for the number of residues (sequence length), with a, b, c, d, e being the regression coefficients. Using this technique to predict the folding rates of 24 two-state proteins can achieve an autocorrelation coefficient of up to 0.91. Since this method only requires the content information of the secondary structure, it is also possible to estimate the protein folding rates using the secondary structure that is calculated from the predicted amino acid sequence. This provides methodological support for predicting protein folding rates based solely on the primary sequence[11-13]. Experimental observations have found that during protein folding, alpha-helices can rapidly fold first, forming certain folded regions (preformed blocks). Taking into account the influence of these pre-formed folded regions on subsequent folding, Ivankov and Finkelstein proposed the concept of effective length in 2004, with the following calculation formula:

$$L_{eff} = L - L_H + l_1 * N_H \quad (3)$$

In the formula, L represents the sequence length, LH is the number of residues in helical conformation, and NH is the number of helices. At this point, it is necessary to consider the entire helical region as l1 main chain residues. From the physical point of view, the size of l1 should be determined based on experimental values and should not exceed 4 residues. When using linear regression to predict the folding rates of 57 proteins, the best prediction accuracy can exceed 80%. Furthermore, this prediction method can be applied to both two-state and three-state proteins.

In recent years, neural network models have been widely applied in protein science, particularly in predicting protein structures, identifying protein-protein interaction sites, and inferring surface protein properties [14-17]. Among them, the back propagation (BP) neural network, trained using the error backpropagation algorithm, is one of the most widely used models. It can learn and store the "best match" relationships between a large number of input and output nodes with different features, without needing prior knowledge to express these matching relationships in mathematical equations.

The topology of the BP neural network model includes input, hidden, and output layers. The core of each neural network, the activation function, plays a crucial role in the success of training the neural network [18-19].

In this paper, by comparing different combinations of activation functions, we select the most suitable activation functions for use in predicting protein folding rates, aiming to unleash the best performance of BPNN. Then, we combine secondary structure information with primary structure information and tertiary structure information respectively, and discuss the most suitable parameter combinations to be used as inputs for the neural network.

2. Methods

2.1 Datasets

The protein folding data in this study were obtained from relevant literature [15, 20-21], comprising specific folding data for 140 proteins. Among them, 88 proteins were classified as bistable proteins, and 52 proteins were classified as non-bistable proteins. All protein structures can be accessed from the Protein Data Bank (http://www.rcsb.org/).

2.2 BP neural network

Although any three-layer BP neural network can adapt to various functions given appropriate training time and sample size, as the number of network layers increases, the model's representational capacity improves but the complexity of the network also increases. Previous studies have indicated that the number of hidden layers is not necessarily the more, the better. The determination of the number of hidden layers varies in different research systems and is validated through computer experiments [9]. In this study, neural network models with 3 to 8 layers were computed, and it was found that the 5-layer neural network model had the highest prediction accuracy and required less fitting time compared to the 3-layer neural network model. Therefore, in this study, the neural network structure was set to five layers, consisting of one input layer, three hidden layers, and one output layer. The number of nodes in the input layer was 3, and in the output layer was 1, representing the predicted value of protein folding rate. The number of nodes in the hidden layers was determined to be 12, 8, and 6 respectively through multiple tests. The structure of the four-layer network is illustrated in Figure 1.

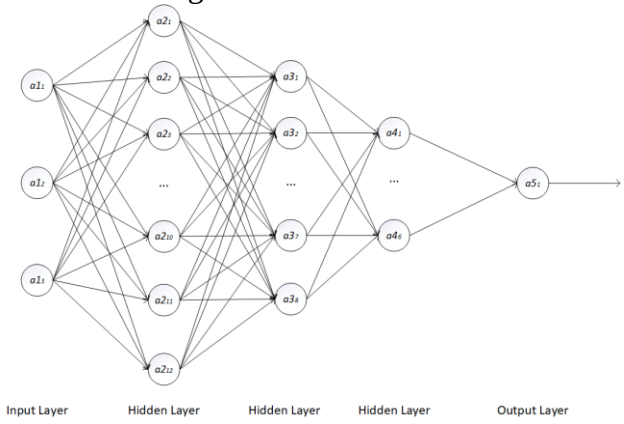

Figure 1: BP Neural Network Structure Diagram

2.3 Evaluation Criteria

In this paper, the evaluation criteria for predicting results are the coefficient of determination (R-squared) and the root mean square error (RMSE).

3. Results

Files separate until after the text has been formatted and styled. Abbreviations and Acronyms

3.1 Choice of activation function

In this paper, activation functions are considered one of the most important factors affecting the efficiency and accuracy of BP neural networks. To assess the performance of the neural network, six functions, namely Sigmoid, Tanh, ReLU, Mish, ELU, and HardSwish, were selected as activation functions to train the same set of protein data. All activation functions are only used between hidden layers, resulting in a comparison of 36 combinations of activation functions. Additionally, the research problem in this paper involves regression prediction, so there is no activation function in the output layer. The three parameters in the input layer are helix content, sheet content, and turn content.

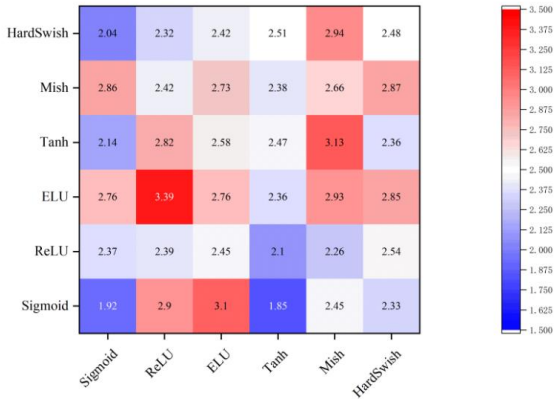

Figure 2: Heatmap of Predicted Values with Different Activation Functions

In this study, the prediction results of 36 combinations of activation functions on the same test set were examined. From Figure 2, we can see that the combination of ELU as the first activation function and ReLU as the second activation function performed the worst in this set of experiments. Conversely, the best combination of activation functions was found to be Sigmoid and Tanh. Several factors contribute to these results.

Firstly, it relates to the nature of the data. The data here represents the percentage of protein secondary structure content, which has some relevance to the characteristics of the Sigmoid and Tanh activation functions (ranging from -1 to 1 for Tanh, and 0 to 1 for Sigmoid). When the output range of the activation function matches the expected range of the network input or output layers, it may benefit network learning.

Secondly, it pertains to the stability of gradients. Sigmoid and Tanh activation functions provide relatively smooth gradients. Although ReLU and ELU were designed to address the gradient vanishing problem in Sigmoid and Tanh, if their parameters are not properly initialized or if there are a large number of negative values in the training data, they may lead to the "dying ReLU" problem, where neuron outputs are consistently 0, resulting in ineffective gradient backpropagation. In contrast, Sigmoid and Tanh can retain more gradient information when the input is between -1 and 1.

Lastly, Sigmoid and Tanh functions limit the output within a fixed range, which aligns with the bounded nature of the input data. Moreover, due to its symmetric nature, Tanh can maintain the directionality of gradients near the point of 0, which may make the learning process more effective for specific tasks compared to using ReLU and ELU.

3.2 Integrating primary and secondary structure information for prediction

This part of the research focuses on predicting the protein folding rate by integrating primary and secondary structure information. The protein chain length or effective chain length, helix content, and sheet content are used as input parameters for the neural network. The final prediction results, as shown in the figure 3 and figure 4, clearly indicate that combining effective chain length with protein secondary structure information leads to better predictions. This outcome suggests that the effective chain length highlights key parts relevant to protein function, which need to fold correctly to perform their functions. Folding typically involves the association of specific regions with the overall three-dimensional structure, and the effective chain length focuses on these regions. Additionally, the effective chain length often contains protein domains or secondary structure elements, which are the basic building blocks in the folding process. During folding, these structural units fold relatively independently and then assemble into the complete tertiary structure. Therefore, the folding rate of the structural units represented by the effective chain length has a certain influence on the overall protein folding rate.

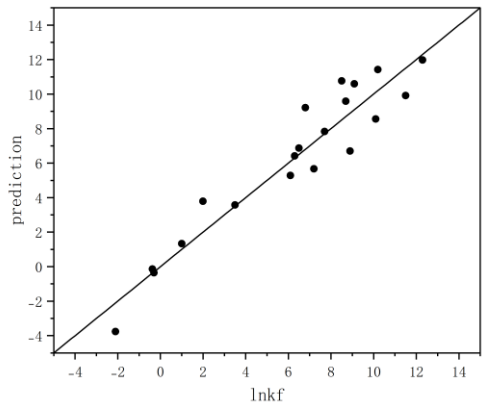

Figure 3: Secondary structure information combined with chain length prediction

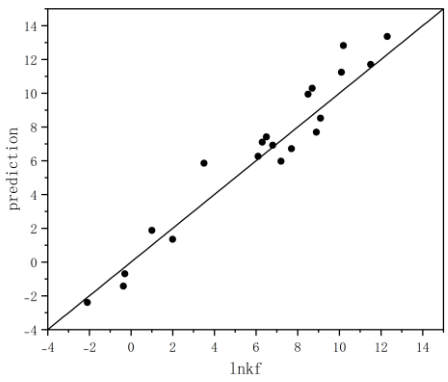

Figure 4: Secondary structure information combined with effective chain length prediction

3.3 Integrating secondary and tertiary structure information.

The main focus of this research section is to analyze the predictive impact of combining CO with secondary structure information under different separation cutoff values. CO, along with helix content and sheet content, is used as input parameters for the neural network. By experimentally studying the effects of different lcut values on the prediction of protein folding rates (the specific prediction results are shown in Figures 5 to 9), the final prediction results indicate that the best prediction is achieved when the lcut value is set to 3, with CO combined with secondary structure.

In protein molecular chains, consecutive four atoms are prone to form torsional dihedral angles, known as rotatable torsion angles, which represent the primary degrees of freedom for proteins. The system potential is influenced by torsion potential interactions. The bond stretching energy arises from interactions between every two adjacent amino acids in the protein chain. Bond bending energy, on the other hand, originates from minor oscillations of the bond angles around their equilibrium positions between every three amino acids due to chemical bonds. These three types of interactions, along with non-bonded interactions, collectively constitute the system potential of proteins. When the cutoff value for residue separation in CO stabilizes at 3, it yields the best prediction results. Therefore, it can be inferred that bond bending energy and torsion energy have a significant impact on the folding rate of proteins within the system potential.

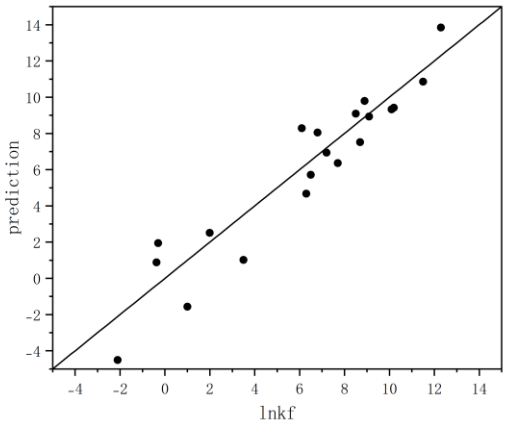

Figure 5: lcut=2 prediction

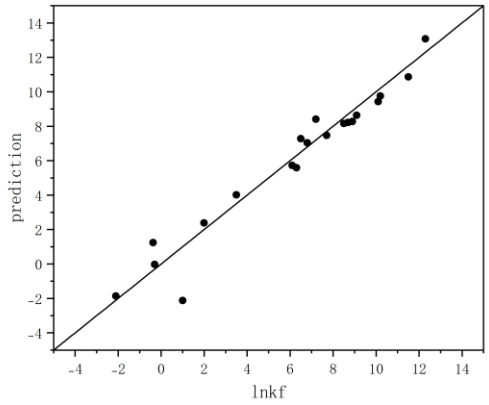

Figure 6: lcut=3 prediction

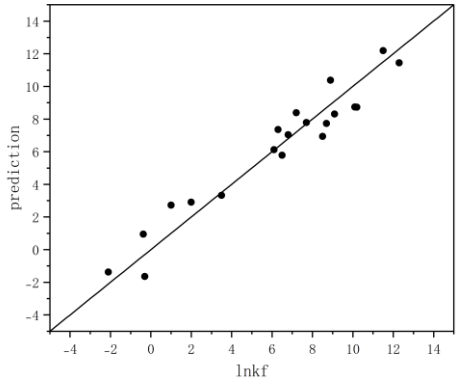

Figure 7: lcut=4 prediction

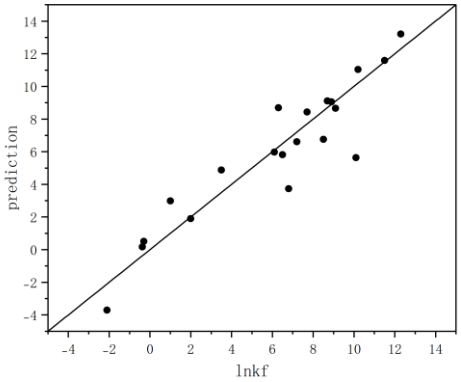

Figure 8: lcut=5 prediction

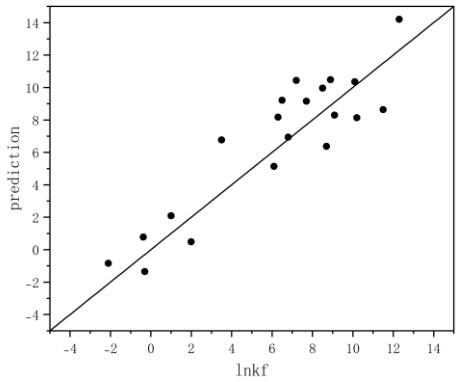

Figure 9: lcut=6 prediction

4. Conclusions

This study initially started from the protein secondary structure information, using BP neural networks to test the impact of different activation function combinations on prediction results. The results indicated that the best activation function combination was Sigmoid and Tanh, with a

resulting RMSE value of 1.85. Subsequently, the prediction was made by combining protein secondary structure information with primary structure information and comparing the prediction results obtained using protein chain length and effective chain length. It was found that the prediction effect was better when combining effective chain length with secondary structure information, with an RMSE value of 1.22. Finally, the prediction effect of combining CO with secondary structure under different lcut values was explored. The experimental results showed that when lcut was set to 3, coupling secondary structure information yielded the best prediction results, with an RMSE value of 0.95. These results are helpful for understanding protein folding.

References

[1] Levinthal C. Are there pathways for protein folding? [J]. Journal de chimie physique, 1968, 65: 44-45.

[2] GALZITSKAYA O V, IVAKOV D N, Finkelstein A V. Folding nuclei in proteins[J]. Molecular Biology, 2001, 35: 605-613.

[3] PLAXCO K W, SIMONS K T, BAKER D. Contact order, transition state placement and the refolding rates of single domain proteins [J]. Journal of Molecular Biology, 1998, 277(4): 985-994

[4] IVANKOV D N, GARBUZYNSKIY S O, ALM E, et al. Contact order revisited: influence of protein size on the folding rate [J]. Protein Science, 2003, 12(9): 2057-2062.

[5] GROMIHA M M, SELVARAJ S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction[J]. Journal of Molecular Biology, 2001, 310(1): 27-32.

[6] ZHOU H, ZHOU Y. Folding rate prediction using total contact distance[J]. Biophysical Journal, 2002, 82(1): 458-463.

[7] MAKAROV D E, KELLER C A, PLAXCO K W, et al. How the folding rate constant of simple, single-domain proteins depends on the number of native contacts[J]. Proceedings of the National Academy of Sciences, 2002, 99(6): 3535-3539.

[8] ZHANG L, LI J, JIANG Z, et al. Folding rate prediction based on neural network model[J]. Polymer, 2003, 44(5): 1751-1756.

[9] HUANG L T, GROMIHA M M. Analysis and prediction of protein folding rates using quadratic response surface models[J]. Journal of Computational Chemistry, 2008, 29(10): 1675-1683.

[10] GROMIHA M M, SELVARAJ S. Inter-residue interactions in protein folding and stability[J]. Progress in Biophysics and Molecular Biology, 2004, 86(2): 235-277.

[11] WAKO H, SCHERAGA H A. Use of distance constraints to fold a protein[J]. Macromolecules, 1981, 14(4): 961-969.

[12] GUO Z, BROOKS III C L, BOCZKO E M. Exploring the folding free energy surface of a three-helix bundle protein[J]. Proceedings of the National Academy of Sciences, 1997, 94(19): 10161-10166.

[13] GROMIHA M M, HUANG L T. Machine learning algorithms for predicting protein folding rates and stability of mutant proteins: comparison with statistical methods[J]. Current Protein and Peptide Science, 2011, 12(6): 490-502.

[14] GARBUZYNSKIY S O, IVANKOV D N, BOGATYREVA N S, et al. Golden triangle for folding rates of globular proteins[J]. Proceedings of the National Academy of Sciences, 2013, 110(1): 147-150.

[15] LANSBURY JR P T. Evolution of amyloid: what normal protein folding may tell us about fibrillogenesis and disease[J]. Proceedings of the National Academy of Sciences, 1999, 96(7): 3342-3344.

[16] WARDAH W, KHAN M G M, SHARMA A, et al. Protein secondary structure prediction using neural networks and deep learning: A review[J]. Computational Biology and Chemistry, 2019, 81: 1-8.

[17] APICELLA A, DONNARUMMA F, ISGRÒ F, et al. A survey on modern trainable activation functions[J]. Neural Networks, 2021, 138: 14-32.

[18] GHOSH A, ELBER R, SCHERAGE H A. An atomically detailed study of the folding pathways of protein A with the stochastic difference equation[J]. Proceedings of the National Academy of Sciences, 2002, 99(16): 10394-10398.

[19] PAN Y, WANG Y, ZHOU P, et al. Activation functions selection for BP neural network model of ground surface roughness[J]. Journal of Intelligent Manufacturing, 2020, 31: 1825-1836.

[20] ZHANG L, SUN T. Folding rate prediction using n-order contact distance for proteins with two-and three-state folding kinetics[J]. Biophysical Chemistry, 2005, 113(1): 9-16.

[21] MUÑOZ V, EATON W A. A simple model for calculating the kinetics of protein folding from three-dimensional structures[J]. Proceedings of the National Academy of Sciences, 1999, 96(20): 11311-11316.