J. Biomedical Science and Engineering, 2009, 2, 136-143 JBiSE

Prediction of protein folding rates from primary sequence by fusing multiple sequential features

Hong-Bin Shen[1,3,*], Jiang-Ning Song[2], Kuo-Chen Chou[1,3]

[1]Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai, 200240, China; [2]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan; [3]Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, USA.
*Corresponding author: hbshen@sjtu.edu.cn

Received 20 May 2009; revised 23 May 2009; accepted 1 June 2009.

ABSTRACT

We have developed a web-server for predicting the folding rate of a protein based on its amino acid sequence information alone. The web-server is called Pred-PFR (Predicting Protein Folding Rate). Pred-PFR is featured by fusing multiple individual predictors, each of which is established based on one special feature derived from the protein sequence. The ensemble predictor thus formed is superior to the individual ones, as demonstrated by achieving higher correlation coefficient and lower root mean square deviation between the predicted and observed results when examined by the jackknife cross-validation on a benchmark dataset constructed recently. As a user-friendly web-server, Pred-PFR is freely accessible to the public at www.csbio.sjtu.edu.cn/bioinf/FoldingRate/.

Keywords: Protein Folding Rate; Ensemble Predictor; Fusion Approach; Web-Server; Pred-PFR

1. INTRODUCTION

Knowledge of protein three-dimensional (3D) structures plays an indispensable role in molecular biology, cell biology, biomedicine, and drug design [1]. However, each protein begins as a polypeptide, translated from a sequence of mRNA as a linear chain of amino acids. A protein can function properly only if it is folded into a correct shape or conformation [2]. Failure to fold into the intended 3D structure usually produces inactive proteins with different properties. Although many efforts have been made trying to understand the mechanism of protein folding (see, e.g., [3,4,5,6]), it still remains one of the most challenging problems in molecular biology. In addition to understanding how a protein chain is folded, it is also important to find the folding rates of proteins from their primary sequences. Protein chains can fold into the functional 3D structures with quite different rates, varying from several microseconds to even an hour [7,8].

Experimentally determining the three dimensional structure of a protein is often very difficult and expensive. However the sequence of that protein is easily known. Therefore, for quite a long time, scientists have tried to use the "least free energy principle" [2,9] to predict the 3D structures of proteins. Unfortunately, owing to the notorious local energy minimum problem, so far it can only be successfully used to address very limited structural characters, such as the handedness tendency and packing arrangement in proteins (see, e.g., [10,11,12]). In the past two decades, various statistical methods have been developed for predicting the structural classes of proteins and their folding patterns according to the sequence information alone (see, e.g., [13,14,15,16,17,18,19,20,21,22,23,24,25,26,27,28] and a review [29]). Encouraged by the results obtained via these statistical approaches, various methods were developed for predicting the folding rates of proteins because the information thus acquired would be very useful for understanding the protein folding mechanism and the sequence-structure-function relationship [8,30]. In this regard, the approaches can be generally categorized into two groups: (1) the prediction of protein folding rates is based on the protein structure information; and (2) the prediction is based on the primary sequence information.

For the first group, the features of proteins are extracted from their 3D structural information and hence the predictions are feasible only after the structures have been determined. Most of the methods in this group tried to derive the statistical significance of the correlation between the protein folding rate and the corresponding structural topological parameters, such as contact order (CO) [31], absolute contact order (Abs_CO) [32], total contact distance (TCD) [33], long-range order (LRO) [34], the fraction of local contact (FLC) [34], the chain

topology parameter (CTP) [35] and the most recent geometric contact number (N_α) [30].

For the second group, the features of proteins are mainly extracted from their primary amino acid sequences, such as the amino acid biochemical properties [36] and the effective folding length (L_{eff}) [8] derived from the sequence-predicted secondary structure. The approaches in the second group are particularly useful when the 3D structural information of the protein concerned is not available.

Although the aforementioned methods in predicting folding rates of proteins each have their own merits, they were all established by focusing on one (or a few) specific feature(s). As is well known, a protein folding system is very complicated that involves many physical and chemical factors. For this kind of complicated biological system, it would be particularly effective to treat it by assembling many individual predictors with each operated based on its own special feature [37,38]. In view of this, the present study was devoted to develop a novel ensemble predictor for predicting the folding rate of a protein chain by incorporating its many different features through an optimal fusion process.

2. MATERIALS AND METHODS

To develop a powerful statistical predictor, the first important thing is to obtain an effective benchmark dataset [39]. To realize this and also for facilitating comparison with the existing prediction methods, we use the benchmark dataset as described below.

2.1. Benchmark Dataset

The large dataset recently constructed by Ouyang and Liang [30] was used in the current study. It contains 80 proteins whose folding rates have been experimentally determined. Of the 80 proteins, 45 belong to the two-state folding behaviors without the visible intermediates while the other 35 belong to the three-state or multi-state folding kinetics that exhibit the obvious intermediate state during the folding process under the experimental conditions. If classified according to their structural classes,18 are all-α proteins, 32 all-β, and the remaining 30 are $\alpha\beta$ proteins (where $\alpha\beta$ means the mix of α/β and $\alpha+\beta$ [40]). The folding rates of the 80 proteins range from $\ln K_f = -6.9$ to $\ln K_f = 12.9$, spanning more than eight orders of magnitude of K_f. For users' convenience, the benchmark dataset, denoted as $\mathbb{S}_{bench}$, is given in the Online Supporting Information A, which can also be downloaded from the web-site at www.csbio.sjtu.edu.cn/bioinf/FoldingRate/. It is instructive to point out that K_f in $\mathbb{S}_{bench}$ is actually an apparent folding rate constant (see Appendix A). Therefore, to develop a statistical method for predicting K_f of a protein according to its sequence information alone, there is no need to discriminate whether the protein is two-state or multi-state folding.

2.2. Sequence Feature Extraction

As mentioned above, although the features extracted from the 3D structures of proteins are very useful for predicting their folding rates, they can be used only when the corresponding PDB codes are available. Owing to such a limit, in this study we will focus on those features that can be derived from the amino acid sequential information alone, either directly or indirectly.

(a) Amino acid properties. Protein is composed of different amino acids, which show different physical, chemical, and conformational properties and hence may have correlations with the folding rates. In this study, the following four amino acid properties were used: α_c, the propensity to be at the C-terminal of α-helix [41]; β_S, the propensity to form β-strand [41]; τ, the compressibility [42]; and SASA, the solvent accessible surface area in an unfolding protein chain [43]. Suppose a protein **P** is expressed by

$$\mathbf{P} = R_1R_2R_3R_4R_5R_6R_7\cdots R_L \tag{1}$$

where R_1 represents the 1st residue of the protein **P**, R_2 the 2nd residue, and so forth. Thus, the protein's scores in the aforementioned four amino acid properties can be formulated as

$$\Phi_i = \frac{\sum_{j=1}^{L}\Phi_{i,j}}{L} \qquad (i = 1,2,3,4) \tag{2}$$

where L represents the protein length, and

$$\Phi_{i,j} = \frac{\Phi^0_{i,j}}{\mathbf{Max}_j\{\Phi^0_{i,j}\} - \mathbf{Min}_j\{\Phi^0_{i,j}\}} \qquad (i = 1,2,3,4;\ \ j = 1,2,\ \cdots,\ 20) \tag{3}$$

where $\Phi^0_{i,j}$ ($i = 1,2,3,4$) respectively represent the original α_c, β_S, τ, and SASA for the j-th ($j = 1,2,\cdots,20$) native amino acid, and their values can be obtained from [41,42,43]; $\mathbf{Max}_j\{\Phi^0_{i,j}\}$ means taking the maximum one among $\Phi^0_{i,1}$, $\Phi^0_{i,2}$, $\cdots$, $\Phi^0_{i,20}$, and $\mathbf{Min}_j\{\Phi^0_{i,j}\}$ the corresponding minimum one. For reader's convenience, the values thus obtained for $\Phi_{i,j}$ ($i = 1,2,3,4$; $j = 1,2,\ \cdots,\ 20$) (cf. **Eq.3**) are given in **Table 1**.

(b) Protein size effect. Many studies have indicated that the protein chain length L and its fractional powers ($L^{1/2}$, $L^{2/3}$, or $L^{3/5}$) or logarithm $\ln(L)$ have a good correlation with the folding rates, suggesting that

L and its various expressions forms could be useful features for predicting protein folding rates [8,30]. In the present study, $\ln(L)$ was adopted.

(c) Information derived from secondary structure prediction. Given a protein sequence, its secondary structure can be predicted by means of various secondary structure prediction tools. In the present study, based on the information thus obtained by using PSIPRED [44], we have the secondary structure content ratios for the protein **P**, as formulated by

$$\Gamma_\alpha + \Gamma_\beta + \Gamma_C = 1 \tag{4}$$

where Γ_α, Γ_β and Γ_C are the ratios of the α-helix, β-sheet, and coiled-coil residues for the protein **P**. Note that although the secondary structure content contains three components ($\Gamma_\alpha, \Gamma_\beta, \Gamma_C$), they were treated as one feature because of the normalized condition imposed by **Eq.4**. Moreover, based on the secondary structure prediction results, the effective protein folding chain length can be derived, as given by [8]:

$$L_{\text{eff}} = L - L_{\text{H}} + L_{\text{h}} \cdot N_{\text{H}} \tag{5}$$

where L is the total number of amino acids for the entire protein chain; L_H the number of predicted helical conformation residues; N_H the number of predicted helices; and L_h the number of an α-helix turn (L_h is generally ≤ 4; for a standard α-helix, $L_h = 3.6$). In the current study, L_h was set at 3, and $\ln(L_{\text{eff}})$ used as the feature input.

2.3. Prediction Algorithm

According to the above section, we have a set of seven different kinds of specific features, as can be summarized by the following equation:

$$\mathbb{S}_{\text{feature}} = \begin{cases} \Phi_1 = \alpha_c \\ \Phi_2 = \beta_S \\ \Phi_3 = \tau \\ \Phi_4 = \text{SASA} \\ \Phi_5 = \ln(L) \\ \Phi_6 = (\Gamma_\alpha, \Gamma_\beta, \Gamma_C) \\ \Phi_7 = \ln(L_{\text{eff}}) \end{cases} \tag{6}$$

To study the folding rate of a protein chain, the key is to determine K_f, the so-called folding rate constant. For reader's convenience, a brief discussion about the role of K_f (or its logarithm $\ln K_f$) on the protein folding rate is provided in **Appendix A**. According to **Eq.6**, we can construct the following seven linear regression models for predicting the protein folding rate constants:

$$\ln K_{\text{f}}^{(1)} = a_1 + b_1 \cdot \alpha_c \tag{7.1}$$

$$\ln K_{\text{f}}^{(2)} = a_2 + b_2 \cdot \beta_S \tag{7.2}$$

$$\ln K_{\text{f}}^{(3)} = a_3 + b_3 \cdot \tau \tag{7.3}$$

$$\ln K_{\text{f}}^{(4)} = a_4 + b_4 \cdot \text{SASA} \tag{7.4}$$

$$\ln K_{\text{f}}^{(5)} = a_5 + b_5 \ln(L) \tag{7.5}$$

$$\ln K_{\text{f}}^{(6)} = a_6 + b_{6,1} \cdot \Gamma_\alpha + b_{6,2} \cdot \Gamma_\beta + b_{6,3} \cdot \Gamma_C \tag{7.6}$$

$$\ln K_{\text{f}}^{(7)} = a_7 + b_7 \ln(L_{\text{eff}}) \tag{7.7}$$

where $K_{\text{f}}^{(i)}$ $(i = 1, 2, \cdots, 7)$ is the protein folding rate constant predicted based on the i-th specific feature Φ_i (cf. **Eq.6**), while a_i and b_i are the corresponding parameters determined by using the regression analysis on a training dataset such as $\mathbb{S}_{\text{bench}}$. For the details of how to use the regression procedures to determine a_i and b_i, refer to [45]. Note that $K_{\text{f}}^{(6)}$ of **Eq.7.6** is involved with more parameters because the 6-th feature Φ_6 contains three sub-features (cf. **Eq.6**).

All the above seven formulae (**Eqs. 7.1–7.7**) can be used to predict the protein folding rates but they each reflect the effect (s) of only one (or one kind) of specific feature (s). To incorporate the effects from all the seven kinds of features, let us consider the following formulation:

$$\ln K_{\text{f}} = \sum_{i=1}^{7} w_i \ln K_{\text{f}}^{(i)} \tag{8}$$

where w_i is the weight that reflects the impact of the i-th specific feature Φ_i on the protein folding rate. If the impacts of the seven features were the same, we should have $w_i = 1/7$ $(i = 1, 2, \cdots, 7)$. Since they are actually not the same, it would be rational to introduce some statistical criterion to reflect their different impacts, as formulated below.

Given a statistical system consisting of N samples, the Pearson Correlation Coefficient (ACC) is defined by

$$\text{PCC} = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left[\sum_{i=1}^{N} (x_i - \bar{x})^2\right]\left[\sum_{i=1}^{N} (y_i - \bar{y})^2\right]}} \tag{9}$$

where x_i and y_i are, respectively, the observed and predicted results for the i-th sample, while $\bar{x}$ and $\bar{y}$ the corresponding mean values for the N samples. Since PCC reflects the correlation of the predicted results with the actual ones, its value can be used to

measure the quality of a prediction method. If all the predicted results are exactly the same as the observed ones, we have the perfect correlation of PCC=1. For different prediction algorithms, **Eq.9** will yield different values of PCC. Therefore, the weight w_i in **Eq.8** can be formulated as

$$w_i = \frac{\mathrm{PCC}(K_{\mathrm{f}}^{(i)})}{\sum_{j=1}^{7} \mathrm{PCC}(K_{\mathrm{f}}^{(j)})} \quad (i = 1, 2, \cdots, 7) \qquad (10)$$

where $\mathrm{PCC}(K_{\mathrm{f}}^{(i)})$ is the Pearson Correlation Coefficient (**Eq.9**) obtained with the i-th folding rate predicting formula in **Eq.7** on the benchmark dataset $\mathbb{S}_{\mathrm{bench}}$ by the jackknife cross-validation.

The prediction method by fusing the seven individual methods as formulated by **Eq.7** is called the **Pred-PFR** (Predictor of Protein Folding Rate).

3. RESULTS AND DICSUSSIONS

In statistical prediction, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling test, and jackknife test [40]. However, as elucidated in [38] and demonstrated by **Eq.5** of [39], among the three cross- validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark dataset, and hence has been increasingly and widely used by investigators to examine the accuracy of various predictors (see, e.g., [46,47,48,49,50,51,52,53, 54]). To demonstrate the quality of **Pred-PFR**, here let us also use the jackknife cross-validation on the benchmark dataset $\mathbb{S}_{\mathrm{bench}}$ (see the Online Supporting Information A).

Now, let us use $\mathrm{PCC}(K_{\mathrm{f}})$ to represent the Pearson Correlation Coefficient (**Eq.9**) obtained with **Pred-PFR** (**Eq.8**) on the benchmark dataset $\mathbb{S}_{\mathrm{bench}}$ by the jackknife cross-validation. For facilitate comparison of the ensemble predictor with the individual predictors, the values of $\mathrm{PCC}(K_{\mathrm{f}})$ and those of $\mathrm{PCC}(K_{\mathrm{f}}^{(i)})$ $(i = 1, 2, \cdots, 7)$ are given in **Table 2.**

Furthermore, to show the accuracy about the prediction in a more intuitive manner, let us introduce the RMSD (Root Mean Square Deviation) as defined by

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{i=1}^{N} (x_i - y_i)^2}{N}} \qquad (11)$$

where x_i, y_i and N have the same meanings as **Eq.9**. Obviously, the smaller the value of RMSD, the more accurate the prediction. If all the predicted results are identical to the corresponding observed ones, we have $\mathrm{RMSD} = 0$.

Similar to the case of PCC, let us use $\mathrm{RMSD}(K_{\mathrm{f}})$ to represent the value of RMSD obtained with the ensemble predictor **Pred-PFR** (**Eq.8**) on the benchmark dataset $\mathbb{S}_{\mathrm{bench}}$ by the jackknife cross-validation, and $\mathrm{RMSD}(K_{\mathrm{f}}^{(i)})$ that by the i-th $(i = 1, 2, \cdots, 7)$ formula of **Eq.7**. All these RMSD values are also given in **Table 2**.

As we can see from the table, the overall PCC value yielded by the ensemble prediction formula (**Eq.8**) is 0.88, which is the closest to 1 in comparison with those by the individual prediction formulae (**Eqs 7.1−7.7**). Such an overall PCC value is even higher than that by the prediction method using the 3D structural information [30] on the same benchmark dataset. Moreover, it can be seen from **Table 2** that the overall RMSD value generated by the ensemble prediction formula is the lowest one in comparison with those by the seven individual prediction formulae. The highest correlation and lowest deviation results indicate that the **Pred-PFR** ensemble predictor formed by the fusing approach is indeed more powerful than the individual predictors.

4. CONCLUSIONS

Pred-PFR is developed for predicting the folding rate of a protein based on its sequence information alone. It is an ensemble predictor formed by fusing multiple individual predictors with each based on one special feature. As expected, the ensemble predictor is superior to the individual predictors. The web-server for **Pred-PFR** is freely accessible to the public at www.csbio.sjtu.edu.cn/bioinf/FoldingRate/.

5. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (Grant no. 60704047), the Science and Technology Commission of Shanghai Municipality (Grant no. 08ZR1410600, 08JC1410600), and sponsored by Shanghai Pujiang Program.

APPENDIX A. THE PROTEIN FOLDING RATE CONSTANT K_{f}

For a given protein, its folding rate is generally reflected by the apparent rate constant K_{f} as defined by the following differential equation

$$\begin{cases} \dfrac{\mathrm{dP}_{\mathrm{unfold}}(t)}{\mathrm{d}t} = -K_{\mathrm{f}} \mathrm{P}_{\mathrm{unfold}}(t) \\ \dfrac{\mathrm{dP}_{\mathrm{folded}}(t)}{\mathrm{d}t} = K_{\mathrm{f}} \mathrm{P}_{\mathrm{unfold}}(t) \end{cases} \qquad (A1)$$

Table 1. The values of the four amino acid properties that have been normalized according to the Max-Min normalization procedure of Eq.3. For more explanation about the four amino acid properties, see the relevant text.

Amino acid code		α_c	β_S	τ	SASA
Single letter	Numerical index j	$\Phi_{1,j}$	$\Phi_{2,j}$	$\Phi_{3,j}$	$\Phi_{4,j}$
A	1	0.58	0.82	0.34	0.21
C	2	0.20	0.25	0.61	0.56
D	3	0.96	0.23	0.12	0.20
E	4	0.90	0.00	0.00	0.29
F	5	0.34	0.12	0.75	0.84
G	6	0.12	0.70	0.28	0.00
H	7	0.09	0.33	0.37	0.51
I	8	0.16	0.33	0.92	0.79
K	9	0.11	0.29	0.27	0.35
L	10	0.10	0.33	0.69	0.69
M	11	0.18	0.38	0.51	0.83
N	12	0.30	0.40	0.39	0.24
P	13	1.00	1.00	0.13	0.23
Q	14	0.45	0.27	0.54	0.39
R	15	0.00	0.73	0.42	0.58
S	16	0.23	0.48	0.28	0.15
T	17	0.47	0.38	0.61	0.27
V	18	0.13	0.42	1.00	0.57
W	19	0.56	0.45	0.75	1.00
Y	20	0.18	0.08	0.82	0.82

Table 2. The jackknife test results by using different formulae on the benchmark dataset $\mathbf{S}_{\mathbf{bench}}$ (see the Online Supporting Information A). [a]Note that PCC may also have negative value (see Eq.9). However, the correlation strength of the predicted results with the observed ones is generally measured by its absolute value.

Prediction formula	PCC [a] (cf. Eq.9)	RMSD (cf. Eq.12)
$\ln K_f^{(1)}$ (see Eq.7.1)	-0.68	3.16
$\ln K_f^{(2)}$ (see Eq.7.2)	0.27	4.17
$\ln K_f^{(3)}$ (see Eq.7.3)	-0.52	3.71
$\ln K_f^{(4)}$ (see Eq.7.4)	-0.39	3.99
$\ln K_f^{(5)}$ (see Eq.7.5)	0.79	2.67
$\ln K_f^{(6)}$ (see Eq.7.6)	0.29	4.14
$\ln K_f^{(7)}$ (see Eq.7.7)	0.85	2.23
$\ln K_f$ (see Eq.8)	**0.88**	**2.03**

where $\mathrm{P}_{\mathrm{unfold}}(t)$ and $\mathrm{P}_{\mathrm{folded}}(t)$ represent the concentrations of its unfolded state and folded state, respectively. Suppose the total protein concentration is C_0, and initially only the unfolded protein is present; i.e., $\mathrm{P}_{\mathrm{unfold}}(t)=C_0$ and $\mathrm{P}_{\mathrm{folded}}(t)=0$ when $t=0$. Subse-quently, the protein system is subjected to a sudden change in temperature, solvent, or any other factor that causes the protein to fold. Obviously, the solution for **Eq.A1** is

$$\begin{cases} \mathrm{P}_{\mathrm{unfold}}(t)=C_0\exp\left(-K_f t\right) \\ \mathrm{P}_{\mathrm{folded}}(t)=C_0\left[1-\exp\left(-K_f t\right)\right] \end{cases} \quad \text{(A2)}$$

It can be seen from the above equation that the larger the K_f, the faster the folding rate will be. However, the actual process is much more complicated than the one as described by **Eq.A1** even if the system concerned consists of only two states. The reason is the folded state may reverse back to the unfolded state, as described by the following equation

$$\mathrm{P}_{\mathrm{unfold}} \underset{k_{21}}{\overset{k_{12}}{\rightleftharpoons}} \mathrm{P}_{\mathrm{folded}} \quad \text{(A3)}$$

where k_{12} is the forward rate constant for $\mathrm{P}_{\mathrm{unfold}}$ converting to $\mathrm{P}_{\mathrm{folded}}$, and k_{21} is the corresponding reverse rate constant. Thus we have the following kinetic equation

$$\begin{cases} \dfrac{\mathrm{dP}_{\mathrm{unfold}}(t)}{\mathrm{d}t}=-k_{12}\mathrm{P}_{\mathrm{unfold}}(t)+k_{21}\mathrm{P}_{\mathrm{folded}}(t) \\ \dfrac{\mathrm{dP}_{\mathrm{folded}}(t)}{\mathrm{d}t}=-k_{21}\mathrm{P}_{\mathrm{folded}}(t)+k_{12}\mathrm{P}_{\mathrm{unfold}}(t) \end{cases} \quad \text{(A4)}$$

Eqs. A3 and **A4** can be expressed by an intuitive graph called directed graph or digraph $\mathbb{G}$ [55,56] as shown in **Fig.1a**. To reflect the variation of the concentrations of unfolded and folded proteins with time, the digraph $\mathbb{G}$ is further transformed to the phase digraph $\tilde{\mathbb{G}}$ as shown in **Fig.1b,** where s is an interim parameter associated with the following Laplace transform

$$\begin{cases} \tilde{\mathrm{P}}_{\mathrm{unfold}}(s) = \int_0^\infty \mathrm{P}_{\mathrm{unfold}}(t)\exp(-ts)\mathrm{d}t \\ \tilde{\mathrm{P}}_{\mathrm{folded}}(s) = \int_0^\infty \mathrm{P}_{\mathrm{folded}}(t)\exp(-ts)\mathrm{d}t \end{cases} \tag{A5}$$

where $\tilde{\mathrm{P}}_{\mathrm{unfold}}$ and $\tilde{\mathrm{P}}_{\mathrm{folded}}$ are the phase concentrations of $\mathrm{P}_{\mathrm{unfold}}$ and $\mathrm{P}_{\mathrm{folded}}$, respectively [55,56]. Thus, using the graphic rule 4 [55,56], also called "Chou's graphic rule for non-steady-state enzyme kinetics" [57], we can immediately obtain the solutions of **Eq.A4**, as given by

$$\begin{cases} \mathrm{P}_{\mathrm{unfold}}(t) = \dfrac{k_{21}C_0}{k_{12}+k_{21}} + \dfrac{k_{12}C_0}{k_{12}+k_{21}}\exp\left[-(k_{12}+k_{21})t\right] \\ \mathrm{P}_{\mathrm{folded}}(t) = \dfrac{k_{12}C_0}{k_{12}+k_{21}} - \dfrac{k_{12}C_0}{k_{12}+k_{21}}\exp\left[-(k_{12}+k_{21})t\right] \end{cases} \tag{A6}$$

Accordingly, it follows

$$\begin{aligned} \frac{\mathrm{dP}_{\mathrm{folded}}(t)}{\mathrm{d}t} &= k_{12}C_0\exp\left[-(k_{12}+k_{21})t\right] \\ &= (k_{12}+k_{21})\mathrm{P}_{\mathrm{unfold}}(t) - k_{21}C_0 \\ &= \left\{\frac{k_{12}(k_{12}+k_{21})}{k_{21}+k_{12}\exp\left[-(k_{12}+k_{21})t\right]}\exp\left[-(k_{12}+k_{21})t\right]\right\}\mathrm{P}_{\mathrm{unfold}}(t) \end{aligned} \tag{A7}$$

Comparing **Eq.A7** with **Eq.A1**, we obtain the following equivalent relation

$$K_{\mathrm{f}} \Leftrightarrow \left\{\frac{k_{12}(k_{12}+k_{21})}{k_{21}+k_{12}\exp\left[-(k_{12}+k_{21})t\right]}\exp\left[-(k_{12}+k_{21})t\right]\right\} \tag{A8}$$

meaning: the apparent folding rate constant K_f is a function of not only the detailed rate constants, but also t. Accordingly, K_f is actually not a constant but will change with time. Only when $k_{12} \gg k_{21}$ and $k_{12} \gg 1$, can **Eq.A8** be reduced to $K_f \simeq k_{12}$ and **Eq.A6** to

$$\frac{\mathrm{dP}_{\mathrm{folded}}(t)}{\mathrm{d}t} \simeq k_{12}\mathrm{P}_{\mathrm{unfold}}(t) = K_f\mathrm{P}_{\mathrm{unfold}}(t) \tag{A9}$$

and K_f be treated as a constant.

It can be imagined that for a three-state or multi-state folding system, K_f will be much more complicated. We can also see from the above derivation that using graphic analysis to deal with kinetic systems is quite efficient and intuitive, particularly in dealing complicated kinetic systems. For more discussions about graphic analysis and its applications to kinetic systems, see [55,58,59,60,61,62].

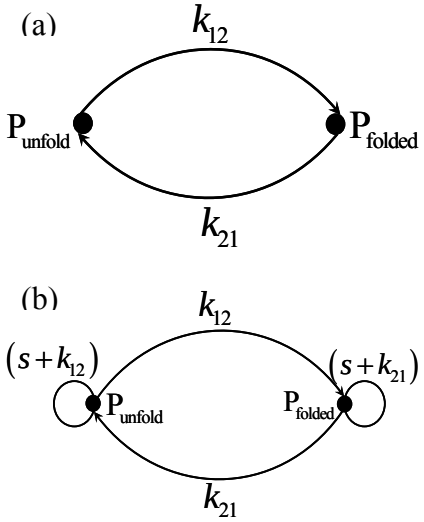

Figure 1. (a) The directed graph or digraph $\mathbb{G}$ [55,56] for the two-state protein folding mechanism as schematically expressed in Eq.A3 and formulated in Eq.A4. **(b)** The phase digraph $\tilde{\mathbb{G}}$ obtained from $\mathbb{G}$ of panel **(a)** according to the graphic rule 4 [55,56], which is also called "Chou's graphic rule for non- steady-state enzyme kinetics" in the literature (see, e.g., [57]). The symbol s in panel **(b)** is an interim parameter (see Eq.A5) and the related text for further explanation).

REFERENCES

[1] Chou, K. C. (2004) Review: Structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry, **11,** 2105−2134.

[2] Anfinsen, C. B. and Scheraga, H. A. (1975) Experimental and theoretical aspects of protein folding. Adv Protein Chem, **29,** 205−300.

[3] Chou, K. C., Nemethy, G., Pottle, M. S. and Scheraga, H. A. (1985) The folding of the twisted beta-sheet in bovine pancreatic trypsin inhibitor. Biochemistry, **24,** 7948−7953.

[4] Creighton, T. E. (1990) Protein folding. Biochem J, **270,** 1−16.

[5] Creighton, T. E. (1995) Protein folding. An unfolding story. Curr Biol, **5,** 353−356.

[6] Scheraga, H. A. (2008) From helix-coil transitions to protein folding. Biopolymers, **89,** 479−485.

[7] Goldberg, M. E., Semisotnov, G. V., Friguet, B., Kuwajima, K., Ptitsyn, O. B. and Sugai, S. (1990) An early immunoreactive folding intermediate of the tryptophan synthease beta 2 subunit is a 'molten globule'. FEBS Lett,

263, 51−56.

[8] Ivankov, D. N. and Finkelstein, A. V. (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA, **101,** 8942−8944.

[9] Anfinsen, C. B. (1973) Principles that govern the folding of protein chains. Science, **181,** 223−230.

[10] Chou, K. C. and Scheraga, H. A. (1982) Origin of the right- handed twist of beta-sheets of poly-L-valine chains. Proceedings of National Academy of Sciences, USA, **79,** 7047−7051.

[11] Chou, K. C., Nemethy, G. and Scheraga, H. A. (1984) Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. Journal of American Chemical Society, **106,** 3161−3170.

[12] Chou, K. C., Maggiora, G. M., Nemethy, G. and Scheraga, H. A. (1988) Energetics of the structure of the four-alpha-helix bundle in proteins. Proceedings of National Academy of Sciences, USA, **85,** 4295−4299.

[13] Klein, P. and Delisi, C. (1986) Prediction of protein structural class from amino acid sequence. Biopolymers, **25,** 1659−1672.

[14] Chou, K. C. and Zhang, C. T. (1992) A correlation coefficient method to predicting protein structural classes from amino acid compositions. European Journal of Biochemistry, **207,** 429−433.

[15] Zhang, C. T. and Chou, K. C. (1992) An optimization approach to predicting protein structural class from amino acid composition. Protein Science, **1,** 401−408.

[16] Chou, J. J. and Zhang, C. T. (1993) A joint prediction of the folding types of 1490 human proteins from their genetic codons. Journal of Theoretical Biology, **161,** 251−262.

[17] Chou, K. C. and Zhang, C. T. (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem, **269,** 22014−22020.

[18] Dubchak, I., Muchnik, I., Holbrook, S. R. and Kim, S. H. (1995) Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci U S A, **92,** 8700−8704.

[19] Chou, K. C. (1995) Does the folding type of a protein depend on its amino acid composition? FEBS Letters, **363,** 127−131.

[20] Chou, K. C. (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function & Genetics, **21,** 319−344.

[21] Bahar, I., Atilgan, A. R., Jernigan, R. L. and Erman, B. (1997) Understanding the recognition of protein structural classes by amino acid composition. PROTEINS: Structure, Function, and Genetics, **29,** 172−185.

[22] Zhou, G. P. (1998) An intriguing controversy over protein structural class prediction. Journal of Protein Chemistry, **17,** 729− 738.

[23] Ding, C. H. and Dubchak, I. (2001) Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics, **17,** 349−358.

[24] Zhou, G. P. and Assa-Munt, N. (2001) Some insights into protein structural class prediction. PROTEINS: Structure, Function, and Genetics, **44,** 57−59.

[25] Ding, Y. S., Zhang, T. L. and Chou, K. C. (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein & Peptide Letters, **14,** 811−815.

[26] Shen, H. B. and Chou, K. C. (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics, **22,** 1717−1722.

[27] Chen, K. and Kurgan, L. (2007) PFRES: protein fold classification by using evolutionary information and predicted secondary structure. Bioinformatics, **23,** 2843−2850.

[28] Shen, H. B. and Chou, K. C. (2009) Predicting protein fold pattern with functional domain and sequential evolution information. Journal of Theoretical Biology, **256,** 441−446.

[29] Chou, K. C. (2005) Review: Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Current Protein and Peptide Science, **6,** 423−436.

[30] Ouyang, Z. and Liang, J. (2008) Predicting protein folding rates from geometric contact and amino acid sequence. Protein Science, **17,** 1256−1263.

[31] Plaxco, K. W., Simons, K. T. and Baker, D. (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol, **277,** 985−994.

[32] Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D. and Finkelstein, A. V. (2003) Contact order revisited: influence of protein size on the folding rate. Protein Science, **12,** 2057−2062.

[33] Zhou, H. and Zhou, Y. (2002) Folding rate prediction using total contact distance. Biophys Journal, **82,** 458−463.

[34] Gromiha, M. M. and Selvaraj, S. (2001) 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 Mol Biol, **310,** 27−32.

[35] Nolting, B., Schalike, W., Hampel, P., Grundig, F., Gantert, S., Sips, N., Bandlow, W. and Qi, P. X. (2003) Structural determinants of the rate of protein folding. J Theor Biol, **223,** 299−307.

[36] Gromiha, M. M., Thangakani, A. M. and Selvaraj, S. (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res, **34,** W70−74.

[37] Wang, D., Keller, J. M., Carson, C. A., McAdo-Edwards, K. K. and Bailey, C. W. (1998) Use of fuzzy-logic-inspired features to improve bacterial recognition through classifier fusion. IEEE Trans Syst Man Cybern B Cybern, **28,** 583−591.

[38] Chou, K. C. and Shen, H. B. (2008) Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols, **3,** 153−162.

[39] Chou, K. C. and Shen, H. B. (2007) Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry, **370,** 1−16.

[40] Chou, K. C. and Zhang, C. T. (1995) Review: Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology, **30,** 275−349.

[41] Chou, P. Y. and Fasman, G. D. (1978) Prediction of secondary structure of proteins from amino acid sequences.

Advances in Enzymology and Related Subjects in Biochemistry, **47,** 45−148.

[42] Iqbal, M. and Verrall, R. E. (1988) Implications of protein folding. Additivity schemes for volumes and compressibilities. J Biol Chem, **263,** 4159−4165.

[43] Oobatake, M. and Ooi, T. (1993) Hydration and heat stability effects on protein unfolding. Prog Biophys Mol Biol, **59,** 237−284.

[44] Jones, D. T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol, **292,** 195−202.

[45] Chou, K. C. (1999) Using pair-coupled amino acid composition to predict protein secondary structure content. Journal of Protein Chemistry, **18,** 473−480.

[46] Zhou, X. B., Chen, C., Li, Z. C. and Zou, X. Y. (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology, **248,** 546−551.

[47] Ding, Y. S. and Zhang, T. L. (2008) Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier. Pattern Recognition Letters, **29,** 1887−1892.

[48] Zhang, G. Y., Li, H. C. and Fang, B. S. (2008) Predicting lipase types by improved Chou's pseudo-amino acid composition. Protein & Peptide Letters, **15,** 1132−1137.

[49] Lin, H. (2008) The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. Journal of Theoretical Biology, **252,** 350−356.

[50] Li, F. M. and Li, Q. Z. (2008) Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach. Protein & Peptide Letters, **15,** 612− 616.

[51] Zhang, G. Y. and Fang, B. S. (2008) Predicting the cofactors of oxidoreductases based on amino acid composition distribution and Chou's amphiphilic pseudo amino acid composition. Journal of Theoretical Biology, **253,** 310−315.

[52] Lin, H., Ding, H., Feng-Biao Guo, F. B., Zhang, A. Y. and Huang, J. (2008) Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition. Protein & Peptide Letters, **15,** 739−744.

[53] Munteanu, C. R., Gonzalez-Diaz, H., Borges, F. and de Magalhaes, A. L. (2008) Natural/random protein classification models based on star network topological indices. Journal of Theoretical Biology, **254,** 775−783.

[54] Rezaei, M. A., Abdolmaleki, P., Karami, Z., Asadabadi, E. B., Sherafat, M. A., Abrishami-Moghaddam, H., Fadaie, M. and Forouzanfar, M. (2008) Prediction of membrane protein types by means of wavelet analysis and cascaded neural networks. Journal of Theoretical Biology, **254,** 817−820.

[55] Chou, K. C. (1989) Graphical rules in steady and non-steady enzyme kinetics. J Biol Chem, **264,** 12074−12079.

[56] Chou, K. C. (1990) Review: Applications of graph theory to enzyme kinetics and protein folding kinetics. Steady and non- steady state systems. Biophysical Chemistry, **35,** 1−24.

[57] Lin, S. X. and Neet, K. E. (1990) Demonstration of a slow conformational change in liver glucokinase by fluorescence spectroscopy. J Biol Chem, **265,** 9670−9675.

[58] Chou, K. C. and Liu, W. M. (1981) Graphical rules for non-steady state enzyme kinetics. Journal of Theoretical Biology, **91,** 637−654.

[59] Zhou, G. P. and Deng, M. H. (1984) An extension of Chou's graphical rules for deriving enzyme kinetic equations to system involving parallel reaction pathways. Biochemical Journal, **222,** 169−176.

[60] Myers, D. and Palmer, G. (1985) Microcomputer tools for steady-state enzyme kinetics. Bioinformatics (original: Computer Applied Bioscience), **1,** 105−110.

[61] Kuzmic, P., Ng, K. Y. and Heath, T. D. (1992) Mixtures of tight- binding enzyme inhibitors. Kinetic analysis by a recursive rate equation. Anal Biochem, **200,** 68−73.

[62] Andraos, J. (2008) Kinetic plasticity and the determination of product ratios for kinetic schemes leading to multiple products without rate laws: new methods based on directed graphs. Canadian Journal of Chemistry, **86,** 342−357.

