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

Hong-Bin Shen¹³,*, Jiang-Ning Song², Kuo-Chen Chou¹³

¹Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, 800 Dongchuan Road, Shanghai, 200240, China; ²Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan; ³Gordon 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/Folding Rate/.

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 \((N_G)\) \([30]\).

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 \((L_{\text{eff}})\) \([8]\) 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-\(\alpha\) proteins, 32 all-\(\beta\), and the remaining 30 are \(\alpha\beta\) proteins (where \(\alpha\beta\) means the mix of \(\alpha/\beta\) and \(\alpha+\beta\) \([40]\)). The folding rates of the 80 proteins range from \(\ln K_t = -6.9\) to \(\ln K_t = 12.9\), spanning more than eight orders of magnitude of \(K_t\). For users’ convenience, the benchmark dataset, denoted as \(S_{\text{bench}}\), 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/](http://www.csbio.sjtu.edu.cn/bioinf/FoldingRate/). It is instructive to point out that \(K_t\) in \(S_{\text{bench}}\) is actually an apparent folding rate constant (see Appendix A). Therefore, to develop a statistical method for predicting \(K_t\) 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: \(\alpha_c\), the propensity to be at the C-terminal of \(\alpha\)-helix \([41]\); \(\beta_s\), the propensity to form \(\beta\)-strand \([41]\); \(\tau\), the compressibility \([42]\); and \(SASA\), the solvent accessible surface area in an unfolding protein chain \([43]\). Suppose a protein \(P\) is expressed by

\[
P = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \cdots R_j
\]

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

\[
\Phi_j = \frac{\phi_{i,j}}{L} \quad (i = 1, 2, 3, 4)
\]

where \(L\) represents the protein length, and

\[
\phi_{i,j} = \frac{\phi_{i,j}^e}{\max_j \{\phi_{i,j}^e\} - \min_j \{\phi_{i,j}^e\}} \quad (i = 1, 2, 3, 4; j = 1, 2, \cdots, 20)
\]

where \(\phi_{i,j}^e\) \((i = 1, 2, 3, 4)\) respectively represent the original \(\alpha_c\), \(\beta_s\), \(\tau\), and \(SASA\) for the \(j\)-th native amino acid, and their values can be obtained from \([41, 42, 43]\): \(\max_j \{\phi_{i,j}^e\}\) means taking the maximum one among \(\phi_{i,j}^e\), \(\phi_{i,j}^e\), \(\cdots\), \(\phi_{i,j}^e\), and \(\min_j \{\phi_{i,j}^e\}\) 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 \((L^{1/2}, L^{2/3}, \text{or } L^{3/5})\) or logarithm \(\ln(L)\) have a good correlation with the folding rates, suggesting that
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 \( \mathbf{P} \), as formulated by

\[
\Gamma_{\alpha} + \Gamma_{\beta} + \Gamma_{C} = 1 \tag{4}
\]

where \( \Gamma_{\alpha}, \Gamma_{\beta} \), and \( \Gamma_{C} \) are the ratios of the \( \alpha \)-helix, \( \beta \)-sheet, and coiled-coil residues for the protein \( \mathbf{P} \). 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]:

\[
L_{\text{eff}} = L - L_{\text{H}} + L_{n} \cdot N_{\text{H}} \quad \tag{5}
\]

where \( L \) is the total number of amino acids for the entire protein chain; \( L_{\text{H}} \) the number of predicted helical conformation residues; \( N_{\text{H}} \) the number of predicted helices; and \( L_{n} \) the number of an \( \alpha \)-helix turn (\( L_{n} \) is generally \( \leq 4 \); for a standard \( \alpha \)-helix, \( L_{n} = 3.6 \)). In the current study, \( L_{n} \) 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:

\[
S_{\text{feature}} = \begin{cases} 
\Phi_1 = a_c & \\
\Phi_2 = \beta_s & \\
\Phi_3 = \tau & \\
\Phi_4 = \text{SASA} & \\
\Phi_5 = \ln(L) & \\
\Phi_6 = (\Gamma_{\alpha}, \Gamma_{\beta}, \Gamma_{C}) & \\
\Phi_7 = \ln(L_{\text{eff}}) &
\end{cases} \tag{6}
\]

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:

\[
\ln K_f^{(i)} = a_i + b_i \cdot \Phi_i \quad i = 1, 2, \ldots, 7 \tag{7}
\]

where \( K_f^{(i)} \) \((i = 1, 2, \ldots, 7)\) is the protein folding rate constant predicted based on the \( i \)-th specific feature \( \Phi_i \) (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 \( S_{\text{bench}} \). For the details of how to use the regression procedures to determine \( a_i \) and \( b_i \), refer to [45]. Note that \( K_f^{(6)} \) of Eq.(7.6) is involved with more parameters because the 6-th feature \( \Phi_6 \) 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:

\[
\ln K_f = \sum_{i=1}^{2} w_i \ln K_f^{(i)} \quad \tag{8}
\]

where \( w_i \) is the weight that reflects the impact of the \( i \)-th specific feature \( \Phi_i \) on the protein folding rate. If the impacts of the seven features were the same, we should have \( w_i = 1/7 \quad (i = 1, 2, \ldots, 7) \). 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

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

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 $\text{PCC}=1$. For different prediction algorithms, Eq. 9 will yield different values of $\text{PCC}$. Therefore, the weight $w_i$ in Eq. 8 can be formulated as

$$w_i = \frac{\text{PCC}(K_i^{(j)})}{\sum_{j=1}^{7} \text{PCC}(K_i^{(j)})} \quad (i = 1, 2, \cdots, 7) \quad (10)$$

where $\text{PCC}(K_i^{(j)})$ is the Pearson Correlation Coefficient (Eq. 9) obtained with the $i$-th folding rate predicting formula in Eq. 7 on the benchmark dataset $S_{\text{bench}}$ 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 DISCUSSIONS

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 $S_{\text{bench}}$ (see the Online Supporting Information A).

Now, let us use $\text{PCC}(K_i)$ to represent the Pearson Correlation Coefficient (Eq. 9) obtained with Pred-PFR (Eq. 8) on the benchmark dataset $S_{\text{bench}}$ by the jackknife cross-validation. For facilitate comparison of the ensemble predictor with the individual predictors, the values of $\text{PCC}(K_i)$ and those of $\text{PCC}(K_i^{(j)})$ ($i = 1, 2, \cdots, 7$) 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

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

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 $\text{RMSD} = 0$.

Similar to the case of $\text{PCC}$, let us use $\text{RMSD}(K_i)$ to represent the value of RMSD obtained with the ensemble predictor Pred-PFR (Eq. 8) on the benchmark dataset $S_{\text{bench}}$ by the jackknife cross-validation, and $\text{RMSD}(K_i^{(j)})$ 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 $K_i$

For a given protein, its folding rate is generally reflected by the apparent rate constant $K_i$ as defined by the following differential equation

$$\begin{align*}
\frac{dp_{\text{unfold}}(t)}{dt} &= -K_i p_{\text{unfold}}(t) \\
\frac{dp_{\text{folded}}(t)}{dt} &= K_i p_{\text{unfold}}(t)
\end{align*} \quad (A1)$$
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.

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

Table 2. The jackknife test results by using different formulae on the benchmark dataset bench (see the Online Supporting Information 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.

<table>
<thead>
<tr>
<th>Prediction formula</th>
<th>PCC^{a} (cf. Eq.9)</th>
<th>RMSD (cf. Eq.12)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ln K^{(1)}_{t}</td>
<td>-0.68</td>
<td>3.16</td>
</tr>
<tr>
<td>ln K^{(2)}_{t}</td>
<td>0.27</td>
<td>4.17</td>
</tr>
<tr>
<td>ln K^{(3)}_{t}</td>
<td>-0.52</td>
<td>3.71</td>
</tr>
<tr>
<td>ln K^{(4)}_{t}</td>
<td>-0.39</td>
<td>3.99</td>
</tr>
<tr>
<td>ln K^{(5)}_{t}</td>
<td>0.79</td>
<td>2.67</td>
</tr>
<tr>
<td>ln K^{(6)}_{t}</td>
<td>0.29</td>
<td>4.14</td>
</tr>
<tr>
<td>ln K^{(7)}_{t}</td>
<td>0.85</td>
<td>2.23</td>
</tr>
<tr>
<td>ln K^{(8)}_{t}</td>
<td>0.88</td>
<td>2.03</td>
</tr>
</tbody>
</table>

where P_{\text{unfold}}(t) and P_{\text{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., P_{\text{unfold}}(t) = C_{0} and P_{\text{folded}}(t) = 0 when t = 0. Subsequently, 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

\[
P_{\text{unfold}}(t) = C_0 \exp(-K_t t)
\]

\[
P_{\text{folded}}(t) = C_0 [1 - \exp(-K_t t)]
\]

It can be seen from the above equation that the larger the K_t, 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

\[
P_{\text{unfold}} \xleftrightarrow{k_{21}} P_{\text{folded}}
\]

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

\[
\begin{align*}
\frac{dP_{\text{unfold}}(t)}{dt} &= -k_{12}P_{\text{unfold}}(t) + k_{21}P_{\text{folded}}(t) \\
\frac{dP_{\text{folded}}(t)}{dt} &= -k_{21}P_{\text{folded}}(t) + k_{12}P_{\text{unfold}}(t)
\end{align*}
\]
Eqs. A3 and A4 can be expressed by an intuitive graph called directed graph or digraph $\mathcal{G}$ [55,56] as shown in Fig. 1a. To reflect the variation of the concentrations of unfolded and folded proteins with time, the digraph $\mathcal{G}$ is further transformed to the phase digraph $\mathcal{G}'$ as shown in Fig. 1b, where $s$ is an interim parameter associated with the following Laplace transform

$$
\tilde{P}_{\text{unfold}}(s) = \int_0^\infty P_{\text{unfold}}(t) \exp(-st) dt \\
\tilde{P}_{\text{folded}}(s) = \int_0^\infty P_{\text{folded}}(t) \exp(-st) dt
$$

(A5)

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

$$
K_f \Rightarrow \left[\frac{k_{12} (k_{12} + 2k_{21})}{k_{21} + k_{12} \exp(-(k_{12} + 2k_{21})t)} \exp(-(k_{12} + 2k_{21})t)\right]
$$

(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 $k_{12} \gg 1$, $K_f$ can 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. 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

$$
P_{\text{unfold}}(t) = \frac{k_{21} C_0}{k_{12} + k_{21}} \exp\left[-(k_{12} + 2k_{21})t\right] + \frac{k_{12} C_0}{k_{12} + k_{21}} \exp\left[-(k_{12} + k_{21})t\right] = \tilde{P}_{\text{unfold}}(s) \mid_{s = -k_{12}}
$$

(A6)

$$
P_{\text{folded}}(t) = \frac{k_{12} C_0}{k_{12} + k_{21}} \exp\left[-(k_{12} + k_{21})t\right] = \tilde{P}_{\text{folded}}(s) \mid_{s = -k_{12}}
$$

(A7)

Accordingly, it follows can Eq. A8 be reduced to $K_f = k_{12}$ and Eq. A6 to

$$
\frac{dP_{\text{folded}}(t)}{dt} = k_{12} P_{\text{unfold}}(t) = K_f P_{\text{unfold}}(t)
$$

(A9)


REFERENCES

263, 51–56.


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


