Predicting Protein Folding Rates Using the Concept of Chou's Pseudo Amino Acid Composition

JIANXIU GUO, NINI RAO, GUANGXIONG LIU, YONG YANG, GANG WANG
School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, People's Republic of China

Received 22 April 2010; Revised 4 November 2010; Accepted 2 December 2010
DOI 10.1002/jcc.21740
Published online 15 February 2011 in Wiley Online Library (wileyonlinelibrary.com).

Abstract: One of the most important challenges in computational and molecular biology is to understand the relationship between amino acid sequences and the folding rates of proteins. Recent works suggest that topological parameters, amino acid properties, chain length and the composition index relate well with protein folding rates, however, sequence order information has seldom been considered as a property for predicting protein folding rates. In this study, amino acid sequence order was used to derive an effective method, based on an extended version of the pseudo-amino acid composition, for predicting protein folding rates without any explicit structural information. Using the jackknife cross validation test, the method was demonstrated on the largest dataset (99 proteins) reported. The method was found to provide a good correlation between the predicted and experimental folding rates. The correlation coefficient is 0.81 (with a highly significant level) and the standard error is 2.46. The reported algorithm was found to perform better than several representative sequence-based approaches using the same dataset. The results indicate that sequence order information is an important determinant of protein folding rates.

© 2011 Wiley Periodicals, Inc. J Comput Chem 32: 1612–1617, 2011

Key words: protein folding rate; pseudo-amino acid composition; sequence-based prediction; jackknife test; linear regression

Introduction

Protein folding is the process by which a protein processes from its denatured state to its specific biologically active conformation, and different proteins have significantly different rates of folding. Some proteins fold within microseconds, whereas other proteins require hours to fold. Studies of protein folding rates can enhance our understanding of the variations in protein folding kinetics, which may be related to several pathologies such as prion and Alzheimer diseases. Understanding the relationship between sequences and the folding rates of proteins remains an important challenge.[1] Although understanding protein folding rates can be characterized using various biochemical experiments,[2–7] it is not only time-consuming but also very costly to carry out such experiments. This is particular the case with the rapid increase in the number of newly found protein sequences, with the gap between the number of newly discovered protein sequences and the characterization of their folding rates widening rapidly. Consequently, the development of an automated and fast method to predict protein folding rates represents an important milestone to achieve.

During the past few years, many efforts have been made in this field. Plaxco et al.[8] found that the contact order (CO) (reflects the relative importance of local and nonlocal contacts to the native structure of a protein) showed a significant correlation with folding rates of small, two-state proteins. Subsequently, many variations of this idea have been studied, which indicated that folding rates also correlated with the long-range order,[9] total contact distance,[10] effective contact order,[11] chain topology parameter[12] and the absolute contact order.[13] These parameters are based on the 3D-structure of proteins. In addition, several structural parameters have been developed to predict the protein folding rates from protein secondary structures (or predicted secondary structures), such as the effective length of a folding chain (L_{eff})[14] and the local secondary structure contents.[15] There have also been several reports detailing the prediction of protein folding rates from primary sequences, for example, the helix parameter,[16] chain length (L),[17] predicted long-range contacts,[18] amino acid properties (Ω),[19] average amino acid property[20] and Fold-Rate.[21] However, these attempts still require some level of structural information, for example, knowledge of the structure classes. Ma et al.[22] constructed an indicator called the composition index (CI) to predict protein folding rates from amino

Additional Supporting Information may be found in the online version of this article.

Correspondence to: N. Rao; e-mail: raonn@uestc.edu.cn.

Contract/grant sponsor: National Natural Science Foundation of China; contract/grant numbers: 30900318, 60801058

acid sequences without any knowledge of the tertiary or secondary structures, or information on the structural class. Unfortunately, this method was only based on the conventional amino acid composition and did not take into account mutations that did not change the occurrence numbers of amino acids but changed order of appearance. These mutations may result in a large change in the folding rates. The recently proposed sequence-based methods[23–28] still did not consider the influence of the sequence order on the folding rates. Inclusion of the sequence order information in the prediction algorithm would lead to an improvement in the prediction quality, and reveal interesting insights into the folding process. Clearly, the sequence order effects are lost if only the conventional amino acid composition is used as the representation for a protein sample.

Pseudo-amino acid composition (PseAAC)[29] was originally introduced to improve the prediction quality for protein subcellular localization[30] and membrane protein type.[31] PseAAC can be used to represent a protein sequence with a discrete model without completely discarding the sequence order information, and hence it has been widely applied for improving the prediction quality of various protein attributes. In the present study, to take into account the sequence order effects, we present an algorithm that adopts the concept of the Chou's PseAAC feature extraction method. This algorithm predicts protein folding rates from amino acid sequences without any knowledge of the tertiary or secondary structures, or structural class information. Evaluated by the jackknife cross-validation test, the predicted folding rates correlate well with the experimental values. The correlation coefficient is 0.81 and the standard error is 2.46. Finally, we compare the performance of our algorithm with several published methods on the same dataset. The experimental results show that our algorithm achieves a higher accuracy than most existing sequence-based methods.

Materials and Methods

Protein Dataset

From the literatures,[13–15,20–25,32–35] we have collected 117 proteins with known experimentally determined folding rates. Protein sequence homology in datasets is known to influence the prediction accuracy, i.e., the prediction accuracy will be overestimated when using highly homologous protein sequences. Thus, to strictly test the current method and facilitate a comparison, the data were screened according to the following procedures. In the first procedure, for proteins with the same name, but from different species, only one was included. In the second procedure, sequences containing ambiguous residue like "X" were excluded. In the third procedure, homologous proteins by comparison with the Uniprot sequence (http://www.uniprot.org/) were removed from our dataset. After these three screening procedures, we obtained a dataset of 99 proteins. The Protein Data Bank codes and the experimental folding rate values $\ln(k_f)$ are listed in Supporting Information Table I. Amino acid sequences of each protein are taken from the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do).

The Pseudo-Amino Acid Composition

To avoid losing key information hidden in protein sequence, the PseAAC[29,36] was used rather than the simple amino acid composition for representing a protein. For a brief introduction on Chou's PseAAC, please visit the Wikipedia web-page at http://en.wikipedia. org/wiki/Pseudo_amino_acid_composition. For a summary about its development and applications, see a recent comprehensive review.[37] Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems, such as predicting subcellular location of proteins,[38–42] subnuclear location of proteins,[43,44] structural classes of proteins,[45] submitochondria localization,[46] protein quaternary structure,[47] GPCR types,[48,49] human papillomaviruses,[50] outer membrane proteins,[51] protein secondary structural content,[52] subcellular localization of mycobacterial proteins,[53] lipase types,[54] DNA-binding proteins,[55] enzyme family class,[56] enzyme subfamily classes,[57] cell wall lytic enzymes,[58] cofactors of oxidoreductases,[59] submitochondria locations,[60] among many other protein attributes and protein related features. However, so far no report whatsoever has been seen that the PseAAC was used for predicting the folding rates of proteins. The present study is devoted to do so, and very encouraging results have been obtained.

Given a protein sequence P with L amino acid resides, it can be formulated as:

$$P = R_1 R_2 R_3 R_4 \cdots R_L \tag{1}$$

where R_1 represents the 1st residue of a protein P, R_2 the 2nd residue, and so forth. According to Chou's PseAAC,[29] the affection of a protein sequence order can be, to some extent, reflected through a set of sequence correlation factors θ_1, θ_2, $\ldots\theta_\lambda$. It original formulation can be briefly described as follows:

$$\begin{cases} \theta_1 = \frac{1}{L-1}\sum\limits_{i=1}^{L-1}\Theta(R_i, R_{i+1}) \\ \theta_2 = \frac{1}{L-2}\sum\limits_{i=1}^{L-2}\Theta(R_i, R_{i+2}) \\ \theta_3 = \frac{1}{L-3}\sum\limits_{i=1}^{L-3}\Theta(R_i, R_{i+3}) \quad (\lambda < L) \\ \quad\cdots \\ \quad\cdots \\ \theta_\lambda = \frac{1}{L-\lambda}\sum\limits_{i=1}^{L-\lambda}\Theta(R_i, R_{i+\lambda}) \end{cases} \tag{2}$$

where θ_1 is called the first-tier correlation factor that reflects the sequence order correlation between all the most contiguous residues along a protein chain, as presented in Figure 1a;[29] θ_2 the second-tier correlation factor that reflects the sequence order correlation between all the second most contiguous residues (Fig. 1b);[29] θ_3 the third-tier correlation factor that reflects the sequence order correlation between all the third most contiguous residues (Fig. 1c);[29] and so forth. A protein sample can be represented by a vector or a point in (20 + λ-D) space. In other words, a protein can be expressed as a vector $X = (x_1, \ldots x_{20},\ x_{20+1}, \ldots x_{20+\lambda})$. The first 20 components reflect the effect of the amino acid composition, whereas the components from 20+1 to 20+λ reflect the effect of the sequence order.

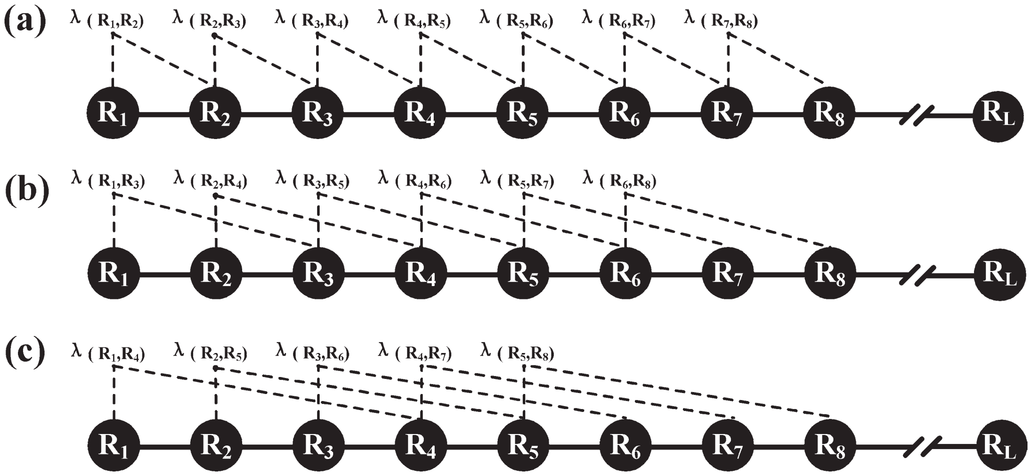

Figure 1. A schematic drawing to show the sequence order correlation mode along a protein sequence. The first-tier panel (a) reflects the correlation mode between all the most contiguous residues, the second-tier panel (b) that between all the second-most contiguous residues, and the third-tier panel (c) that between all the third-most contiguous residues.

Sequence Feature Extraction

Many studies[17,61–63] have indicated that the protein chain length L and its fractional powers ($L^{1/2}$ and $L^{2/3}$) or logarithm $\ln(L)$ show a good correlation with the folding rates, which indicates that L and its various expression forms can be useful features for predicting folding rates. In our study, $\ln(L)$ is adopted as the first part of the feature vector.

Ma et al.[22] have reported that the amino acid composition of a protein sequence is strongly correlated with the folding rate. In this work, for the first 20 components in the vector X, the occurrence frequencies of amino acid A, G, N, Q, T and V are selected as the second part of the feature vector according to the correlation coefficient between protein folding rates and its amino acid contents, where we take 0.3 as the relatively significant threshold. This result is consistent with previously reported work.[22]

Since the affection of a protein sequence order can be approximately reflected with a set of sequence correlation factors $\theta_1,\theta_2,\ldots\theta_\lambda$ (the components from $20+1$ to $20+\lambda$ in the vector X), the third part of the feature vector can be selected from the sequence correlation factor θ. In our algorithm, the correlation function in Eq. (2) is given by

$$\Theta(R_i,R_j) = \left|H(R_j) - H(R_i)\right| \qquad (3)$$

where $H(R_j)$ is the original hydrophobic value of an amino acid R_j that is taken from Tanford.[64] Kauzmann[65] has pointed out that hydrophobic interactions are the dominant force driving protein folding. Hence, the hydrophobic values may be used to effectively reflect the sequence order effects. For $\theta_1,\theta_2,\ldots,\theta_\lambda$, which represents the dominant factor correlated with the folding rates and which represents redundant information? The correlation-based feature selection process[66] is performed as follows:

Initially, the full sequence correlation factor θ is used to develop a linear regression model using the jackknife test and the corresponding correlation coefficient between the predicted and the actual folding rates is computed. Next, a given factor is removed only if the removal does not decrease the correlation coefficient of the prediction using the reduced set of factors. After the factor is removed, the process is continually repeated until no factor can be removed. This feature selection method has also been successfully applied to design the PPFR method.[25] The aim of this procedure is to ensure that the selected features do not overfit the dataset and improve the predicted accuracy.

Previous investigations[29] indicated that the optimal value for λ should be the one that results in the best overall jackknife test (will be described in the next section). We have tried different values of λ in our method, and finally found $\lambda = 10$ can be used as the optimal value for the dataset. The above feature selection process is performed for the 10 features (the full sequence correlation factor θ). The final retained features are $\theta_2,\theta_4,\theta_9$ and θ_{10}. They are taken as the third part of the feature vector involving some sequence order information.

Evaluation Method

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.[67] However, as elucidated previously by Chou[68] and demonstrated by Eq. (50) in ref. 69, among the three cross-validation methods, the jackknife test is regarded as one of the most effective and objective methods for cross-validation in statistics because it can always yield an unique result and obtain a more accurate estimation of the prediction accuracy for a given benchmark dataset, and hence has been increasingly used and widely recognized by investigators for examining the accuracy of various predictors.[40–42,49,51–53,57,58,60,70,71] Accordingly, in this study the jackknife test was adopted to evaluate the prediction method as well. In the jackknife test, each protein in the dataset is singled out in turn as an independent testing sample, and all the rule parameters are calculated without using this protein.

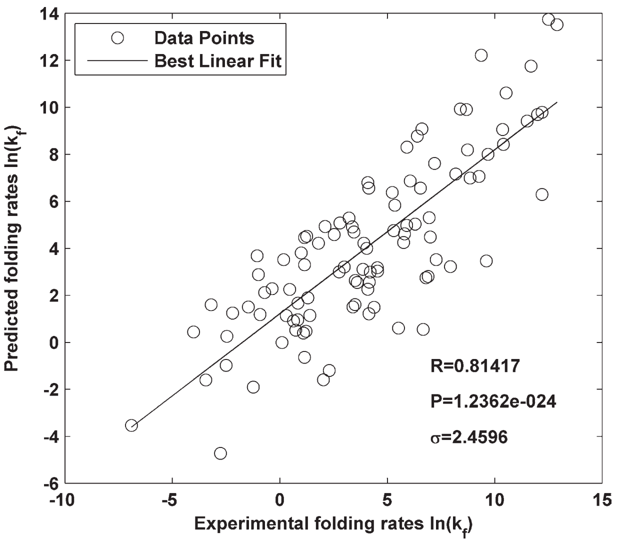

Figure 2. Relationship between the experimental and predicted folding rates using linear regression model with jack-knife test for a set of 99 proteins.

Results and Discussion

Prediction Results

According to the above analysis, we used a 1 + 6 + 4 = 11 dimension vector to represent a protein sequence, which was analyzed via a simple linear regression model. The jackknife test was performed to demonstrate the quality of the method. The predicted folding rates of all the proteins in the dataset are shown in Supporting Information Table I. The relationship between the experimental and predicted folding rates is illustrated in Figure 2. The final correlation coefficient R is 0.81 with a highly significant level ($p = 1.23e{-}24$) and the standard error σ is 2.46. The results show that the method achieves a good prediction performance. If the method were trained on 99 proteins, the fitting equation would be given as follows:

$$\begin{aligned}\text{Folding\_rate} = {} & 23.06f_{\rm A} + 6.68f_{\rm G} + 23.86f_{\rm N} + 19.26f_{\rm Q} \\ & -12.18f_{\rm T} - 21.49f_{\rm V} + 10.06\theta_2 + 3.64\theta_4 - 7.71\theta_9 \\ & - 4.01\theta_{10} - 6.32\ln(L) + 28.04 \qquad (4)\end{aligned}$$

where $f_{\rm i}$ is the occurrence frequencies of amino acid i. The folding rate of a new protein can be computed by it.

In this study, the sequence correlation factors, θ_2,θ_4,θ_9 and θ_{10}, play a dominant role in the folding rate of a protein. Interestingly, we observed a strong correlation between folding rate and the factor θ_2 as well as the factor θ_4. The two factors may represent the effects of the local interactions between residues (short and medium range) along the polypeptide chain. θ_2 may represent the sum of the inter-residue interaction in the sequence level in the short range (=±2 residues), and θ_4 may represent the sum of the interactions in the medium range (=±4 residues). The recent works[72–75] are manifested here. However, the factors θ_9 and θ_{10} were found to not correlate well with protein folding rates, but play an important role in increasing the final correlation coefficient. If we remove factors θ_9 and θ_{10} from the final feature vector X, the prediction accuracy decreased to 0.77 (with a standard error of 2.71) for the same dataset. Besides the θ_9 and θ_{10} may represent the effect of the long range interactions, this result is perhaps consistent with previously reported results[76] which suggest that only increases in the hydrophobicity at positions which are not involved in the folding nucleus can accelerate the folding process. Recent efforts[77–79] have reported that seventeen amino acids were prone to fold into a helix structure; a particular residue only maintained a correlation with eight sequential residues N- and C-terminal to this residue. As such, the ninth and tenth residues are not involved in the folding nucleus. In this work, the sequence correlation factors θ_9 and θ_{10} are suitably the sum of the position information of the ninth and tenth residues, respectively.

Comparison with Different Methods

A direct comparison of the correlation coefficients obtained in this work and other methods is not appropriate due to the use of different datasets. The prediction quality of the method herein is measured based on the correlation coefficient between the predicted and experimental folding rates using the jackknife test on 99 proteins, whereas four existing representative sequence-based measures, including CI,[22] Fold-rate,[21] QRSM[23] and N_α^{24} were designed and tested on 62 proteins, 77 proteins, 77 proteins (there are some error folding rate values and repeat proteins in the original literature) and 80 proteins (we note that the number of proteins in the text is 80, while the actual number is 78 in the table in the original literature), respectively. To facilitate a comparison, the same dataset should be used here. For comparing the performance of the predictors, the following two ways were used. (1) Based on the jackknife test. All the methods are performed on the dataset used in this study. The results of the prediction quality are summarized in Table 1. (2) Based on the independent dataset test. Our method is trained on the data used in other work and tested on the remaining data, and the performances of the four existing representation sequence-based measures are also tested on the remaining data. The prediction qualities are shown in Table 2. For example, CI was designed and tested on 62 proteins (D62) in the original literature. The dataset used in this study includes all sequences from the dataset D62 and 37

Table 1. Performances of Different Methods in Predicting Protein Folding Rates Using Jackknife Test.

Method	R-value	P-value	σ-value
Fold-rate	0.23	0.023	5.45
CI	0.71	3.63e-17	3.93
QRSM	0.45	2.58e-6	11.80
N_α	0.40	3.24e-5	8.45
Current	0.81	1.24e-24	2.46

All methods are tested on the same testing dataset (99 proteins). R-value denotes the correlation coefficient. P-value denotes the significant level. σ-value denotes the standard error.

Table 2. Performances of Different Methods in Predicting Protein Folding Rates Using Independent Dataset Test.

Method (dataset)	R-value	P-value	σ-value
CI (62)[a]	0.73	2.51e-17	3.79
Current (62)[a]	0.80	3.30e-9	2.62
Fold-rate (77)[b]	0.26	0.19	7.70
QRSM (77)[b]	0.33	0.096	12.85
Current (77)[b]	0.90	1.56e-10	2.11
N_α (80)[c]	0.35	0.11	7.47
Current (80)[c]	0.51	0.012	3.41

R-value denotes the correlation coefficient; P-value denotes the significant level; σ-value denotes the standard error.

[a] The methods are trained on 62 proteins and tested on the left 37 chains from 99 proteins. The result of CI is from the web server at http://sdbi.sdut.edu.cn/FDserver.

[b] The methods are trained on 77 proteins and tested on the left 22 chains from 99 proteins. The result of fold-rate is from the web server at http://psfs.cbrc.jp/fold-rate/. The result of QRSM is from the web server at http://210.60.98.17/FOLDRATE20r/foldrate20.htm.

[c] The methods are trained on 80 proteins and tested on the left 19 chains from 99 proteins. The result of N_α is from the web server at http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html.

chains (D37) are left. So the current method and CI are trained on D62 and tested on D37. Similarly, the dataset used in this study include all sequences from the other three datasets.

Tables 1 and 2 report that the correlation coefficient (R) of our approach is higher than the results of the other methods using the same benchmark datasets. Moreover, the significant level (P) and the standard error (σ) are the lowest using the presented method. The highest correlation coefficient and the lowest prediction deviation indicate that the prediction quality of our method is remarkably improved after taking into account the sequence order information. Thus, this method can serve as an effective tool to predict protein folding rates from amino acid sequences without any explicit structural information.

Limitations of the Present Method and Possible Improvements

Although the current method provides an accurate means to predict protein folding rates, there are a number of limitations and possible improvements to be implemented. (1) The folding rates of proteins depend on experimental conditions, which are not considered in the present work. For example, the folding rate of the protein 1YCC (the deviation of the linear fitting between the predicted and experimental folding rates is the largest and the experimental value is larger than the predicted value), was measured at an elevated temperature of 40°C rather than 25°C. At room temperature, it is anticipated that the fit to the regression line in Figure 2 would be significantly improved; the proteins 1PGB:B and 1L2Y:A (the most extreme outliers in Fig. 2) are two short artificial peptides. The folding rates of these peptides are different when compared to nature proteins. The correlation in Figure 2 may be improved if these proteins are removed from the dataset as carried in previous studies.[22,23,25–27] (2) Only one kind of physicochemical property, hydrophobicity, was used in this approach. More sequence order information would be involved if we selected more than one type of physicochemical property. Overall, a remarkable improvement in terms of prediction quality should be observed using a larger and more "standard"[7] experimental dataset and by selecting additional physicochemical properties. (3) In the present investigation, the main purpose was to develop a general, fast and accurate method to predict the folding rates of proteins based only on protein sequence, without any explicit structural information, or information of the structural class, and without prior knowledge of the folding type. However, two and three state protein folding involves different processes which may be governed by different factors. Consequently, revealing the influence of these different factors on the rates of protein folding represent future research efforts. (4) Since user-friendly and publicly accessible web servers represent a future direction for developing practically more useful predictors,[80] we are currently investigating the use of a web server for the method presented in this article.

Conclusions

Current algorithms for predicting protein folding rates do not use sequence order information. This is due to the extremely large numbers of sequence order patterns in proteins and their diverse lengths have made it very difficult to accommodate the sequence order effects. To tackle this issue, according to the concept of Chou's PseAAC, a new method was presented to approximate the sequence order effects. The method was shown to predict protein folding rates from amino acid sequences without any explicit structural information (including known native structure, known/predicted secondary structure, structural class information and prior knowledge of the folding type). The research results show that the correlation coefficient between the experimental and predicted folding rates for 99 nonhomologous proteins reaches 0.81 (with a highly significant level) and the standard error is 2.46 based on the jackknife test. This accuracy illustrates that this method is better than most of the sequence-based only predictors.

Acknowledgments

The authors thank the two anonymous referees whose constructive comments are very helpful for improving the manuscript.

References

1. Eaton, W. A.; Muñoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Annu Rev Biophys Biomol Struct 2000, 29, 327–359.
2. Zeeb, M.; Balbach, J. Methods 2004, 34, 65–74.
3. Fabian, H.; Naumann, D. Methods 2004, 34, 28–40.
4. Zarrine-Afsara, A.; Davidson, A. R. Methods 2004, 34, 41–50.
5. Maity, H.; Maity, M.; Krishna, M. M. G.; Mayne, L.; Englander, S. W. Proc Nat Acad Sci USA 2005, 102, 4741–4746.
6. Xiao, H.; Hoerner, J. K.; Eyles, S. J.; Dobo, A.; Voigtman, E.; Mel'cuk, A. I.; Kaltashov, I. A. Protein Sci 2005, 14, 543–557.
7. Maxwell, K. L.; Wildes, D.; Zarrine-Afsar, A.; Rios, M. A. D. L.; Brown, A. G.; Friel, C. T.; Hedberg, L.; Horng, J.-C.; Bona, D.; Miller, E. J.; Vallée-Bélisle, A.; Main, E. R. G.; Bemporad, F.; Qiu,

L.; Teilum, K.; Vu, N.-D.; Edwards, A. M.; Ruczinski, I.; Poulsen, F. M.; Kragelund, B. B.; Michnick, S. W.; Chiti, F.; Bai, Y.; Hagen, S. J.; Serrano, L.; Oliveberg, M.; Raleigh, D. P.; Wittung-Stafshede, P.; Radford, S. E.; Jackson, S. E.; Sosnick, T. R.; Marqusee, S.; Davidson, A. R.; Plaxco, K. W. Protein Sci 2005, 14, 602–616.
8. Plaxco, K. W.; Simons, K. T.; Baker, D. J Mol Biol 1998, 277, 985–994.
9. Gromiha, M. M.; Selvaraj, S. J Mol Biol 2001, 310, 27–32.
10. Zhou, H.; Zhou, Y. Biophys J 2002, 82, 458–463.
11. Weikl, T. R.; Dill, K. A. J Mol Biol 2003, 332, 953–963
12. Nölting, B.; Schälike, W.; Hampel, P.; Grundig, F.; Gantert, S.; Nicole, S.; Wolfhard, B.; Phoebe, X. Q. J Theor Biol 2003, 223, 299–307.
13. Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. Protein Sci 2003, 12, 2057–2062.
14. Ivankov, D. N.; Finkelstein, A. V. Proc Nat Acad Sci USA 2004, 101, 8942–8944.
15. Gong, H.; Isom, D. G.; Srinivasan, R.; Rose, G. D. J Mol Biol 2003, 327, 1149–1154.
16. Shao, H.; Peng, Y.; Zeng, Z. H. Protein Pept Lett 2003, 10, 277–280.
17. Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. Proteins 2003, 51, 162–166.
18. Punta, M.; Rost, B. J Mol Biol 2005, 348, 507–512.
19. Huang, J. T.; Jing, T. Proteins 2006, 63, 551–554.
20. Gromiha, M. M. J Chem Inf Model 2005, 45, 494–501.
21. Gromiha, M. M.; Thangakani, A. M.; Selvaraj, S. Nucleic Acids Res 2006, 34, 70–74.
22. Ma, B. G.; Guo, J. X.; Zhang, H. Y. Proteins 2006, 65, 362–372.
23. Huang, L. T.; Gromiha, M. M. J Comput Chem 2008, 29, 1675–1683.
24. OuYang, Z.; Liang, J. Protein Sci 2008, 17, 1256–1263.
25. Jiang, Y.; Iglinski, P.; Kurgan, L. J Comput Chem 2009, 30, 772–783.
26. Gao, J.; Zhang, T.; Zhang, H.; Shen, S.; Ruan, J.; Kurgan, L. Proteins 2010, 78, 2114–2130.
27. Shen, H. B.; Song, J. N.; Chou, K. C. J Biomed Sci Eng 2009, 2, 136–143.
28. Chou, K. C.; Shen, H. B. Open Bioinformatics J 2009, 3, 31–50.
29. Chou, K. C. Proteins 2001, 43, 246–255.
30. Chou, K. C.; Elrod, D. W. Protein Eng 1999, 12, 107–118.
31. Chou, K. C.; Elrod, D. W. Proteins 1999, 34, 137–153.
32. Debe, D. A.; William A. Goddard, I. J Mol Biol 1999, 294, 619–625.
33. Zhang, L. X.; Sun, T. T. Biophys Chem 2005, 113, 9–16.
34. Fulton, K. F.; Devlin, G. L.; Jodun, R. A.; Silvestri, L.; Bottomley, S. P.; Fersht, A. R.; Buckle, A. M. Nucleic Acids Res 2005, 33, 279–283.
35. Capriotti, E.; Casadio, R. Bioinformatics 2007, 23, 385–386.
36. Chou, K. C. Bioinformatics 2005, 21, 10–19.
37. Chou, K. C. Curr Proteomics 2009, 6, 262–274.
38. Lin, H.; Wang, H.; Ding, H.; Chen, Y. L.; Li, Q. Z. Acta Biotheor 2009, 57, 321–330.
39. Zhang, S. W.; Zhang, Y. L.; Yang, H. F.; Zhao, C. H.; Pan, Q. Amino Acids 2008, 34, 565–572.
40. Jiang, X. Y.; Wei, R.; Zhang, T. L.; Gu, Q. Protein Pept Lett 2008, 15, 392–396.
41. Li, F. M.; Li, Q. Z. Protein Pept Lett 2008, 15, 612–616.
42. Chou, K. C.; Shen, H. B. PLoS One 2010, 5, e9931.
43. Jiang, X.; Wei, R.; Zhao, Y.; Zhang, T. Amino Acids 2008, 34, 669–675.
44. Ding, Y. S.; Zhang, T. L. Pattern Recognit Lett 2008, 29, 1887–1892.
45. Li, Z. C.; Zhou, X. B.; Dai, Z.; Zou, X. Y. Amino Acids 2009, 37, 415–425.
46. Nanni, L.; Lumini, A. Amino Acids 2008, 34, 653–660.
47. Zhang, S. W.; Chen, W.; Yang, F.; Pan, Q. Amino Acids 2008, 35, 591–598.
48. Qiu, J. D.; Huang, J. H.; Liang, R. P.; Lu, X. Q. Anal Biochem 2009, 390, 68–73.
49. Gu, Q.; Ding, Y. S.; Zhang, T. L. Protein Pept Lett 2010, 17, 559–567.
50. Esmaeili, M.; Mohabatkar, H.; Mohsenzadeh, S. J Theor Biol 2010, 263, 203–209.
51. Lin, H. J Theor Biol 2008, 252, 350–356.
52. Chen, C.; Chen, L. X.; Zou, X. Y.; Cai, P. X. Protein Pept Lett 2009, 16, 27–31.
53. Lin, H.; Ding, H.; Guo, F. B.; Zhang, A. Y.; Huang, J. Protein Pept Lett 2008, 15, 739–744.
54. Zhang, G. Y.; Li, H. C.; Gao, J. Q.; Fang, B. S. Protein Pept Lett 2008, 15, 1132–1137.
55. Fang, Y. P.; Guo, Y. Z.; Feng, X. Y.; Li, M. L. Amino Acids 2008, 34, 103–109.
56. Qiu, J. D.; Huang, J. H.; Shi, S. P.; Liang, R. P. Protein Pept Lett 2010, 17, 715–722.
57. Zhou, X. B.; Chen, C.; Lia, Z. C.; Zou, X. Y. J Theor Biol 2007, 248, 546–551.
58. Ding, H.; Luo, L. F.; Lin, H. Protein Pept Lett 2009, 16, 351–355.
59. Zhang, G. Y.; Fang, B. S. J Theor Biol 2008, 253, 310–315.
60. Zeng, Y. h.; Guo, Y. z.; Xiao, R. q.; Yang, L.; Yu, L. z.; Li, M. l. J Theor Biol 2009, 259, 366–372.
61. Thirumalai, D. J Phys I 1995, 5, 1457–1467.
62. Gutin, A. M.; Abkevich, V. I.; Shakhnovich, E. I. Phys Rev Lett 1996, 77, 5433–5436.
63. Finkelstein, A. V.; Badretdinov, A. Y. Fold Des 1997, 2, 115–121.
64. Tanford, C. J Am Chem Soc 1962, 84, 4240–4274.
65. Kauzmann, W. Adv Protein Chem 1959, 14, 1–63.
66. Landwehr, N.; Hall, M. A.; Frank, E. Mach Learn 2005, 59, 161–205.
67. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275–349.
68. Chou, K. C.; Shen, H. B. Nat Protocol 2008, 3, 153–162.
69. Chou, K. C.; Shen, H. B. Anal Biochem 2007, 370, 1–16.
70. Zhou, G. P. J Protein Chem 1998, 17, 729–738.
71. Zhou, G. P.; Doctor, K. Proteins 2003, 50, 44–48.
72. Kumarevel, T.; Gromiha, M. M.; Selvaraj, S.; Gayatri, K.; Kumar, P. Biophys Chem 2002, 99, 189–198.
73. Jiang, Z. T.; Zhang, L. X.; Chen, J.; Xia, A. G.; Zhao, D. L. Polymer 2002, 43, 6037–6047.
74. Gromiha, M. M.; Selvaraj, S. J Biol Chem 1997, 23, 151–162.
75. Gromiha, M. M.; Selvaraj, S. Prog Biophys Mol Biol 2004, 86, 235–277.
76. Viguera, A. R.; Vega, C.; Serrano, L. Proc Nat Acad Sci USA 2002, 99, 5349–5354.
77. Lee, S.; Lee, B. C.; Kim, D. Proteins 2006, 62, 1107–1114.
78. Cuff, J. A.; Barton, G. J. Proteins 2000, 40, 502–511.
79. Geourjon, C.; Deléage, G. Comput Appl Biosci 1995, 11, 681–684.
80. Chou, K. C.; Shen, H. B. Nat Sci 2009, 1, 63–92.