Amino Acids (2012) 43:567–572
DOI 10.1007/s00726-011-1189-3

ORIGINAL ARTICLE

Relationship between protein folding kinetics and amino acid properties

Jitao T. Huang · Dajie J. Xing · Wei Huang

Received: 19 June 2011 / Accepted: 29 November 2011 / Published online: 13 December 2011

Abstract The successful prediction of protein-folding rates based on the sequence-predicted secondary structure suggests that the folding rates might be predicted from sequence alone. To pursue this question, we directly predict the folding rates from amino acid sequences, which do not require any information on secondary or tertiary structure. Our work achieves 88% correlation with folding rates determined experimentally for proteins of all folding types and peptide, suggesting that almost all of the information needed to specify a protein's folding kinetics and mechanism is comprised within its amino acid sequence. The influence of residue on folding rate is related to amino acid properties. Hydrophobic character of amino acids may be an important determinant of folding kinetics, whereas other properties, size, flexibility, polarity and isoelectric point, of amino acids have contributed little to the folding rate constant.

Keywords Protein folding kinetics · Folding rate constants · Amino acid composition · Hydrophobic character · Statistical analysis

Introduction

Protein structure (topological pattern, contacts, secondary structures, etc.) is an important determinant of protein-folding mechanisms (Dobson 2003; Baker 2000; Plaxco et al. 1998, 2000; Makarov et al. 2002; Gromiha and Selvaraj 2001). The empirical relationship between protein-folding rate and secondary structure content has been reported for two-state proteins (Gong et al. 2003; Huang et al. 2007). Moreover, Ivankov and Finkelstein (2004) presented a general method to predict folding rates of proteins of all folding kinds from the amino acid sequence-predicted secondary structure. Their results imply that folding rates may be predicted from sequence using secondary structure prediction methods.

However, accuracy of secondary structure prediction seems to exist the superior limits (Huang and Wang 2002; Russell and Barton 1993; Levin 1997), although the algorithms just keep on improving and the accuracy's record is broken frequently (Simossis and Heringa 2004; Rost 2001). In this case, why not directly using amino acid sequence to predict folding rates? Over the past few years, the folding rates of proteins were predicted using amino acid sequences (Gromiha 2005; Huang and Tian 2006; Ma et al. 2006; Zheng and Jie 2008; Chang et al. 2010; Xi et al. 2010; Lin et al. 2010; Gao et al. 2010; Li and Li 2011; Guo and Rao 2011). It has been shown, however, that the sequence by itself determines folding rates of only two-state folding proteins and fails to predict those for multi-state proteins. Also, it is not clear whether the physico-chemical properties of amino acids are a cause of the correlation between folding kinetics and sequence.

Here, we estimate folding rates of proteins of all folding kinds (two- and multi-state kinetics) from their amino acid sequences alone, which do not require any result on secondary structure prediction. It achieves 88% correlation with experiment over all 67 proteins (including 41 two-state and 26 multi-state proteins) collected up to now. The comparison of contribution of each amino acid on folding

Electronic supplementary material The online version of this article (doi:10.1007/s00726-011-1189-3) contains supplementary material, which is available to authorized users.

J. T. Huang (✉) · D. J. Xing · W. Huang
State Key Laboratory of Elemento-Organic Chemistry,
College of Chemistry, Nankai University, Tianjin 300071, China
e-mail: jthuang@nankai.edu.cn

rates and properties of the amino acid shows that the hydrophobic interaction plays the dominant role in determining the protein folding kinetics, whereas the other chemical natures of the amino acids play a secondary role.

Results and discussion

Simple proteins often fold according to the first-order rate equation:

$$-\frac{\mathrm{d}[U]}{\mathrm{d}t} = k[U] \tag{1}$$

where $[U]$ is the concentration of unfolded proteins at time t, and k is a folding rate constant in aqueous solution. Using multiple linear regressions, the correlation between the number of residues in each amino acid type and folding rate constants determined experimentally are examined for 67 two- and multi-state proteins. The best-fit linear relationship between ln k and the number of amino acid residues is:

$$\begin{aligned}\ln k = {} & -0.12A - 0.58C + 0.03D - 0.13E - 0.36F \\ & + 0.03G + 0.25H - 0.3I + 0.3K - 0.37L - 0.02M \\ & + 0.51N + 0.71P + 0.12Q + 0.18R + 0.13S \\ & - 0.17T - 0.3V - 0.55W - 0.69Y + 9.33\end{aligned} \tag{2}$$

where A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y is the number of alanine, cysteine, aspartate, glutamate, phenylalanine, glycine, histidine, isoleucine, lysine, leucine, methionine, asparagines, proline, glutamine, arginine, serine, threonine, valine, tryptophan and tyrosine of a protein sequence, respectively. Figure 1 plots these results for the two- and multi-state folding proteins. The line represents linear fit with correlation coefficient, r, of 0.88; $p < 0.0001$, indicating that amino acid composition is an important determinant of protein folding rates. In Fig. 1 two slow folding proteins, tryptophan synthase β chain (1QOP_B) and C-terminal domain of 3-phosphoglycerate kinase (pgk) (1PHP), have very low estimated values, i.e., conform to Eq. 2. If the two proteins are excluded, the correlation coefficient shows a modest decrease ($r = 0.84$; $p < 0.0001$). In addition, relatively weaker correlation is also observed between ln k and amino acid compositions (the occurrence frequency of each amino acid) ($r = 0.78$; $p < 0.0001$).

Using Eq. 2, the strong correlations are also observed between ln k and the number of amino acids for two-state proteins ($r = 0.78$; $p < 0.0001$) and multi-state proteins ($r = 0.86$; $p < 0.0001$), respectively. While the relationship between ln k and length of overall protein chain is weak ($r = -0.55$; $p < 0.0001$).

We performed a jackknife test, in which the error was calculated by refitting the data 67 times, omitting a protein in each cycle. The correlation coefficient is 0.89 (±0.003); $p < 0.0001$. Results from the jackknife analysis can be used to estimate the mean and standard deviation for each parameter in Eq. 2, as shown in Eq. 3:

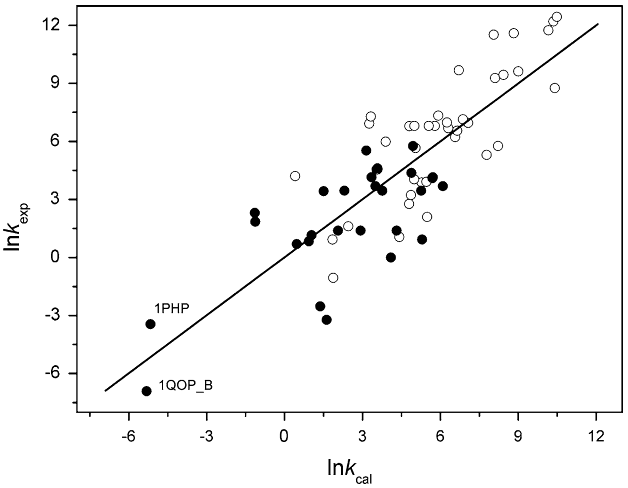


Fig. 1 Calculated folding-rate constant versus experimentally determined folding-rate constant for the set of 67 proteins (41 two-state proteins and 26 multi-state proteins). The calculated rate constant is based on the number of amino acids with weights from Eq. 2. The regression line is given by the equation: ln k_{exp} = ln k_{cal} with correlation coefficient 0.88, where ln k_{cal} and ln k_{exp} are the logarithm of folding rates calculated theoretically and measured experimentally, respectively. *Open circle* two-state proteins, *closed circle* multi-state proteins

$$\begin{aligned}\ln k = {} & -0.119(\pm0.009)A - 0.525(\pm0.04)C \\ & + 0.028(\pm0.012)D - 0.132(\pm0.001)E \\ & - 0.356(\pm0.019)F + 0.028(\pm0.01)G \\ & + 0.252(\pm0.015)H - 0.297(\pm0.013)I \\ & + 0.299(\pm0.007)K - 0.364(\pm0.009)L \\ & - 0.017(\pm0.02)M + 0.509(\pm0.014)N \\ & + 0.707(\pm0.018)P + 0.124(\pm0.013)Q \\ & + 0.181(\pm0.008)R + 0.124(\pm0.001)S \\ & - 0.171(\pm0.01)T - 0.304(\pm0.011)V \\ & - 0.536(\pm0.02)W - 0.688(\pm0.021)Y \\ & + 9.325(\pm0.079)\end{aligned} \tag{3}$$

Thus, we observe a statistically significant correlation between folding rates and amino acid number for all kinds of proteins. Using Eq. 2, the protein-folding rates can be predicted from their primary sequences. The coefficients (weights) of the regression equation give a measure of the sensitivity of the folding reaction to the residue. As shown in Fig. 2, the order of the amino acids by descending their weight is P, N, K, H, R, S, Q, D, G, M, A, E, T, I, V, F, L, W, C, Y. Based on the values of weight, the amino acids are divided into two groups of P, N, K, H, R, S, Q, D, G being defined as folding-accelerated amino acids, and Y, C, W, L, F, V, I, T, E, A, M being defined as folding-inhibited amino acids. The amino acids with amide groups (N and Q),

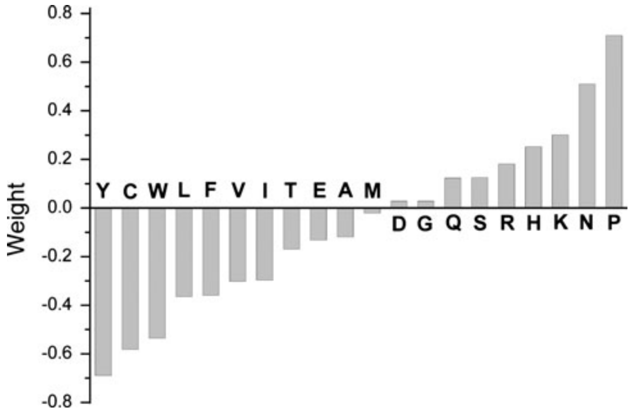


Fig. 2 The contribution of amino acid residues to protein-folding rate. The data are derived from the weights of Eq. 2.

cationic groups (*K* and *R*) and amphoteric group (*H*) facilitate protein folding and maybe even catalysis. Whereas the amino acids with aromatic group (*Y*, *W* and *P*) and hydrophobic sidechain (*L*, *V*, *I* and *A*) are an inhibitor of folding reaction. Proline (*P*) and cysteine (*C*) have large absolute values of the weight. The most common interpretation is that prolines are usually found on the protein surface, where the polypeptide chain must change direction (Richardson 1981), thereby decreasing the number of possible conformations. In addition, there are only three proteins with disulphide bridge in our dataset, chicken egg white lysozyme (PDB id: 1HEL), cell adhesion molecule CD2 (PDB id: 1HNG) and α-amylase inhibitor tendamistat (PDB id: 2AIT). The influence of the number of disulphide bridges (Creighton 1992, 1997) on the folding of protein remains unclear.

There is significant relationship between folding kinetics of a protein and hydrophobic character of the residues. The weights of Eq. 3 is correlated with average surrounding hydrophobicities of amino acids ($r = -0.7$; $p < 0.0001$; Fig. 3a). The average surrounding hydrophobicity is the sum of the hydrophobic indices assigned to the various residues that appear within an 8 Å radius volume in the protein crystal (Manavalan and Ponnuswamy 1978). As the number of hydrophobic residues is increased the rate of folding is slower. This suggests that interaction of water molecules to the amino acids leads to speeding up of folding process of protein.

A significant correlation is also observed between the weight and hydrophobicity of amino acids in folded form ($r = -0.68$; $p < 0.0001$; Fig. 3b). This scale as another hydrophobic index provides information with regard to hydrophobic domains, loop sites and nucleation sites in protein molecules (Ponnuswamy et al. 1980). The calculated result consistently explains the correlation between weight and average surrounding hydrophobicity, supporting the view that the hydrophobic properties of amino acids are factors to determine folding rate of the protein.

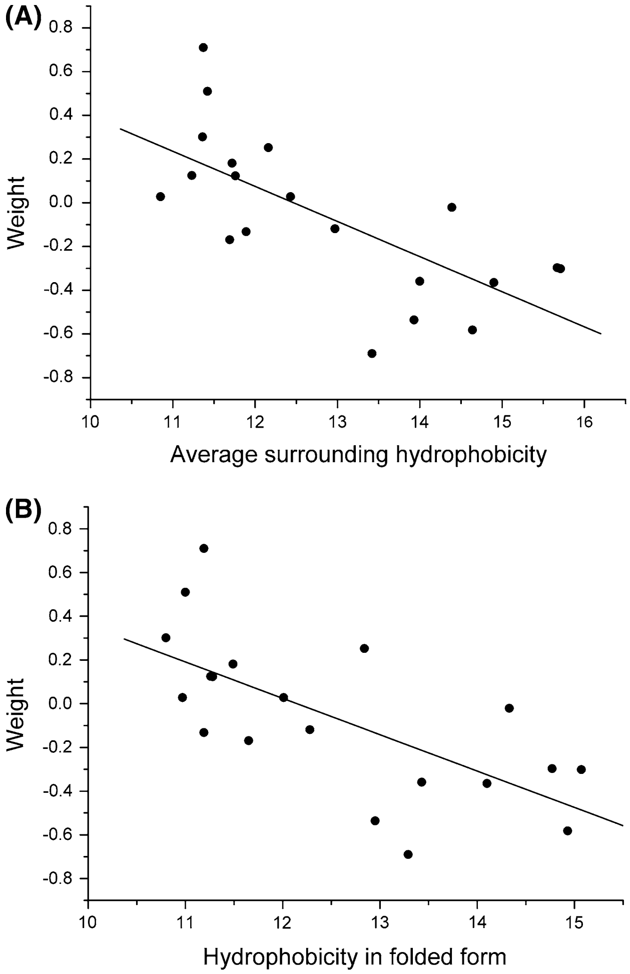
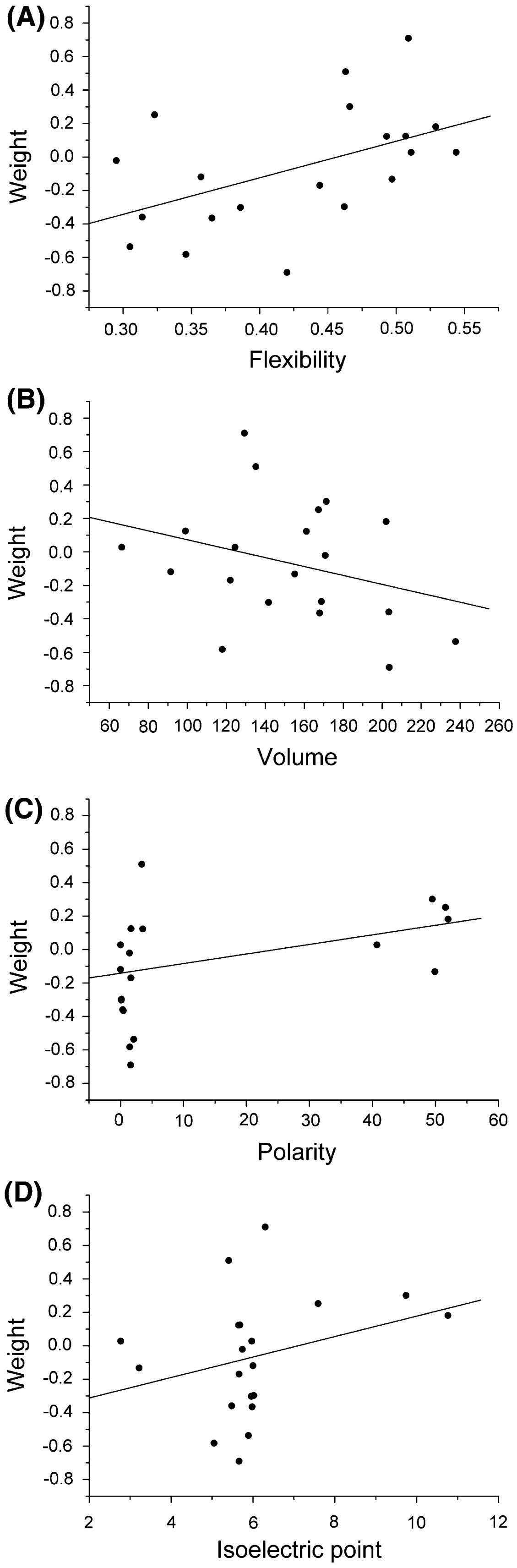


Fig. 3 The relationship between the weight of Eq. 3 and average surrounding hydrophobicities of amino acids. The regression line is given by the equation: $W = -0.16(\pm 0.04) \times \mathrm{ASH} + 2(\pm 0.5)$ with correlation coefficient −0.7; $p < 0.0001$, where *W* is the weight and ASH is the average surrounding hydrophobicity. The relationship between the weight of Eq. 3 and hydrophobicity of amino acids in folded form. The regression line is given by the equation: $W = -0.17(\pm 0.04) \times \mathrm{HFF} + 2.02(\pm 0.54)$ with correlation coefficient −0.67; $p < 0.0001$, where *W* is the weight and HFF is the hydrophobicity in folded form

Correlations between the weights and size or mechanical properties of amino acids are weak or non-existent. A weak correlation ($r = 0.5$; $p = 0.03$; Fig. 4a) hints that amino acid flexibility may be related to the weights. While no correlation ($r = -0.31$; $p = 0.18$; Fig. 4b) is observed between the weights and amino acid volume. The result indicates that flexibility of amino acid has relatively little influence on the folding rates. The flexibility of amino acid is the average fluctuational amplitudes of residues in the protein secondary structures (Bhaskaran and Ponnuswamy



◀ **Fig. 4** The relationship between the weight of Eq. 3 and flexibility, volume, polarity or isoelectric point of amino acids. The correlation coefficients are 0.50 ($p = 0.03$), −0.31 ($p = 0.18$), 0.33 ($p = 0.15$) and 0.3 ($p = 0.2$), respectively. These properties are not related to protein folding kinetics

1988). Size (volume) of amino acid itself (Chothia 1984) has no influence on the folding kinetics.

Also, correlation between the weights and electrostatic aspects (e.g. polarity and isoelectric point) of amino acids is non-existent. No correlation ($r = 0.33$; $p = 0.15$; Fig. 4c) is observed between the weights and amino acid polarity (Zimmerman et al. 1968). Also not apparent in our dataset is any significant correlation ($r = 0.3$; $p = 0.2$; Fig. 4d) between the weights and isoelectric points of amino acids (Zimmerman et al. 1968). However, as mentioned previously, the amino acids with positively charged group may facilitate protein folding.

Several remarks in conclusion:

1. Although additions to or loss of the residues with large absolute value of the weight in a protein sequence could make changes to their folding rates, we do not consider that the prediction works well for sequence-designed peptides and engineering proteins. These proteins frequently encounter misfolding and aggregation during the folding process.
2. The result is not really useful to estimate the effect of point- or multi-mutation on the folding rate since the statistical analysis is based on the natively packed proteins only.
3. The same amino acid might play different roles depending on structural context. For example, the influence of three proline residues (Pro39, Pro85 and Pro105) to folding rate is different in protein apo-CRABP (Eyles and Gierasch 2000). If one uses amino acid composition to represent the sequence of a protein, thus all the sequence-order effects are missed. The prediction method might be inevitably approximate, because it does not take into account those residue-order permutations that do not change the amino acid composition.

Materials and methods

Data

The experimental data for the ln k were from the reports by Ivankov and Finkelstein (2004) and Gromiha et al. (2006). The data for the more recently characterized proteins were from our earlier datasets (Huang et al. 2007; Huang and

Cheng 2007, 2008). Of them, 1MJC, 1SHF, 1CIS, 1C8C, 1PSE, 1HDN, 1HZ6, 1ARQ, 1BNI and 1SHG were omitted from our dataset, because they were homologous to 3MEF, 1NYF, 1COA, 1BNZ, 1PSF, 1POH, 2PTL, 1ARR, 1BRS and 1AEY, respectively. The ln k values of all 67 non-homologous proteins were listed in Table S1 (see Supplementary Data). It includes 41 two-state folding proteins and 26 multi-state proteins. It is noteworthy, however, that two polypeptide segments (2 and 19 entries in Table S1) are C-terminal domain (chain length is 200) and N-terminal domain (chain length is 154) in the same protein 3-phosphoglycerate kinase (pgk) (PDB code is 1PHP), respectively. Two different polypeptide chains (7 and 36 entries in Table S1) which have the same PDB code (1AON) are GroEL apical, domain: residues 191–345 (multi-state kinetics) and SH3-domain (fyn), domain 1: residues 136–191 (two-state kinetics), respectively. Subsequently, a set of representative protein sequences was taken from the PDB data bank (http://www.rcsb.org/pdb). The number of residues in each amino acid type was counted.

Analysis

Pearson correlation coefficient, r, is a measurement of the strength and direction of a linear relationship between two groups of base contents. It takes values ranging from -1 (perfect inverse correlation) and 1 (perfect positive correlation). Strong correlation is determined for an absolute r value >0.7; weak correlation >0.5, and statistical significance is determined for a p value <0.05 for all tests. The bi-variate correlation analysis was carried out by the R statistical package (version 2.13.0; http://www.r-project.org/) (Gonzalez et al. 2007). The linear regression analysis was performed by the online statistics and forecasting software (version 1.1.23-r6; http://www.wessa.net/slr.wasp).

Acknowledgments The authors wish to thank Prof. Jinpeng Cheng of Nankai University and Prof. Chunting Zhang of Tianjin University for his continuous support and encouragement. We are indebted to our colleagues Prof. Ao Yu and Prof. Jiaqi He for very useful discussions. This work was supported by the National Natural Science Foundations of China (NSFC-20721062 and NSFC-20972078).

References

Baker DA (2000) The surprising simplicity to protein folding. Nature 405:39–42

Bhaskaran R, Ponnuswamy PK (1988) Positional flexibilities of amino acid residues in globular proteins. Int J Pept Protein Res 32:242–255

Chang L, Wang J, Wang W (2010) Composition-based effective chain length for prediction of protein folding rates. Phys Rev E 82:051930

Chothia C (1984) Principles that determine the structure of proteins. Ann Rev Biochem 53:537–572

Creighton TE (1992) Protein folding pathways determined using disulphide bonds. Bioessays 14:195–199

Creighton TE (1997) Protein folding coupled to disulphide bond formation. Biol Chem 378:731–744

Dobson CM (2003) Protein folding and misfolding. Nature 426:884–890

Eyles SJ, Gierasch LM (2000) Multiple roles of prolyl residues in structure and folding. J Mol Biol 301:737–747

Gao J, Zhang T, Zhang H, Shen S, Ruan J, Kurgan L (2010) Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. Proteins 78:2114–2130

Gong H, Isom DG, Srinivasan R, Rose GD (2003) Local secondary structure content predicts folding rates for simple, two-state folding proteins. J Mol Biol 327:1149–1154

Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V (2007) SNPassoc: an R package to perform whole genome association studies. Bioinformatics 23:654–655

Gromiha MM (2005) A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model 45:494–501

Gromiha MM, Selvaraj S (2001) Comparison between long-range interactions and contact order in determining the folding rate of two-state folding proteins: application of long-range order to folding rate prediction. J Mol Biol 310:27–32

Gromiha MM, Thangakani AM, Selvara S (2006) FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucl Acid Res 34:W70–W74

Guo JX, Rao NN (2011) Predicting protein folding rate from amino acid sequence. J Bioinf Comput Biol 9:1–13

Huang JT, Cheng JP (2007) Prediction of folding transition-state position (βT) of small, two-state proteins from local secondary structure content. Proteins 68:218–222

Huang JT, Cheng JP (2008) Differentiation between two-state and multi-state folding proteins based on sequence. Proteins 72:44–49

Huang JT, Tian J (2006) Amino acid sequence predicts folding rate for middle-size two-state proteins. Proteins 63:551–554

Huang JT, Wang MT (2002) Secondary structural wobble: the limits of protein prediction accuracy. Biochem Biophys Res Commun 294:621–625

Huang JT, Cheng JP, Chen H (2007) Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins 67:12–17

Ivankov DN, Finkelstein AV (2004) Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA 101:8942–8944

Levin JM (1997) Exploring the limits of nearest-neighbour secondary structure prediction. Protein Eng 10:771–776

Li RF, Li H (2011) The influence of protein coding sequences on protein folding rates of all-β proteins. Gen Physiol Biophys 30:154–161

Lin GN, Wang Z, Xu D, Cheng J (2010) SeqRate: sequence-based protein folding type classification and rates prediction. BMC Bioinf 11(Suppl 3):S1

Ma BG, Guo JX, Zhang HY (2006) Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. Proteins 65:362–372

Makarov DE, Keller CA, Plaxco KW, Metiu H (2002) How the folding rate constant of simple, single-domain proteins depends on the number of native contacts. Proc Natl Acad Sci USA 99:3535–3539

Manavalan P, Ponnuswamy PK (1978) Hydrophobic character of amino acid residues in globular proteins. Nature 275:673–674

Plaxco KW, Simons KT, Baker D (1998) Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 277:985–994

Plaxco KW, Simons KT, Ruczinski I, Baker D (2000) Topology, stability, sequence, and length: defining the determinants of two-state folding protein folding kinetics. Biochemistry 39:11177–11183

Ponnuswamy PK, Prabhakarana M, Manavalan P (1980) Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. Biochim Biophys Acta 623:301–316

Richardson JS (1981) Anatomy and taxonomy of protein structures. Adv Protein Chem 34:167–339

Rost B (2001) Review: protein secondary structure prediction continues to rise. J Struct Biol 134:204–218

Russell RB, Barton GJ (1993) The limits of protein secondary structure prediction accuracy from multiple sequence alignment. J Mol Biol 234:951–957

Simossis VA, Heringa J (2004) Integrating protein secondary structure prediction and multiple sequence alignment. Curr Protein Pept Sci 5:249–266

Xi LL, Li SY, Liu HX, Li JZ, Lei BL, Yao XJ (2010) Global and local prediction of protein folding rates based on sequence autocorrelation information. J Theor Biol 264:1159–1168

Zheng OY, Jie L (2008) Prediction of protein folding rates from geometric contact and amino acid sequences. Protein Sci 17:1256–1263

Zimmerman JM, Eliezer N, Simha R (1968) The characterization of amino acid sequences in proteins by statistical methods. J Theor Biol 21:170–201