PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Reduced alphabet for protein folding prediction

Jitao T. Huang,* Titi Wang, Shanran R. Huang, and Xin Li

Department of Chemistry and National Laboratory of Elemento-Organic Chemistry, Nankai University, Tianjin, 300071, People's Republic of China

ABSTRACT

What are the key building blocks that would have been needed to construct complex protein folds? This is an important issue for understanding protein folding mechanism and guiding de novo protein design. Twenty naturally occurring amino acids and eight secondary structures consist of a 28-letter alphabet to determine folding kinetics and mechanism. Here we predict folding kinetic rates of proteins from many reduced alphabets. We find that a reduced alphabet of 10 letters achieves good correlation with folding rates, close to the one achieved by full 28-letter alphabet. Many other reduced alphabets are not significantly correlated to folding rates. The finding suggests that not all amino acids and secondary structures are equally important for protein folding. The foldable sequence of a protein could be designed using at least 10 folding units, which can either promote or inhibit protein folding. Reducing alphabet cardinality without losing key folding kinetic information opens the door to potentially faster machine learning and data mining applications in protein structure prediction, sequence alignment and protein design.

Proteins 2015; 83:631–639.

Key words: protein folding; reduced alphabet; prediction; folding unit.

INTRODUCTION

Protein folding is a process that going down the funnel-like free energy landscape through multiple parallel pathways toward the bottom of the funnel.[1–6] The details of the landscape surface of a folding funnel are unique for a specific protein structure, possibly depending only on simple basic parts.

Amino acids are the basic molecular building blocks or the chemical components of proteins. Hydrogen bonds can link several amino-acid residues together to form an integral α-helix, β-sheet or other secondary structure elements. They are also viewed as modularized building blocks. These protein components are not equally important in folding process, and the question which of them should be accounted as essential elements capable of constructing protein 3D structures?

From the work of several groups studying protein design and folding, it is heavily suggested that protein folding can be achieved with significantly fewer components than 20 commonly found amino acids. Reducing (simplifying) the 20 amino acids into a smaller number of representative amino acids have been proved to be a successful strategy in machine learning, data mining, and numerical optimization. The resulting reduced amino acid alphabets have been applied to protein design,[7–10] sequence alignment,[11–14] protein classification,[15,16] protein evolution,[17,18] and protein structure prediction.[19–22] For protein folding problem, Baker *et al.*[20] used a reduced alphabet containing five amino acids (Ala, Glu, Gly, Ile, and Lys) to design a Src SH3 protein of 57 residues. Wang and Wang[21] found the same alphabet based on assessment of mismatch between a reduced interaction matrix and the Miyazawa or Jernigan matrix. The lattice chains obtained provide a better understanding on the simplification of natural proteins.[22] The effective length based on specific amino acid types is well correlated with folding rates.[23] Levy and coworkers[24] proposed that reduced alphabets of 10–12 amino acids can be used to design folding sequences for a large

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

Grant sponsor: Natural Science Foundations of China; Grant number: NSFC-20972078.

*Correspondence to: Jitao T. Huang, Department of Chemistry and National Laboratory of Elemento-Organic Chemistry, Nankai University, Tianjin 300071, People's Republic of China. E-mail: jthuang@nankai.edu.cn

Received 15 September 2014; Revised 7 November 2014; Accepted 21 December 2014

Published online 16 January 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24762

number of proteins. Akanuma *et al.*[7] observed that 88 % of the residues in a phosphoribosyltransferase can be replaced by nine amino acids (Ala, Glu, Gly, Leu, Pro, Arg, Thr, Val, and Tyr) without losing key biochemical information. Shortly afterward, Fan and Wang[12] used completely different methods to draw a consistent conclusion that the minimum number of amino acids required to encode a protein fold is ~10. For these reduced alphabets, a pertinent question is whether it is possible to accurately predict protein folding rates from a small number of amino acids.

Knowledge of the protein structure is helpful in interpreting protein folding kinetic data. Analyzing the folding kinetics can reveal the underlying folding mechanism of the protein. For instance, prediction of folding rate can determine the role of each amino acids and secondary structures in forming proper protein structure. Plaxco *et al.* and Baker *et al.*[25–27] first found a correlation between the topological structures and the folding rates of the small proteins. Subsequently, efforts have further been made to predict protein folding rates from their secondary structure assignment[28,29] or amino acid sequence[30–38] to mine the necessary information to make up the correct 3D structure. In our previous study, we also observed that folding rates, transition state position and folding kinetic type of proteins are well correlated to their secondary structure content[39,40] or amino acid composition.[41–44]

These prompt the question of whether there is a minimal alphabet with could be used to predict protein folding rates. The folding rates of the simplified proteins are compared to that of the corresponding wild type proteins. If the simplified proteins refold at almost the same rate as wild type, the letters of the minimal alphabet can all be viewed as potential structural units for protein folding and the design of artificial proteins. Furthermore, we try to merge 20 amino acids with secondary structure features to create a superset of the descriptors. This superset provides a more detailed description for the protein folding kinetics, which might offer a platform that facilitates the simplification of protein folding model.

MATERIALS AND METHODS

Folding rates

The folding rates determined experimentally of 94 proteins of all folding kinetic types were taken from the reports by Jackson,[45] Ivankov and Finkelstein,[29] and Gromiha et al.[31] The data for the more recently characterized proteins were from the reports by Lu and coworkers[46] and Protein Kinetic Database (http://kineticdb.protres.ru/).[47] These proteins are non-homologous, which their sequence identities are <95 % BLAST identity cut-off (BLAST alignment http://blast.ncbi.nlm.nih.gov/). The folding rates of these proteins were listed in Supporting Information Table SI and SII.

Amino acid sequences

Protein (domain) sequences were taken from the Protein Data Bank (PDB) (http://www.rcsb.org/pdb).[48] The number of residues in each amino acid type was counted, as listed in Supporting Information Table SI.

Secondary structure assignment

Secondary structures were assigned as described by Kabsch and Sander.[49] Secondary structure sequences were from the DSSP dataset (http://swift.cmbi.ru.nl/gv/dssp/) or PDB databank (http://www.rcsb.org/pdb).[48] The number of residues in each secondary structure type was counted, as listed in Supporting Information Table SII.

Manual stepwise regression

Linear regression was performed between 28 predictor variables and folding rates to obtain the partial significance level (partial P values) of Student's t test of each variable. After removing one variable with the largest partial P values, the regression analysis was performed again between residual 27 variables and folding rates to obtain 27 new t test partial P values. After further removing one variable with maximal partial P values obtained recently, the regression analysis was performed again. Such process was repeated until all variables are eliminated. The statistical correlations, R value, F test, P value, F value, and SD value of each regression round were recorded. The multiple linear regression analysis was performed by the Xurus Website Regression Tools (http://www.xuru.org/rt/MLR.asp) or SPSS software (version 16; SPSS).

Manual simultaneous regression

An additional manual method was the use of a Student's t test to simultaneously estimate the partial significance levels of 28 predictor variables, that is, to obtain 28 partial P values. All variables were sequentially removed in order of decreasing partial P values, and the R values, F test, P values, F values, and SD values of residual variables were obtained by linear regression analysis.

Automated stepwise regression

Automated stepwise regression is an algorithm of building a model by successively adding or removing variables based on their partial significance level.[50] In typical implementation of the algorithm, stepwise regression was performed with an entry criterion of t test partial $P < 0.82$ and a removal criterion of partial $P > 0.98$. The 28 predictor variables were automatically entered into

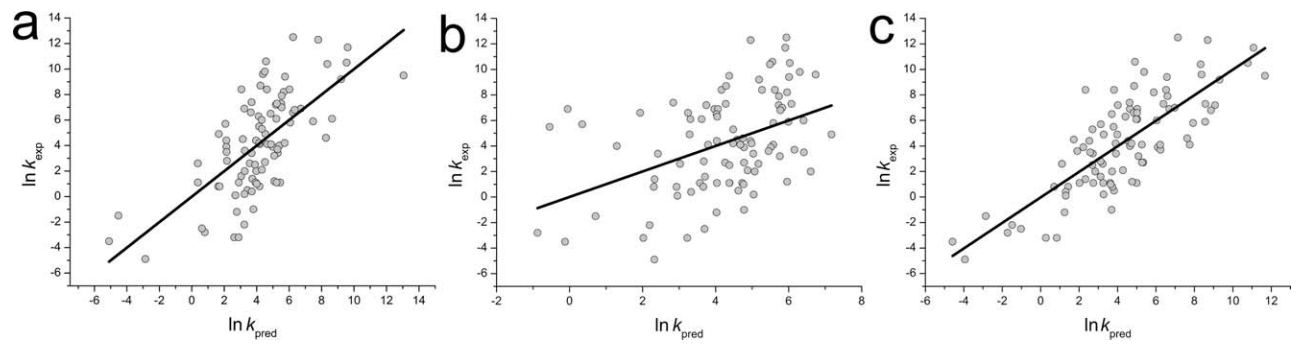

Figure 1

Predicted folding rates versus experimentally determined folding rates of proteins. The predicted folding rates are estimated from different alphabets: (**a**) the 20-letter alphabet (including 20 common amino acids), (**b**) the 8-letter alphabet (including β-bridge ([B]), no secondary structure assigned ([C]), β-sheet ([E]), 3_{10}-helix ([G]), α-helix ([H]), π-helix ([I]), bend ([S]), and turn ([T]) secondary structures), (**c**) the merged 28-letter alphabet (including 20 amino acids and 8 secondary structures). The three corresponding regression lines are given by the relationships: ln k_{exp} = ln k_{pred} with correlation coefficients of 0.667 ($P = 2.29 \times 10^{-4}$), 0.436 ($P = 9.271 \times 10^{-3}$), and 0.787 ($P = 2.4 \times 10^{-6}$), respectively.

the model and the corresponding correlations (R value, F test, P value, F value, and SD value) are estimated. The effective alphabet of predictors was obtained if it achieved a significance of $P < 0.01$ and correlation coefficient of $R > 0.7$. The stepwise regression analyses were performed by SPSS software (version 16; SPSS).

Statistical correlation

In a multiple regression, Pearson's correlation coefficient (R value) is a measurement of the intensity of a linear relationship between multiple-variable predictors and folding rates. A strong correlation was defined as a correlation coefficient of > 0.7; a weak correlation as $0.4 - 0.7$; and no correlation as < 0.4. A two-tailed probability (P value) of < 0.01 for the F value was considered as significant for all statistical analysis in this study. The bi-variate correlation analyses were carried out by the R statistical package (version 2.13.0; http://www.r-project.org/).[51] The multiple linear regression analyses were performed by the Xurus Website Regression Tools (http://www.xuru.org/rt/MLR.asp) or SPSS software (version 16; SPSS).

Jackknife test

A jackknife test was used to carry out sensitivity analysis to the final comparative model. For 94 protein samples, the errors were computed by refitting the data 94 times, deleting a protein in each cycle. Results generated from these leave-one-out processes were used to calculate the mean and standard deviation for each factor of regression line. If the mean correlation coefficient does

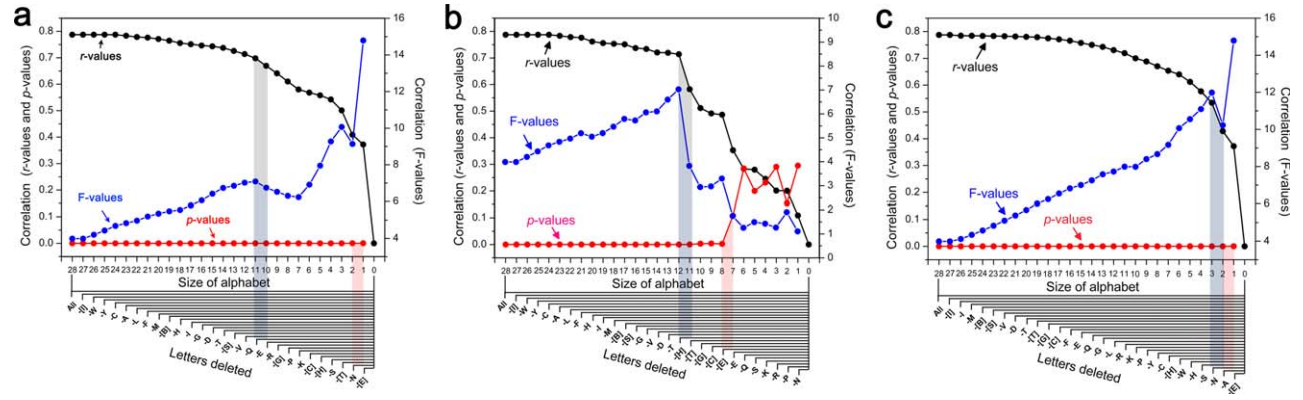

Figure 2

The dependence of correlation parameters on alphabet size. All letters are removed one-by-one from the 28-letter alphabet. The correlations (R values, F test P values and F values) between the residual letters and folding rates are estimated at each round. The R value denotes the Pearson correlation coefficient; the P value denotes the significance level of F test; the F value denotes the ratio between the mean of squares effect and the mean of squares error. (**a**) Manual stepwise regression, (**b**) manual simultaneous regression, and (**c**) automated stepwise regression.[50,52,53]

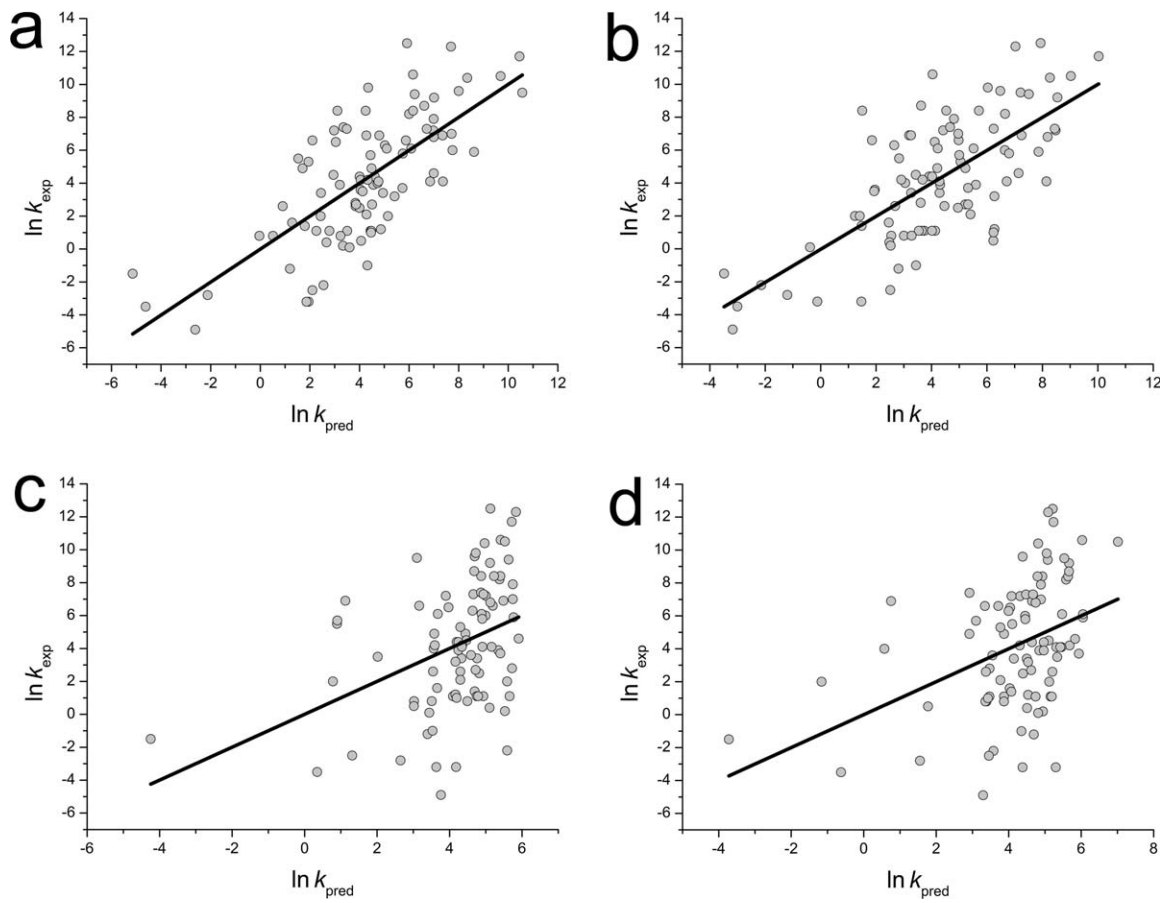

Figure 3

Predicted folding rates versus experimentally determined folding rates of proteins. The predicted folding rates are estimated from different reduced alphabets: (**a**) the reduced 12-letter alphabets, {N P R K S Q E [E] [C] [G] [T] [H]} and {[E] N [T] S [H] [C] K P [G] R E Q}. The two alphabets have the same correlation with folding rates although they are obtained from two different regression methods; (**b**) the reduced 10-letter alphabet, {[E] A N S H W [H] C Y P}, (**c**) the reduced 5-letter alphabet, {A E G I K}, from reports by Baker *et al.*[20] and Wang *et al.*[21] (**d**) the reduced 9-letter alphabet, {A D G L P R T V Y}, by reported of Akanuma *et al.*[7] The four corresponding regression lines are given by the relationships: ln $k_{exp} = \ln k_{pred}$ with correlation coefficients of 0.714 ($P = 1.43 \times 10^{-8}$), 0.70 ($P = 7.1 \times 10^{-9}$), 0.391 ($P = 0.011$), 0.407 ($P = 0.071$), respectively.

not decline dramatically after the test, the regression model would be expected to be reliable.

RESULTS AND DISCUSSION

Protein folding kinetics is considered to be determined by its sequence and structure, because the logarithm (ln k) of folding rate constant can be predicted from its amino acid composition, as well as secondary structure composition. We collect 94 non-homologues proteins for which folding rate constants (k) have been determined experimentally, as detailed in the Supporting Information Table SI. These proteins are used to reexamine the relationship between the folding rates and amino acid compositions or secondary structures, prior to be used in further statistical analyses.

For 94 proteins in the dataset, there is a significant but weak correlation [correlation coefficient $R = 0.677$; F test $P = 2.289 \times 10^{-4}$; $F = 3.09$; SD = 3.182; Fig. 1(a)] between folding rates and the occurrence frequency of 20 common amino acids in the protein (Supporting Information Table SI). A weaker correlation [$R = 0.436$; F test $P = 9.271 \times 10^{-3}$; $F = 2.89$; SD = 3.59; Fig. 1(b)] is also observed between folding rates and the occurrence frequency of eight secondary structures in the protein (Supporting Information Table SII). According to Kabsch and Sander,[49] the eight secondary structures are α-helix, 3_{10}-helix, π-helix, β-strand, β-bridge, reverse turn, bend and straight chain without hydrogen-bonded structure. We combine 20 amino acids with eight secondary structures to construct a 28-letter alphabet. A significant and strong relationship [$R = 0.787$; F test $P = 2.4 \times 10^{-6}$; $F = 3.99$; SD = 2.8; Fig. 1(c)] is observed between folding rates and the 28-letter alphabet. Thus, the introduction of some structural factor to amino acid alphabet may be a better description for protein folding kinetics.

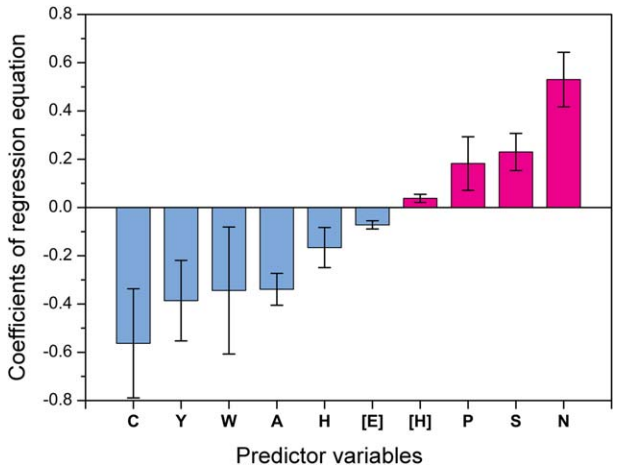

Figure 4

The contribution of reduced primary and secondary structures to protein-folding rate. These data are derived from the coefficients of regression equation [Eq. (2)]. The weight coefficients of Cys, Tyr, Trp, Ala, His, and β-sheet are negative; the weight coefficients of Asn, Ser, Pro, and α-helix are positive.

It is noteworthy, however, that a high correlation coefficient might also be due to the fact that too many alphabet letters are used in prediction. In order to avoid overfitting, the alphabet is reduced by the removal of those letters that have less contribution to folding rates. Primary and secondary structures are not equally important for protein structure. Only some of which are essential for protein folding.

We try to reduce the number of letters to a minimal set or so while still retaining as much predictive power as possible. However, it is quite difficult to enumerate all the possible combinations of 28 letters and get the best one. A simple and efficient method is to evaluate the relative importance of 28 letters for folding rate using statistical Student's *t* test. Then, unimportant letters are removed. Reducing variables has three different methods: (1) manual stepwise regression, (2) manual simultaneous regression, and (3) automatic stepwise regression. The partial *P* value is the smallest significance level for which the observed sample data result in rejection of the null hypothesis, where the letter with a lower partial *P* value is relatively more important.

In the manual stepwise regression, the regression coefficients, *R* values, *F* test, *P* values and *F* values, are repeatedly estimated 28 times, removing one letter with the lowest highest *P* value at each round. The alphabet-size profile of the obtained regression coefficients are shown in Figure 2(a). As can be seen from this figure, when the alphabet size is >12 letters, a satisfying correlation [$R \geq 0.7$; *F* test $P < 0.01$; Fig. 3(a)] with experimental folding rates can be achieved.

In manual simultaneous regression process, we employ a student's *t* test to estimate the partial *P* values of 28 letters. Letters are removed one by one from the alphabet, according to the decreasing order of these partial *P* values. The variation trend of *R* values, *F* test *P* values, and *F* values is shown in Figure 2(b). In this figure, the same correlation is also appeared between folding rates and the 12-letter alphabet [Fig. 3(a)]. This alphabet contains the same 12 letters as above reduced alphabet from manual stepwise regression, but letter order is not completely identical.

The automated stepwise regression[50,52] is a common algorithm of building a model by successively adding one letter at a time if it is statistically significant, or removing one letter that are not statistically significant. These steps are carried out automatically by the SPSS program and the corresponding *R*-, *P*-, and *F*-values are listed in Supporting Information Table SIII. The alphabet size dependence of these statistal parameters is shown in Figure 2(c). As in this figure, even if the size of alphabet is reduced to 10 letters, it still retains the predictive power of model with $R \geq 0.7$ and *F* test $P < 0.01$. This reduced alphabet includes [H] [E] A N C H P S W Y, that is, α-helix, β-sheet, alanine, asparagine, cysteine, histidine, proline, serine, tryptophan, and tyrosine.

To test this minimal alphabet, we use these ten letters to estimate the folding rates measured experimentally for 94 proteins. A least-square fit of the data with the linear equation gives

$$\begin{aligned} \ln k_{\text{pred}} = {} & -0.56(\pm 0.23)\times C - 0.39(\pm 0.17)\times Y - 0.34(\pm 0.26)\times W - 0.34(\pm 0.07)\times A \\ & - 0.17(\pm 0.08)\times H - 0.07(\pm 0.02)\times [E] + 0.04(\pm 0.02)\times [H] + 0.18(\pm 0.11)\times P \\ & + 0.23(\pm 0.08)\times S + 0.53(\pm 0.12)\times N + 6.5(\pm 0.52), \end{aligned} \quad (1)$$

with correlation coefficient *R* of 0.7; *F* test *P* value of 7.1×10^{-9}; *F* value of 7.985; SD value of 2.896 [Fig. 3(b)]. Where *C Y W A H* [*E*] [*H*] *P S N* are the number of residues in cysteine, tyrosine tryptophan, alanine, histidine, β-sheet, α-helix, proline, serine, and asparagine, respectively. These eight amino acids and two secondary structures can be thought to be essential folding units for all proteins.

To examining reliability of this 10-letter alphabet, we use the jackknife (cross-validation) test to estimate the performance of the alphabet in predicting folding rates for 94 proteins, in which the correlation coefficients are

calculated by refitting the data 94 times, omitting one protein sample in each cycle. The obtained regression equation is parametrized by the average of polynomial coefficients for all cycles, that is

$$\begin{aligned}\ln k_{\text{pred}} = {} & -0.563(\pm 0.023)\times C - 0.386(\pm 0.02)\times Y - 0.345(\pm 0.033)\times W - 0.339(\pm 0.007)\times A \\ & -0.165(\pm 0.015)\times H - 0.072(\pm 0.002)\times [E] + 0.038(\pm 0.002)\times [H] + 0.181(\pm 0.016)\times P \\ & +0.229(\pm 0.008)\times S + 0.529(\pm 0.013)\times N + 6.453(\pm 0.056),\end{aligned} \tag{2}$$

with mean correlation coefficient R of 0.7 ($\pm$0.006); F test P value of 1.11×10^{-8} ($\pm 8.82 \times 10^{-9}$); F value of 7.91 ($\pm$0.27); SD value of 2.95 ($\pm$0.5), suggesting that the statistical method described by Eq. (1) is robust.

According to the sign of the polynomial coefficients of regression equation [Eq. (1)], the predictor variables are divided into two groups of C Y W A H [E] being defined as folding-inhibitory factor (negative polynomial coefficients), and N S P [H] being defined as folding-triggered factor (positive polynomial coefficients) (Fig. 4). Natural proteins may fold into the ordered structures by tuning the rate of each folding step.

Such a correlation is not simply a statistical coincidence, but has a physical meaning. The height of the histogram in Figure 4 indicates the relative contribution of ten predictor variable to folding rates. As can be seen from this figure, only regular secondary structures, α helix and β sheet have significant influence on folding rate. Having more α helices can speed up folding rate and whereas more β sheets can slow down folding rate. The amino acids with modest hydrophobicity and polarity, C (2.5), A (1.8), S (−0.8), W (−0.9), Y (−1.3), P (−1.6), H (−3.2), N (−3.4), play an important role in protein folding. Highly hydrophobic amino acids, I (4.5), V (4.2), L (3.8), F (2.8), and hydrophilic (polar) amino acids, K (−3.9), R (−4.5), D (−3.5), E (−3.5), does not contribute significantly to folding rate, where numbers in bracket are Kyte and Doolittle hydrophobic indexes.[54] As far as the reduced alphabet is concerned, roughly speaking, the folding-promoting ability increases progressively with increase in hydrophobic character of the amino acids (Fig. 5), which is consistent with our previous observations.[43,44]

Cysteine is a hydrophobic amino acid having a thiol group; two thiol groups are susceptible to oxidation to form the disulfide bridge that is a rate-limiting step in many folding reactions. Thus, the cysteine residues may significantly slow folding rate. In contrast, asparagine side-chain has a propensity to form hydrogen bonding with the peptide backbone, thus these amino acids help to stabilize the protein structure in the two ends of secondary structure segments and may facilitate protein folding. Serine is an amino acid with a primary hydroxyl group; it can act as a hydrogen bond donor or acceptor. Proline is a structural disruptor of α helices and β sheets because of its irregular geometry. These residues are often spread on the surface of the protein to promote the formation of turns.

Several basic ideas on reducing amino acid alphabet have been presented to construct stable protein structure. However, we observe little correlation [$R = 0.391$; F test $P = 0.011$; SD = 3.625; Fig. 3(c)] between folding rates and the simplified five-letter alphabet {A E G I K} from reports by Riddle *et al.*[20] and Wang and Wang.[21] Also not apparent in the dataset is any significant correlation [$R = 0.407$; F test $P = 0.071$; SD = 3.684; Fig. 3(d)] between folding rates and the simplified 9-letter alphabet {A D G L P R T V Y} from report by Akanuma *et al.*[7] These suggest that, although these five-letter and nine-letter alphabets are insufficient to accurately predict folding kinetic rates of proteins, they still have good kinetic accessibility in model studies. Our reduced alphabet is potentially associated with the variant five groups, (CMFI) (LVWY) (ATGS) (NQDE) (HPRK), reported by Wang and Wang.[21] The underlines and bullet points are used to highlight the letters of Wang-Wang (W-W) reduced alphabet and our simplified alphabet, respectively. Only letter representatives

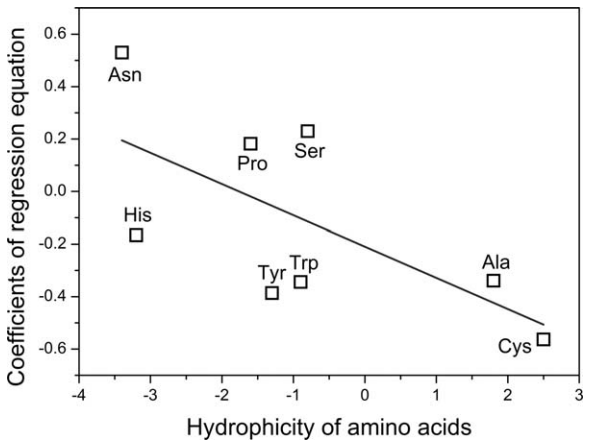

Figure 5

The relationship between amino acid hydrophobicity (H) and foldability (F) described by the coefficients of Eq. (2). The regression line is given by the equation: $F = -0.12(\pm 0.06) \times H + 0.21(\pm 0.12)$ with correlation coefficient −0.66; $P = 0.07$, where H is Kyte and Doolittle hydrophobic index.[54]

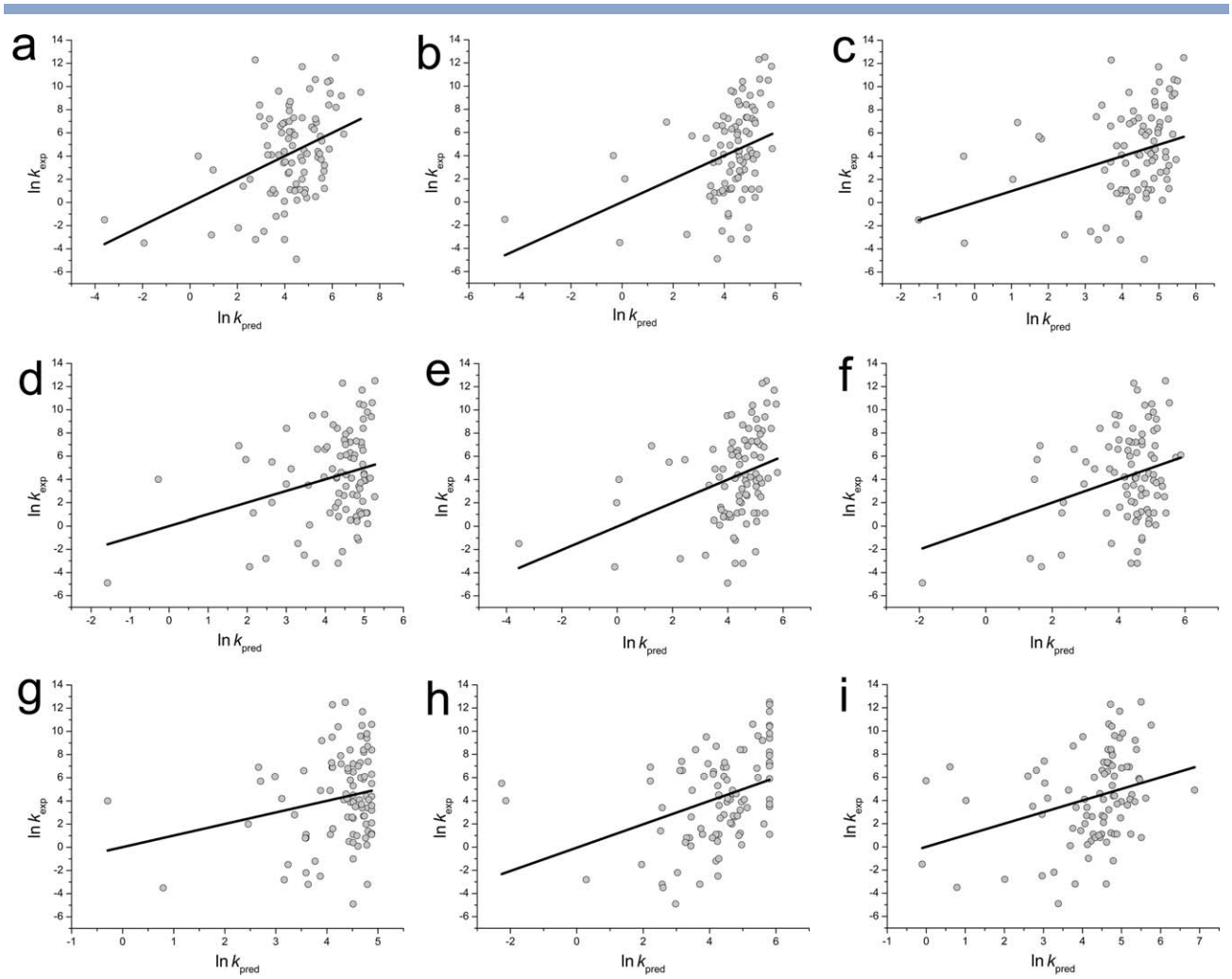

Figure 6

Predicted folding rates versus experimentally determined folding rates of proteins. The predicted folding rates are estimated from different subset of amino acids and secondary structures: (**a**) polar amino acids, {C D E M S T W Y}; (**b**) hydrophobic amino acids, {A F G I L P V}; (**c**) acidic amino acids, {D E}; (**d**) basic amino acids, {R K H}; (**e**) aliphatic amino acids, {A G I L V}; (**f**) aromatic amino acids, {F H W Y}; (**g**) helix secondary structures, {[H] [G] [I]}; (**h**) β secondary structures, {[E] [B]}; and (**i**) loop secondary structures, {[T] [S] [C]}. The nine corresponding regression lines are given by the relationships: $\ln k_{exp} = \ln k_{pred}$ with correlation coefficients of 0.415 ($P = 0.034$), 0.37 ($P = 0.071$), 0.331 ($P = 0.0052$), 0.289 ($P = 0.049$), 0.363 ($P = 0.027$), 0.303 ($P = 0.07$), 0.212 ($P = 0.123$), 0.38 ($P = 8.153 \times 10^{-4}$), and 0.311 ($P = 0.027$), respectively.

of the groups are different from each other, I ↔ C, L ↔ WY, A ↔ A, E ↔ N, and K ↔ HP. Our result qualitatively supports the amino acid grouping scheme of W–W model.[21,22]

Most of the reduced amino acid representations have been derived based on common physicochemical properties shared by different amino acids. The common way to design a reduced amino acid alphabet consists to cluster amino acid into groups according to specific features, for instance, polar/hydrophobic amino acids, acidic/basic amino acids, and aliphatic/aromatic amino acids. Correlations between folding kinetics and specific amino acid groups are examined as below.

A very weak correlation [$R = 0.415$; F test $P = 0.034$; $F = 2.209$; SD = 3.647; Fig. 6(a)] is found between folding rates and polar amino acids {C D E M S T W Y} for the proteins of our dataset. However, no significant relationship [$R = 0.37$; F test $P = 0.071$; $F = 1.954$; SD = 3.701; Fig. 6(b)] is observed between the folding rates at which they fold and hydrophobic amino acids {A F G I L P V}. Polar amino acids have a somewhat larger influence to folding rate than nonpolar amino acids.

No correlation [$R = 0.331$; F test $P = 0.0052$; $F = 5.581$; SD = 3.656; Fig. 6(c)] is observed between folding rates and acidic amino acids {D E}. Also not apparent in the dataset is any significant correlation [$R = 0.289$; F test $P = 0.049$; $F = 2.718$; SD = 3.73; Fig. 6(d)] between folding rates and basic amino acids {H K R}. This suggests that electric charge of molecules does not affect the protein folding.

Folding rates of proteins in our dataset has not shown correlation [$R = 0.363$; F test $P = 0.027$; $F = 2.671$; SD = 3.67; Fig. 6(e)] with aliphatic amino acids {A G I L V}. These folding rates are also not associated [$R = 0.303$; F test $P = 0.07$; $F = 2.247$; SD = 3.733; Fig. 6(f)] with aromatic (and heterocyclic) amino acids {F H W Y}.

We also examine correlations between folding kinetics and specific secondary structure groups. We are observed a significant but not strong relationship [$R = 0.38$; F test $P = 8.153 \times 10^{-4}$; $F = 7.698$; SD = 3.582; Fig. 6(g)] between folding rates and β-structures, {[E] [B]}. As mentioned above, this statistical significance is derived from the correlation between folding rates and [E]. In addition, no significant correlation [$R = 0.212$; F test $P = 0.123$; $F = 2.145$; SD = 3.785; Fig. 6(h)] is apparent between folding rates and protein's helix structures, {[H] [G] [I]}. We are also not aware of any relationship [$R = 0.311$; F test $P = 0.027$; $F = 3.21$; SD = 3.702; Fig. 6(i)] between folding rates and loop structures, {[T] [S] [C]}.

As indicated above, these subsets of amino acids and secondary structures have relatively little influence on folding kinetic rates, suggesting that they are the nonessential folding blocks. These structural units might play important role in various biological functions, for example, enzyme catalysis.

ACKNOWLEDGMENT

The authors thank Prof. Jinpeng Cheng of Nankai University and Prof. Chunting Zhang of Tianjin University for his continuous support and encouragement.

REFERENCES

1. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels pathways and the energy landscape of protein folding: a synthesis. Proteins 1995;21:167–195.
2. Dill KA, Chan HS. From levinthal to pathways to funnels. Nat Struct Biol 1997;4:10–19.
3. Karplus M. Behind the folding funnel diagram. Nat Chem Biol 2011;7:401–404.
4. Koga N, Tatsumi-Koga R, Liu G, Xiao R, Acton TB, Montelione GT, Baker D. Principles for designing ideal protein structures. Nature 2012;491:222–227.
5. Dill KA, MacCallum JL. The protein-folding problem 50 years on. Science 2012;338:1042–1046.
6. Lapidus LJ. Exploring the top of the protein folding funnel by experiment. Curr Opin Struct Biol 2013;23:30–35.
7. Akanuma S, Kigawa T, Yokoyama S. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc Natl Acad Sci USA 2002;99:13549–13553.
8. Walter KU, Vamvaca K, Hilvert D. An active enzyme constructed from a 9-amino acid alphabet. J Biol Chem 2005;280:37742–37746.
9. Launay G, Mendez R, Wodak S, Simonson T. Recognizing protein-protein interfaces with empirical potentials and reduced amino acid alphabets. BMC Bioinform 2007;8:270.
10. Luthra A, Jha AN, Ananthasuresh GK, Vishveswara S. A method for computing the inter-residue interaction potentials for reduced amino acid alphabet. J Biosci 2007;32:883–889.
11. Cannata N, Toppo S, Romualdi C, Valle G. Simplifying amino acid alphabets by means of a branch and bound algorithm and substitution matrices. Bioinformatics 2002;18:1102–1108.
12. Fan K, Wang W. What is the minimum number of letters required to fold a protein?. J Mol Biol 2003;328:921–926.
13. Li T, Fan K, Wang J, Wang W. Reduction of protein sequence complexity by residue grouping. Protein Eng 2003;16:323–330.
14. Melo F, Marti-Renom MA. Accuracy of sequence alignment and fold assessment using reduced amino acid alphabets. Proteins 2006; 63:986–995.
15. Peterson EL, Kondev J, Theriot JA, Phillips R. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics 2009;25:1356–1362.
16. Albayrak A, Out HH, Sezerman UO. Clustering of protein families into functional subtypes using relative complexity measure with reduced amino acid alphabets. BMC Bioinform 2010;11:428.
17. Etchebest C, Benros C, Bornot A, Camproux AC, de Brevern AG. A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. Eur Biophys J 2007;36: 1059–1069.
18. Reetz MT, Wu S. Greatly reduced amino acid alphabets in directed evolution: making the right choice for saturation mutagenesis at homologous enzyme positions. Chem Commun (Camb) 2008; 5499−5501.
19. Dill KA. Theory for the folding and stability of globular proteins. Biochemistry 1985;24:1501–1509.
20. Riddle DS, Santiago JV, Bray-Hall ST, Doshi N, Grantcharova VP, Yi Q, Baker D. Functional rapidly folding proteins from simplified amino acid sequences. Nat Struct Biol 1997;4:805–809.
21. Wang J, Wang W. A computational approach to simplifying the protein folding alphabet. Nat Struct Biol 1999;6:1033–1038.
22. Wang J, Wang W. Modeling study on the validity of a possibly simplified representation of proteins. Phys Rev E 2000;61:6981–6986.
23. Chang L, Wang J, Wang W. Composition-based effective chain length for prediction of protein folding rates. Phys Rev E 2010;82: 051930.
24. Murphy LR, Wallqvist A, Levy RM. Simplified amino acid alphabets for protein-fold recognition and implications for folding. Protein Eng 2000;13:149–152.
25. Plaxco KW, Simons KT, Baker D. Contact order transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
26. Plaxco KW, Simons KT, Ruczinski I, Baker D. Topology stability sequence and length: defining the determinants of two-state folding protein folding kinetics. Biochemistry 2000;39:11177–11183.
27. Baker D. A surprising simplicity to protein folding. Nature 2000; 405:39–42.
28. Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content predicts folding rates for simple two-state folding proteins. J Mol Biol 2003;327:1149–1154.
29. Ivankov DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA 2004;101:8942–8944.
30. Gromiha MM. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model 2005;45:494–501.
31. Gromiha MM, Thangakani AM, Selvaraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucl Acid Res 2006;34:W70–W74.
32. Ma BG, Guo JX, Zhang HY. Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. Proteins 2006;65:362–372.
33. Ouyang Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci 2008;17:1256–1263.
34. Jiang Y, Iglinski P, Kurgan L. Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem 2009;30:772–783.

35. Fang Y, Ma D, Li M, Wen Z, Diao Y. Investigation of the proteins folding rates and their properties of amino acid networks. Chemom Intell Lab Syst 2010;101:123–129.
36. Guo J, Rao N. Predicting protein folding rate from amino acid sequence. J Bioinform Comput Biol 2011;9:1–13.
37. Cheng X, Xiao X, Wu ZC, Wang P, Lin WZ. Swfoldrate: predicting protein folding rates from amino acid sequence with sliding window method. Proteins 2013;81:140–148.
38. Adhikari AN, Freed KF, Sosnick TR. Simplified protein models: predicting folding pathways and structure using amino acid sequences. Phys Rev Lett 2013;111:028103.
39. Huang JT, Cheng JP, Chen H. Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins 2007;67:12–17.
40. Huang JT, Cheng JP. Prediction of folding transition-state position (β_T) of small two-state proteins from local secondary structure content. Proteins 2007;68:218–222.
41. Huang JT, Tian J. Amino acid sequence predicts folding rate of middle-size two-state proteins. Proteins 2006;63:551–554.
42. Huang JT, Cheng JP. Differentiation between two-state and multi-state folding proteins based on sequence. Proteins 2008;72:44–49.
43. Huang JT, Xing DJ, Huang W. Relationship between protein folding kinetics and amino acid properties. Amino Acids 2012;43:567–572.
44. Huang JT, Huang W, Huang SR, Li X. How the folding rates of two- and multi-state proteins depend on the amino acid properties. Proteins 2014;82:2375–2382.
45. Jackson SE. How do small single-domain proteins fold? Fold Des 1998;3:R81–R91.
46. Dong R, Hu XH, Lu J. The constitution and analysis of the protein folding rate dataset. Chin Acta Biophys Sin 2012;28:509–519.
47. Bogatyreva NS, Osypov AA, Ivankov DN. KineticDB: a database of protein folding kinetics. Nucl Acid Res 2009;37:D342–D346.
48. Deshpande N, Addess KJ, Bluhm WF, Merino-Ott JC, Townsend-Merino W, Zhang Q, Knezevich C, Xie L, Chen L, Feng Z, Green RK, Flippen-Anderson JL, Westbrook J, Berman HM, Bourne PE. The RCSB protein data bank: a redesigned query system and relational database based on the mmCIF schema. Nucl Acid Res 2005; 33:D233–D237.
49. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
50. Howitt D, Cramer D. A guide to computing statistics with SPSS for windows. London: Prentice-Hall; 2001.
51. Gonzalez JR, Armengol L, Sole X, Guino E, Mercader JM, Estivill X, Moreno V. SNPassoc: an R package to perform whole genome association studies. Bioinformatics 2007;23:654–655.
52. Draper NR, Smith H. Applied regression analysis, 2nd ed. New York: Wiley; 1981.
53. Chatterjee S, Hadi AS, Price B. Regression analysis by example, 3rd ed. New York: Wiley; 2000.
54. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–132.