Prediction of Protein Folding Rates from Primary Sequences Using Hybrid Sequence Representation

YINGFU JIANG, PAUL IGLINSKI, LUKASZ KURGAN
Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, Canada, T6G 2V4

Received 21 May 2008; Accepted 8 July 2008
DOI 10.1002/jcc.21096
Published online 27 August 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: The ability to predict protein folding rates constitutes an important step in understanding the overall folding mechanisms. Although many of the prediction methods are structure based, successful predictions can also be obtained from the sequence. We developed a novel method called prediction of protein folding rates (PPFR), for the prediction of protein folding rates from protein sequences. PPFR implements a linear regression model for each of the mainstream folding dynamics including two-, multi-, and mixed-state proteins. The proposed method provides predictions characterized by strong correlations with the experimental folding rates, which equal 0.87 for the two- and multistate proteins and 0.82 for the mixed-state proteins, when evaluated with out-of-sample jackknife test. Based on in-sample and out-of-sample tests, the PPFR's predictions are shown to be better than most of other sequence only and structure-based predictors and complementary to the predictions of the most recent sequence-based QRSM method. We show that simultaneous incorporation of several characteristics, including the sequence, physiochemical properties of residues, and predicted secondary structure provides improved quality. This hybridized prediction model was analyzed to reveal the complementary factors that can be used in tandem to predict folding rates. We show that bigger proteins require more time for folding, higher helical and coil content and the presence of Phe, Asn, and Gln may accelerate the folding process, the inclusion of Ile, Val, Thr, and Ser may slow down the folding process, and for the two-state proteins increased β-strand content may decelerate the folding process. Finally, PPFR provides strong correlation when predicting sequences with low similarity.

 J Comput Chem 30: 772–783, 2009

Key words: protein folding rates; two-state folding kinetics; multistate folding kinetics; sequence-based prediction; linear regression

Introduction

Protein folding is the physical process by which a polypeptide molecule, which at the basic level consists of a linear chain of amino acids, folds into its natural, biologically active three-dimensional conformation. The determination of kinetics of protein folding and the determination of the protein's native state are challenging problems in modern molecular biology. The ability to understand and predict protein folding rates is an important and interesting step in understanding the overall folding mechanisms. Traditionally, experimental techniques used to determine folding kinetics include spectroscopic methods, mass spectrometry, NMR, hydrogen exchange, and laser-induced temperature jumps.[1–6] Two different folding mechanisms are defined: two-state and multistate kinetics. Two-state proteins fold in an "all-or-none" process,[7] while multistate proteins fold with at least one intermediate state. Usually two-state kinetics are characteristic of small proteins, while larger proteins undergo a multistate folding in which intermediates accumulate during the early stages and follow a stepwise assembly procedure.[4,7,8] Some proteins switch their folding behavior between two state and multistate by point mutation(s) brought about by changing conditions such as the salt concentration or temperature,[7,9] which blurs the division line.[10]

The experimental studies resulted in an accumulation of a sufficient amount of data[11] to build computational models. The main advantage of computational (in-silico) models is that they provide high-throughput analysis that can help in coping with the large amount of unprocessed protein sequences, i.e., while experimental studies of folding rates have been performed on several dozens proteins, there are millions of proteins which

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

Correspondence to: L. Kurgan; e-mail: lkurgan@ece.ualberta.ca

Contract/grant sponsor: NSERC Canada

await such analysis. Past years have seen several attempts toward building computational models that revealed several factors which are correlated with protein folding rates for two- and multistate proteins, and a mixture of the two categories. Initially, three-dimensional information was examined and used to predict the folding rates. Plaxco and colleagues in 1998 demonstrated that the topological complexity of the native state, defined by a parameter called relative contact order (CO), correlates well with protein folding rates.[12] Subsequently, based on the concept of contact order, long-range order (LRO),[13] total contact order,[14] and absolute contact order (Abs_CO)[10] were proposed. Other topological properties were also used by various researchers.[15–17] More recently, several attempts were made to predict folding rates from partial structural information, such as the knowledge of structural classes[18,19] and the knowledge of secondary structure assigned with DSSP.[20] We note that the structure-based methods use parameters and properties that can be computed for a relatively small number of proteins (the current version of the Protein Data Bank, which stores proteins with known structure, includes about 50,000 structures), when compared with the much larger number of known proteins (currently about 5.4 million proteins are stored in the RefSeq database). More recently, a few approaches have been developed that use only the amino acid sequences to predict folding rates. Ivankov and Finkelstein reported in 2004 that the effective chain length (Leff), which is computed using predicted secondary structure, is characterized by a significant correlation with the folding rate.[21] Huang and Tian[22] and Ma et al.[23] proposed in 2006 that the combination of several parameters, such as amino acid rigidity (*R*), secondary structure parameter (D), composition vectors (CV), chain length (L), amino acid weight (*W*), degeneracy (*D*), and composition index (CI) provide a model that allows the prediction of folding rates which are well correlated with the actual folding rates. Most recently, Huang and Gromiha developed an accurate sequence-based predictor that is based on 49 physicochemical, energetic, and conformational properties of constituent amino acids.[24]

After a thorough analysis of the existing predictors, including a recent comprehensive review concerning computational analysis and prediction of folding rates,[25] we found that no attempts were made to combine the individual, prior results to build a more accurate prediction model. To this end, we investigate whether combining information coming from multiple sources, such as the protein sequence, physicochemical properties of amino acids, and predicted secondary structure, could provide improved quality of the prediction and predictions that are complementary to outputs of the existing, top-performing methods. We implemented a new predictor, named PPFR (prediction of protein folding rates), which incorporates prediction models for three cases: two-, multi-, and mixed-state proteins (proteins for which the folding dynamics are unknown and may include two state and multistate). PPFR is based on a careful design that includes three main steps: (1) collection and optimization of both existing and innovative parameters (features) that correlate with the folding rates; (2) selection of the most promising subset of features from Step 1; and (3) design of a prediction method that takes as an input the set of features from Step 2. The first step is motivated by the fact that several of the recently proposed methods explore different types of features that are correlated with the folding rates and that can be derived from the sequences,[21–24] while no attempts were made to combine these potentially complementary parameters. We also investigate new features that are shown to provide promising results. The second step aims at the selection of a small subset of features that exhibit low correlation with each other (that are complementary to each other) and which together can be used to build an accurate model for prediction of folding rates. In the last step, we train a prediction model using the selected features and a dataset of proteins for which the folding rates were measured experimentally. We also discuss and analyze our prediction model to reveal which complementary factors govern the folding rates.

Materials and Methods

Primary sequences for a set of proteins with known experimental folding rates are used as the dataset to design and test the proposed prediction method. The sequences are first converted into a set of features, which are used to generate the prediction models after the feature selection process. Three linear regression models for the two-, multi-, and mixed-state proteins were developed.

Datasets

A benchmark dataset that includes 62 proteins, called *D62*, with known experimental folding rates that was introduced by Ivankov and Finkelstein[21] is used in this study. The dataset includes 37 two-state proteins and 25 multistate proteins. Two short artificial peptides are removed from the original dataset, as in the recent study by Ma et al.[23] Experimental folding rates are represented as the decimal logarithms of protein folding rates in water in the absence of denaturant ($\log(k_f)$), which are negatively correlated with the actual folding time. D62 dataset allows for a direct comparison with six structure-based methods[10,12–14,20,26] and two sequence-based methods[21,23] that were tested with the same dataset. This dataset includes several sequences with relatively high similarity, i.e., 2, 10, and 21 sequences share identity above 80, 50, and 35%, respectively.

In contrast to other contributions that usually test the developed predictor with one dataset, we developed a second dataset that serves as an independent benchmark. This set was created using a complete list of 77 proteins with known folding rates that was published in.[25] To create the second dataset, we removed all chains that share at least 35% pairwise sequence identity with the sequences in the D62 dataset. This allows us to verify whether the proposed method can produce accurate results when nonredundant proteins are used. We also removed one sequence for which the K-fold method,[26] which is included in the above-mentioned study, generated an error and thus could not produce predictions. As a result, the second nonredundant dataset includes eight chains and is referred to as D8.

Experimental Setup

We applied three test types to evaluate the quality of the proposed method and to compare it with competing methods: (1) resubstitution (in-sample test) on D62 dataset, (2) jackknife (out-of-sample test) on the D62 dataset, and (3) independent test

Table 1. Weight (w), Degeneracy (d), Flexibility (f), p_α, p_β, and p_{turn} Values.

AA	w	d	f	p_α	p_β	p_{turn}	AA	w	d	f	p_α	p_β	p_{turn}
A (Ala)	89.0935	4	0.95	1.25	0.89	0.78	M (Met)	149.2124	1	0.86	1.43	0.99	0.39
C (Cys)	121.1590	2	0.88	1.12	0.85	0.80	N (Asn)	132.1184	2	1.01	0.87	0.86	1.28
D (Asp)	133.1032	2	1.09	1.03	0.74	1.41	P (Pro)	115.1310	4	1.08	0.60	0.71	1.91
E (Glu)	147.1299	2	1.04	1.45	0.65	1.00	Q (Gln)	146.1451	2	1.03	1.24	0.82	0.97
F (Phe)	165.1900	2	0.91	1.08	1.22	0.58	R (Arg)	174.2017	6	1.03	0.99	1.02	0.88
G (Gly)	75.0669	4	1.04	0.57	0.93	1.64	S (Ser)	105.0930	6	1.05	0.82	0.96	1.33
H (His)	155.1552	2	0.95	1.25	1.04	0.69	T (Thr)	119.1197	4	1.05	0.81	1.13	1.03
I (Ile)	131.1736	3	0.89	0.94	1.41	0.51	V (Val)	117.1469	4	0.93	0.88	1.48	0.47
K (Lys)	146.1882	2	1.08	1.24	0.81	0.96	W (Trp)	204.2262	1	0.92	1.03	1.15	0.75
L (Leu)	131.1736	6	0.96	1.32	1.03	0.59	Y (Tyr)	181.1894	2	0.93	0.75	1.25	1.05

(out-of-sample test) in which the prediction method is trained on the D62 dataset and tested using the D8 dataset. The resubstitution test checks the quality of the generated model on the (training) data used to establish the model. This test verifies whether the model accurately describes the training dataset. At the same time, it does not show how well the prediction model can perform on data which are not included in the training set. Although this test may overestimate the quality, it was extensively used to evaluate prior prediction methods[10,12–14,20,21,23,25] and thus is also included in this contribution. The jackknife test, also called the leave-one-out test, is an n-fold crossvalidation, where n is the total number of instances in the dataset. The test is repeated n times, each time using one sequence to test the prediction model, and the remaining n-1 sequences to establish the model. This test assures that the results are not biased toward the training datasets and that the estimated prediction performance reflects future application of the prediction model. The jackknife test was used to evaluate only some of the recent methods.[23,24] Although no prior attempts were made to evaluate folding rate prediction methods using a set of nonredundant sequences, we include the third test type due to some overlap with respect to the sequence similarity within the D62 dataset.

Secondary Structure Prediction

Secondary structures predicted from the protein primary sequences are used to derive some of the features that are utilized by the proposed method. Prior folding rate prediction methods used the predicted secondary structure to derive the number of residues in α-helical conformation and the number of α-helices in the sequence.[21] In contrast to this approach, we use the predicted structure to derive information about the three major secondary structures that include α-helices, β-strands, and coils. We applied two secondary structure predictors: (1) PROTEUS, which was recently shown to provide superior prediction accuracies when compared with nine other prediction methods[27] and (2) PSIPRED,[28,29] which was extensively used in numerous protein structure prediction methods.[30–32]

Features

The features used to encode protein sequences are divided into four categories: composition-based features, property-based features, predicted secondary structure-based features, and sequence length-based features.

Composition-Based Features

We consider CVs composed of 20 features defined as:

$$\text{CV\_}X = \frac{N_X}{L}$$

where N_X, $X = 1, 2, \ldots, 20$, is the count of the occurrences of amino acid X in a given sequence (see Table 1 for one-letter encoding), and L is the total number of residues in a sequence.

Three composite composition-based features and three custom-designed features proposed in[23] are defined as follows:

$$\begin{aligned}
\text{C_NQIVT} &= (N_{\text{N}} + N_{\text{Q}} - N_{\text{I}} - N_{\text{V}} - N_{\text{T}})/L \\
\text{C_AG} &= (N_{\text{A}} + N_{\text{G}})/L \\
\text{C_NQTVC} &= (N_{\text{N}} + N_{\text{Q}} - N_{\text{T}} - N_{\text{T}} - N_{\text{V}} - N_{\text{C}})/L \\
\text{CI_NQIVT} &= \text{C_NQIVT} + 0.23\ W/(L^*D) \\
\text{CI_AG} &= \text{C_AG} + 7.42\ W/(L^*D) \\
\text{CI_NQTVC} &= \text{C_NQTVC} + 0.46\ W/(L^*D)
\end{aligned}$$

where W is the weight value and D is the degeneracy value; both are described below. We note that the original formula for C_NQTVC[23] shows N_{I}, while the erratum corrects it to N_{T}.

Property-Based Features

Mean molecular weight and degeneracy (which is related to translation speed in the ribosome) of a given protein as proposed by Ma et al.[23] are defined as follows:

$$D = \frac{\sum_{i=1}^{L} d_i}{L} \qquad W = \frac{\sum_{j=1}^{L} w_i}{L}$$

where w_i is the weight and d_i is the degeneracy value of the ith residue, see Table 1.

Flexibility (which is related to temperature factors, i.e., B-factors, of the C_α atoms) and three normalized frequencies of occurrence of the residues in α-helix, β-sheet, and reverse turn

(which are suggested to act as a nucleation center during protein folding) secondary structure conformations from[22] are defined as:

$$F = \frac{\sum_{i=1}^{L} f_i}{L} \quad P_\alpha = \frac{\sum_{i=1}^{L} p_{\alpha i}}{L} \quad P_\beta = \frac{\sum_{i=1}^{L} p_{\beta i}}{L} \quad P_{turn} = \frac{\sum_{i=1}^{L} p_{\mathrm{turn}i}}{L}$$

where f_i, $p_{\alpha i}$, $p_{\beta i}$, $p_{\mathrm{turn}i}$ are the property values of the ith residue, see Table 1.

A set of 49 normalized amino acid properties proposed by Gromiha et al.[18,19] were also used in this study. The average amino acid property for each protein, P_g was computed as

$$P_g = \frac{\sum_{i=1}^{L} p_{ig}}{L}$$

where p_{ig} is the property value of the ith residue and $g = 1, 2, \ldots, 49$ is the index of a given property.

Predicted Secondary Structure-Based Features

The three-state secondary structure predictions computed using PSIPRED and PROTEUS prediction servers were used to generate a set of features for each sequence. The motivation to compute these features comes from the work of Ivankov and Finkelstein,[21] who have shown that size (length) and number of secondary structure segments (in their case α-helix segments only) provides useful information that helps in the prediction of folding rates. On the basis of this observation, we designed a wide range of features that are based on all three secondary structures:

1. Composition vector, CV_y for y = {h, e, c} and where h denotes α-helix, e denotes β-strand, and c denotes coil. CV_h and CV_e are equivalent to the secondary structure content.
2. Composition moment vector

$$\mathrm{CMV}k\_y = \frac{\sum_{i=1}^{L} n_{yi}^{k}}{\prod_{d=1}^{k} (L-d)}$$

where y = {h,e,c}, n_{yj} represents the jth position of the yth secondary structure, and $k = 1, 2, \ldots, 5$ is the order of the CMV. For $k = 0$ CMVk_y reduces to CV_y.

3. Normalized count of segments that include at least k residues

$$N\mathrm{Count}Lk\_y = \frac{\sum_{j=k}^{20} \mathrm{count}_y^i}{\sum_{\substack{w=\{\mathrm{h,e,c}\}\ if\ y=\mathrm{c} \\ w=\{\mathrm{h,e}\}\ if\ y=\{\mathrm{h,e}\}}} \mathrm{total}_w}$$

where y = {h,e,c}, $k = 2, 3, \ldots, 20$ for β-strand and coil segments and $k = 2, 3, \ldots, 20$ for α-helix segments, count_y^j denotes the number of α-helix, β-strand, and coil segments of length j, respectively, and total_w denotes total number of all segments belonging to the wth secondary structures. The smallest α-helix segment is assumed to include at least three residues. The count of coil segments is normalized by the total number of all segments, while the counts of β-strand and α-helix segments are normalized by the total number of β-strand and α-helix segments. Different normalizations accommodate for the all-α and all-β structural classes that may not include any β-strand and α-helix segments, respectively.

4. Length of the longest segment *MaxSeg_y* for y = {h,e,c}
5. Normalized length of the longest segment

$$N\mathrm{LongestSegment}\_y = \mathrm{MaxSegment}\_y/L \quad \text{for } y = \{\mathrm{h,e,c}\}$$

6. Average length of the segment AvgSegment_y for y = {h,e,c}
7. Normalized average length of the segment

$$N\mathrm{AvgSegment}\_y = \mathrm{AvgSegment}\_y/L \quad \text{for } y = \{\mathrm{h,e,c}\}$$

Length-Based Features

Length (L) is defined as the number of residues in a protein sequence chain.

The effective length and derived features from Ivankov and Finkelstein[21] are calculated as follows:

$$L_{\mathrm{eff}}\_pm = L - L_{\mathrm{H}} + l_1 \times N_{\mathrm{H}}$$

$$L\mathrm{eff}\_pm\_P = (L_{\mathrm{eff}})^P$$

where pm = {*proteus*, *psipred*} defines the *p*rediction *m*ethod (pm) used to derive the secondary structure, L_{H} is the predicted number of residues in helical conformation, N_{H} is the predicted number of helical segments, l_1 is a constant, and p is an optimal exponent value. As suggested in ref. 21, $l_1 = 3$ and $p = 0.1$ were used.

Since Ivankov and Finkelstein have optimized the exponent value assuming only integer values, we experimented with floating point exponents and different predicted secondary structures. We applied a two-step procedure to find a suitable exponent and secondary structure prediction method:

1. The value of *Leff_pm_P* was computed for all considered p values and both the PROTEUS and PSIPRED predicted secondary structures, the resulting features were merged, and a feature selection was performed.
2. The selected features were used, one at the time, to develop linear regression models to predict the folding rate, and the feature (and its corresponding p value and secondary structure prediction methods) that gave the highest correlation between the experimental (actual) and the predicted folding rate was retained.

Table 2. Results of Finding Floating Point p Value and Secondary-Structure Prediction Method for *Leff_pm_P* Feature.

Leff_pm_P feature		Number of cross validation folds	
p	Secondary-structure prediction method	in which the feature was selected by CFSS	Correlation coefficient
0.0001	PROTEUS	53	0.668
0.0002	PSIPRED	57	0.725
0.0011	PROTEUS	58	0.750
0.0011	PSIPRED	61	0.752
0.0010	PSIPRED	62	0.747
0.0020	PROTEUS	14	0.741

To avoid overfitting, both steps were performed using jackknife tests on the D62 dataset. The feature selection in step 1 was performed using the correlation-based feature subset selection (CFSS) method described in the Feature Selection section. The regression model from step 2 is explained in the Prediction Model section. We assumed p values ranging between 0.0001 and 0.2 (values below 0.0001 and above 0.2 were rejected by the feature selection in step 1) and assumed that step 2 received only those features that were selected by CFSS in at least 10 cross validation folds. Table 2 lists the *Leff_pm_P* features that were selected by CFSS. It shows the p value and prediction method for the selected features, together with the number of cross validation folds in which the feature was selected and the corresponding correlation coefficient when this feature was used to predict the folding rate.

The two configurations that have correlations above 0.75 correspond to $p = 0.0011$, for both the PROTEUS and PSIPRED predicted secondary structures. Thus, these two corresponding features were selected.

We also used the effective length formula by considering β-strand segments:

$$L_{\text{eff\_}}pm\_E = L - L_{\text{H}} + l_1 \times N_{\text{H}} - L_{\text{E}} + l_2 \times N_{\text{E}} \quad \text{for } pm = \{proteus, psipred\}$$

where L_{H}, l_1, N_{H} are as described earlier, L_{E} is the number of residues in β-strand conformation, N_{H} is the number of β-strands, and l_2 is a constant. The value of $l_2 = 15$ was selected following the same procedure as for the selection of p.

Prediction Model

The folding rate prediction was performed using a linear regression predictor

$$\text{folding\_rate}_s = \sum_{j=1}^{k_s} w_{sj} x_{sj} + w_{s0}$$

where $s = \{\text{two state, multistate, mixed-state}\}$ corresponds to the folding dynamics type, x_{sj} corresponds to the jth feature for the sth folding dynamics type, K_s is the number of features for the sth folding dynamics type, and w_{sj} is the jth feature's regression coefficient for the sth folding dynamics type.

The values of the regression coefficients were estimated from the data using Weka, which is a comprehensive open-source library of machine learning methods.[33] Linear regression was also used to develop two other recent folding rate prediction methods.[22,23]

Feature Selection

As this study combines features developed in several existing prediction methods with a set of newly proposed features (in total 270 features), a feature selection was performed to reduce the dimensionality. The feature selection process was divided into two steps.

1. Removal of weak features.
2. Removal of correlated features among the features selected in step 1.

In step 1 we applied the CFSS method,[34] which was also successfully applied to develop a recent logistic linear regression based method for structural class prediction.[35] CFSS evaluates a given subset of features (determined using a best-first search based on hill-climbing with backtracking) by considering the individual predictive ability of each feature along with the degree of redundancy between them. The feature selection was performed using jackknife tests and only the features that are selected in at least one cross validation fold are passed to step 2, while the remaining features are removed. The cross validation performed to select features should assure that the selected features do not overfit the dataset.

Step 2 is performed by removing individual features among the feature subset that was produced in step 1. Initially, the full set of features selected in step 1 was used to develop a linear regression model using jackknife tests and the corresponding correlation between the predicted and the actual folding rates was computed. Next, a given feature was removed only if the removal did not decrease the correlation coefficient of the jackknife-based prediction using the reduced set of features. After the feature was removed, the process was repeated again until no feature could be removed. This allows obtaining a small set of complementary features that in tandem provide accurate predictions.

The above feature selection was performed separately for each of the three datasets, i.e., two-, multi-, and the mixed states. As a result, the initial set of 270 features was reduced in Step 1 to 39, 144, and 22 features, respectively, for the two-, multi-, and the mixed-state datasets. After step 2, the corresponding number of retained features was 10, 10, and 8, respectively.

The selected features, which are divided into four categories for each dataset, are shown in Table 3. The majority of the selected features were from among those proposed in this paper. The features proposed in the past works are denoted in Table 3, and account for 30, 20, and 37% of the features used for the two-, multi- and the mixed-state proteins, respectively. We observe that the selected number of features is comparable with the number features included in several competing sequence-based methods. More specifically, the most recent method

Table 3. Selected Features for the Three Types of Folding Kinetics Including Their Correlation with the Folding Rates.

Folding kinetics	Feature name	Category	Correlation
Two state	Leff_proteus_0011	Length	−0.74
	C_NQIVT[a]	Composition	0.68
	proteus_CMV4_e	Predicted secondary structure	−0.63
	proteus_CountL10_e	Predicted secondary structure	−0.53
	proteus_NLongestSegment_c	Predicted secondary structure	0.47
	psipred_CountL10_e	Predicted secondary structure	−0.46
	Pf_s[a]	Property	0.32
	proteus_CountL3_c	Predicted secondary structure	0.30
	psipred_CountL19_c	Predicted secondary structure	−0.2
	CV_S[a]	Composition	−0.06
Multistate	L[a]	Length	−0.80
	Leff_proteus[a]	Length	−0.75
	psipred_NAvgSegment_c	Predicted secondary structure	0.58
	psipred_LongestSegment_h	Predicted secondary structure	−0.30
	proteus_NLongestSegment_e	Predicted secondary structure	0.30
	psipred_CountL6_c	Predicted secondary structure	0.25
	proteus_AvgSegment_e	Predicted secondary structure	−0.22
	psipred_CountL20_c	Predicted secondary structure	−0.17
	proteus_CountL20_c	Predicted secondary structure	0.13
	psipred_CountL3_c	Predicted secondary structure	−0.10
Mixed state	Leff_proteus_0011	Length	−0.77
	Leff_psipred_0011	Length	−0.77
	CV_Q[a]	Composition	0.37
	proteus_CountL10_c	Predicted secondary structure	0.33
	proteus_CountL4_c	Predicted secondary structure	0.27
	CV_N[a]	Composition	0.25
	psipred_CountL11_e	Predicted secondary structure	−0.16
	CV_F[a]	Composition	0.13

For each folding kinetics type, the features are ordered by decreasing absolute value of the correlation with the folding rate on D62 dataset.

[a] Denotes features used in existing prediction methods.

QRSM method includes 49 features to predict mixed-mode proteins,[24] while the regression-based method by Ma et al. uses eight features for the two-state model, five features for the multistate model, and eight features in the case of the mixed-state model.[23] The largest portion of the selected features was computed based on statistics drawn from the predicted secondary structure. The features derived from sequence length augmented with information from the predicted secondary structure were also found useful for predictions of all three folding types. The sequence composition features are selected only for the two- or mixed-state type proteins, while only one physicochemical property-based feature is used for the two-state proteins. Our results with respect to using effective length formulae that consider β-strand segments are consistent with Ivankov and Finkelstein et al.[21] The corresponding features were not selected for prediction of mixed-state proteins, which shows that they would not help in improving these predictions. Similarly as in ref. 21, we hypothesize that the reason could be that β-strand and α-helix contents are strongly negatively correlated.

Table 3 also includes the correlation coefficients between the selected features and the folding rates in the D62 dataset. This allows establishing relative importance of features, although we note that features with lower correlation should not be considered less useful since they are complementary to features with higher correlation. We observe that length and effective lengths features are characterized by strong correlation of above 0.7. These are followed by features based on the predicted secondary structure and sequence composition for which correlation values above 0.5 are obtained.

Results and Discussion

The proposed PPFR method is based on three linear regression models where each model corresponds to different folding kinetics. In case the user is uncertain about which model should be used, i.e., the predicted sequence could be either two state or multistate, (s)he should use the model for mixed-state proteins.

Prediction Results

Using the jackknife test on the D62 dataset, Table 4 compares predictions obtained with linear regression models computed for

Table 4. Comparison of Correlation Coefficients Between the PPFR Predicted and the Experimental Folding Rates on D62 Dataset Using the Jackknife Test for All Features, and Features Selected in Step 1 and Step 2 of the Feature Selection.

Features used to represent sequences	Correlation coefficient		
	Two state	Multistate	Mixed state
All features	0.72	0.44	0.69
Features after step 1 of the feature selection	0.81	0.41	0.78
Features after step 2 of the feature selection (final feature sets)	0.87	0.87	0.82

the original set of 270 features, for the features selected in Step 1 of the feature selection procedure, and for the final set of features. The table shows correlation coefficients between the predicted and the actual (experimental) folding rates. We observe that the feature selection helps to better tune the prediction models, i.e., the final feature sets are not only substantially smaller than the original set, but most importantly they result in a better quality of folding rate predictions. PPFR predictions for all three types of folding kinetics are strongly correlated with the actual folding rates. In the cases of the two- and multi-state types, the correlation coefficient equals 0.87, while for the mixed-state proteins it equals 0.82. We note that although the increase in number of features in some prediction methods leads to improved accuracy/correlation, in our case additional features may result in lowering the quality. We believe that this behavior is specific to regression models that behave poorly when the input features are characterized by strong correlation with each other. The smaller feature sets behave in a more complementary fashion to yield more accurate predictions.

Table 5 lists the predictions obtained with the mixed-state model of the PPFR method for all three types of tests including resubstitution and jackknife tests on the D62 dataset and the test on the D8 dataset. The mean average error (MAE) for the resubstitution test on D62 equals 0.88 and for the jackknife it equals 0.93, while for the test on the independent dataset, the MAE is 1.18. The corresponding correlation coefficients equal 0.85, 0.82, and 0.76 for the resubstitution, jackknife and independent tests, respectively. The strong correlations obtained for all the three tests indicate that PPFR obtains comparable results with out-of-sample and in-sample tests, which in turns supports a claim that our design does not result in overfitting the D62 dataset.

Factors Governing the Folding Rates

The linear regression models for the two-, multi-, and mixed-state proteins, which were computed using the D62 dataset, are shown in Table 6. The individual models show regression coefficients for each of the input features. The sign of the coefficient indicates whether a given feature is positively or negatively correlated with the experimental folding rate. We observe that our model not only indicates which features (factors) are related to the folding rate, but most importantly it indicates which of these factors are complementary with each other, i.e., which could be used in tandem to improve predictions. We concentrate our discussion upon those features that have a correlation coefficient with the folding rates of at least 0.5 in absolute value (see Table 3).

Considering that the experimental folding rates are negatively correlated with the folding time we observe that sequence length (L feature) and effective sequence length (Leff_proteus_0011, Leff_psipred_0011, and Leff_proteus features), which considers predicted α-helix segments, are positively correlated with the folding time. This agrees with results shown in ref. 21 and shows that bigger proteins require more time for folding and that high helical content may accelerate the folding process. The main reasons for the latter relation is that some preformed helices may already exist in the unfolded state of the chain and/or because the helices are rapidly formed in the course of folding.[21]

Feature C_NQIVT in the two-state model (see the definition in the Features section) shows that increased amounts of Asn (N) and Gln (Q) are negatively correlated with the folding time and it also shows that a positive correlation between the folding time and the content is observed for Ile (I), Val (V), and Thr (T). These relations are consistent with the results of Ma et al.[23] This is also supported by inclusion of the sequence composition-based features of Phe (F), Asn (N), and Gln (Q) amino acids in the mixed-state model that are characterized by a negative correlation with the folding time (positive regression weights and positive correlation shown in Table 3) and for Ser (S) in the two-state model that is positively correlated with the folding time. The amide amino acids Asn and Gln could be implicated in accelerating the folding process in two-state proteins because of more probable hydrogen exchanges with the solvent that stabilizes the transition state structure and lowers the energy barrier in the folding path.[36–38] At the same time, Ile, Thr, and Val are characterized by branched side chains, which may result in slowing down the folding process by enlarging the number of potential conformations.[17,39,40] Similarly as in ref. 23 we note that hydrophobicity, which is one of the potential factors that affects folding time due to stabilizing effects of hydrophobic regions,[41,42] was not confirmed by our results. For instance, hydrophobic amino acids Ile, Val, and Ser are positively correlated, while Phe which is another hydrophobic amino acid is negatively correlated with the folding time. Also, among the three hydrophilic amino acids, Asn and Gln are negatively correlated while Thr is positively correlated.

The features based on the predicted secondary structure show that for the two-state proteins the increased count of β-strand segments (proteus_CountL10_e) and positional composition (proteus_CMV4_e) are positively correlated with the folding time (negatively correlated with the folding rate), which could mean that formation of β-sheets slows down the folding process. Finally, we observe that increased coil content (psipred_NAvgSegment_c, proteus_NLongestSegment_c, proteus_CountL10_c, and proteus_CountL3_c features) is negatively correlated with folding time (positively correlated with the folding rate) indicating that formation of coils may accelerate the folding process. The other selected coil-based features (i.e., psipred_CountL3_c, proteus_CountL4_c, psipred_CountL6_c, psipred_CountL19_c,

Table 5. Test Results Obtained with the PPFR Method on the D62 Dataset Based on Resubstitution and Jackknife Tests and Based on the Nonredundant D8 Dataset When Using D62 for Training.

Dataset	PDBid	Folding-kinetics type	Experimental folding rate $\log_{10}(k_f)$	Predicted $\log_{10}(k_f)$	
				Resubstitution	Jackknife
D62	1PIN	two state	4.1	2.836	2.753
	2PDD	two state	4.3	2.756	2.652
	2ABD	two state	2.9	2.472	2.458
	256B	two state	5.3	2.969	2.854
	1IMQ	two state	3.2	2.581	2.557
	1LMB	two state	3.7	2.498	2.449
	1FNF(90)	two state	−0.4	1.428	1.519
	1WIT	two state	0.2	1.059	1.091
	1TEN	two state	0.5	1.257	1.290
	1SHG	two state	0.6	1.617	1.650
	1SRL	two state	1.7	1.824	1.826
	1PNJ	two state	−0.5	1.253	1.307
	1SHF	two state	2	1.671	1.662
	1PSF	two state	1.4	2.149	2.213
	1CSP	two state	2.9	2.498	2.478
	1C9O	two state	3.1	2.564	2.533
	1G6P	two state	2.7	1.844	1.780
	1MJC	two state	2.3	2.052	2.043
	1LOP	two state	2.9	1.259	1.137
	1C8C	two state	3	2.140	2.089
	1HZ6	two state	1.8	1.629	1.499
	1PGB(57)	two state	2.6	1.968	1.942
	1FKB	two state	0.7	1.030	1.037
	2CI2	two state	1.7	1.807	1.809
	1AYE	two state	3	2.151	2.119
	1URN	two state	2.5	1.823	1.798
	1APS	two state	−0.7	1.226	1.337
	1RIS	two state	2.6	1.710	1.675
	1POH	two state	1.2	1.849	1.877
	1DIV	two state	2.6	2.730	2.736
	2VIK	two state	3	1.477	1.422
	1L2Y	two state	5.4	4.374	4.407
	1VII	two state	5	3.660	3.456
	1BDD	two state	5.1	3.839	3.688
	1ENH	two state	4.6	3.385	3.317
	2ACY	two state	0.4	1.505	1.625
	1L8W	two state	0.7	0.745	0.734
	1A6N	multistate	0.5	1.924	1.995
	1CEI	multistate	2.5	2.470	2.471
	2CRO	multistate	1.6	2.256	2.308
	2A5E	multistate	1.5	1.251	1.258
	1TIT	multistate	1.6	1.400	1.396
	1HNG	multistate	0.8	1.158	1.179
	1FNF(94)	multistate	2.4	1.304	1.241
	1IFC	multistate	1.5	0.864	0.817
	1EAL	multistate	0.6	0.915	0.926
	1OPA	multistate	0.6	1.426	1.460
	1CBI	multistate	−1.4	0.784	0.865
	1QOP(268)	multistate	−1.1	0.543	0.608
	1AON	multistate	0.3	0.706	0.722
	1BRS	multistate	1.5	2.012	2.029
	3CHY	multistate	0.4	1.269	1.303
	2RN2	multistate	0	1.061	1.109
	1RA9	multistate	2	0.691	0.644

(*Continued*)

Table 5. *(Continued)*

Dataset	PDBid	Folding-kinetics type	Experimental folding rate $\log_{10}(k_f)$	Predicted $\log_{10}(k_f)$ Resubstitution	Jackknife
1QOP(396)	multistate	−3	0.109	0.287	
	1PHP(175)	multistate	1	0.722	0.708
	1PHP(219)	multistate	−1.5	0.536	0.627
	1BNI	multistate	1.1	1.551	1.562
	2LZM	multistate	1.8	1.501	1.492
	1UBQ	multistate	2.6	1.833	1.805
	1SCE	multistate	1.8	1.544	1.533
	1GXT	multistate	1.9	1.634	1.593
				Predicted $\log_{10}(k_f)$ nonredundant dataset	
D8	1HRC	two state	3.8	2.085	
	1YCC	two state	4.18	2.089	
	1NYF	two state	1.97	1.985	
	1PKS	two state	−0.46	1.415	
	2AIT	two state	1.8	1.336	
	2HQI	two state	0.08	1.732	
	1PBA	two state	3	2.243	
	1HX5	multistate	0.32	1.210	

psipred_CountL20_c, proteus_CountL20_c, and psipred_CountL3_c) are characterized by low, i.e. < 0.3, correlation coefficient values, see Table 3.

Comparison with Competing Prediction Methods

The prediction quality of the proposed PPFR method, which is measured based on the correlation coefficient between the predicted and the experimental (actual) folding rates using the resubstitution and jackknife tests on the D62 dataset, is compared with six structure-based methods, including CO,[12] Abs_CO,[10] LRO,[13] TCD,[14] SSC,[20] K-Fold,[26] and two sequence-based methods Leff[21] and CI[23]; see Tables 7 and 8. Table 8 also includes results of the recently proposed QRSM method[24] which was designed and tested using a larger set of 77 proteins. It includes all chains from the D62 dataset and 15 other sequences, which is why resubstitution results on this dataset cannot be compared against results reported in Table 7.

For all three folding kinetics and both the resubstitution and the jackknife tests, PPFR shows improvements when compared with most of the existing methods including structure- and sequence-based methods. When the resubstitution test is applied to the two-, multi-, and mixed-state proteins, PPFR achieves correlation coefficients of 0.92, 0.92, and 0.85, respectively, which are 0.13, 0.15, and 0.12 higher than the correlation coefficients obtained by any of the competing methods. Using the jackknife test, our method obtains correlation coefficients of 0.87, 0.87, and 0.82, which are 0.14, 0.17, and 0.09 higher than the result of the recently proposed sequence-based CI method for the two-, multi- and mixed-state proteins, respectively. Similar improvements are observed when comparing our results with a recently proposed structure-based K-Fold method that predicts folding rates for mixed-state proteins.

We observe that PPFR results fall somewhat short of the results obtained with the recently proposed QRSM method.[24] We note that the jackknife results for QRSM method were computed with a dataset of 77 sequences that contains 10 sequences

Table 6. Prediction Models for Two-, Multi-, and Mixed-State Proteins.

Two-state proteins

$$\text{folding\_rate}_{\text{two state}} = -8.2790*\text{CV\_S} + 3.7617*\text{C\_NQIVT} + 7.0623*\text{Pf\_s} - 572.4288*\text{Leff\_proteus\_0011} - 5.3806*\text{proteus\_CMV4\_e} + 2.0119*\text{proteus\_CountL3\_c} - 2.1168*\text{proteus\_CountL10\_e} + 1.9751*\text{proteus\_NLongestSegment\_c} - 11.0500*\text{psipred\_CountL19\_c} - 2.6765*\text{psipred\_CountL10\_e} + 576.2708$$

Multistate proteins

$$\text{folding\_rate}_{\text{multistate}} = -0.0032 * \text{L} - 0.0083*\text{Leff\_proteus} + 4.8614*\text{proteus\_CountL20\_c} + 11.6171*\text{proteus\_NLongestSegment\_e} - 0.1965*\text{proteus\_AvgSegment\_e} - 5.1997*\text{psipred\_CountL3\_c} + 2.3357*\text{psipred\_CountL6\_c} - 13.2128*\text{psipred\_CountL20\_c} - 0.0217*\text{psipred\_LongestSegment\_h} + 14.9313*\text{psipred\_NAvgSegment\_c} + 3.8541$$

Mixed-state proteins

$$\text{folding\_rate}_{\text{mixed state}} = 3.2268*\text{CV\_F} + 4.8074*\text{CV\_N} + 6.5496*\text{CV\_Q} - 551.0237*\text{Leff\_proteus\_0011} + 1.1418*\text{proteus\_CountL4\_c} + 1.9093*\text{proteus\_CountL10\_c} - 561.0221*\text{Leff\_psipred\_0011} - 2.0846*\text{psipred\_CountL11\_e} + 1117.9965$$

Table 7. Comparison of Correlation Coefficients Between the Predicted Folding Rates (Using the Resubstitution Test) and the Actual Folding Rates for Different Prediction Methods for the Two-, Multi- and Mixed-State Proteins.

Folding kinetics	CO[a]	Abs_CO[a]	LRO[a]	TCD[a]	SSC[a]	Leff[a]	*CI*[b]	PPFR
Two state	−0.57	−0.64	−0.79	−0.79	0.64	−0.61	0.73	0.92
Multistate	0.43	−0.44	−0.34	0.23	−0.01	−0.77	0.70	0.92
Mixed state	0.12	−0.57	−0.61	−0.19	0.42	−0.73	0.72	0.85

All methods were tested on the D62 dataset.
[a] Denotes results from ref. 23.
[b] Denotes results from ref. 23 that were corrected based on personal communication with the authors.

with 100% pairwise similarity with respect to the other sequences in the dataset, 19 with above 80% identity, 28 with above 50% identity, and 33 with above 35% identity. In contrast, the D62 dataset is more difficult as it includes only two sequences that share identity of above 80%, 10 with above 50% identity, and 21 with above 35% identity. Most importantly, we observe that the prediction of PPFR and QRSM are complementary. We extracted ten chains for which PPFR obtains the biggest MAE (1QOP(396), 256B, 1CBI, 1PHP(219), 1YCC, 1APS, 1FNF(90), 1PKS, 1PNJ, and 1LOP). The MEA of our method for these chains equals 2.1 while the MAE of QRSM for these chains equals 0.7. At the same time, for the ten chains for which the QRSM method obtains the largest MAE (1L8W, 1PKS, 1PSF, 1PIN, 1HX5, 1VII, 1ENH, 2HQI, 1BRS, and 1DIV), the MAE of PPFR and QRSM equal 1.0 and 1.8, respectively. The only chain that appears in both groups of high MAEs is 1PKS, for which our method obtained a slightly lower 1.9 MAE when compared with a 2.0 MAE obtained by QRSM. The reason for this complementarity comes (1) from different inputs used by each of the methods, i.e., PPFR uses effective sequence length, predicted secondary structure, and sequence composition while QRSM is based on indices that describe physicochemical, energetic, and conformational properties of the constituent amino acids and (2) from the different prediction models used, i.e., PPFR applies linear regression while QRSM implements a quadratic response surface model. We also note that QRSM introduces mixed-state models that are designed for specific structural classes (all-α, all-β, and mixed), and these models are shown to be characterized by very high correlation with the folding rates (although usage of these models requires a priori knowledge of a structural class for a given input sequence). At the same time, PPFR includes models for two- and multistate proteins while QRSM predictions are based only on the mixed-state model.

Table 9 shows the results obtained on the D8 dataset. We compare the results obtained by PPFR with the results obtained by K-Fold[26] and QRSM.[24] We note that the results of PPFR and K-Fold are consistent, i.e., both methods were training with the D62 dataset and tested on the D8 dataset, while the results for QRSM are based on the jackknife predictions for the chains from the D8 dataset, i.e., QRSM was developed with dataset of 77 chains that already incorporates the chains from D8 dataset. The table shows that PPFR is capable of producing folding rates that are strongly correlated with the actual rates for proteins with low identity. This is in contrast to the K-Fold method that obtained a correlation of 0.14, which could be explained by the simplicity of this method that uses only one input feature. We also note that the results of QRSM also suggest that this method can provide high quality predictions for nonredundant sequences, although in this case the results are based on a larger dataset that includes a large fraction of similar sequences. Because of the limited size of the D8 dataset (we note that no additional sequences could be added to this dataset since D8 and D62 together represent all sequences with known folding rates), we also analyzed jackknife results on the D62 dataset when limiting them to those sequences that share a given maximal pairwise identity with other sequences in this dataset. The corresponding correlation coefficients equal 0.82, 0.82, 0.74, and 0.75 when using all sequence and sequences with at most 50, 35, and 30% identity, respectively. These results are consistent with the results obtained on the D8 dataset, confirming that the proposed

Table 8. Comparison of Correlation Coefficients Between the Predicted Folding Rates (Using the Jackknife Test) and the Actual Folding Rates for Different Prediction Methods for the Two-, Multi- and Mixed-State Proteins.

Folding kinetics	CI[a]	*K*-Fold[b]	QRSM[c]	PPFR
Two state	0.73	N/A	N/A	0.87
Multistate	0.70	N/A	N/A	0.87
Mixed state	0.73	0.74	0.89	0.82

[a] Denotes results from ref. 23.
[b] Denotes results from ref. 26 where 5-fold cross validation was performed and only the mixed-state model was developed.
[c] Denotes results from ref. 24 where a different set of 77 proteins was used and only the mixed-state model was developed.

Table 9. Comparison of Correlation Coefficients Between the Predicted Folding Rates and the Actual Folding Rates for Different Prediction Methods Using a Mixed-State Model and the D62 Dataset to Train the Model and Nonredundant D8 Dataset for Testing.

Folding kinetics	*K*-Fold	QRSM[a]	PPFR
Mixed state	0.14	0.81	0.76

[a] Denotes jackknife test results from ref. 24 where chains from the D8 dataset were included in the set of 77 proteins used in the jackknife test.

method is capable of providing high quality predictions for nonredundant chains.

The results obtained with PPFR are shown to be consistent over the three types of tests, i.e., the model is not only capable of capturing the relation between the input features and the folding rate (which is shown by the resubstitution test) but most importantly it can be also used to perform successful predictions for unseen proteins (which is shown by the jackknife test) and for unseen proteins that share low similarity with proteins used to develop the model (which is shown by the test on the D8 dataset). We emphasize that PPFR does not utilize actual (experimentally determined) secondary structural information to perform the predictions. Similarly as in case of the QRSM method, this allows PPFR to be used in wide-spread applications where only the amino acid sequence is known.

Limitations

Although the proposed method provides high-prediction quality, it does not take into account two important factors, namely, mutations and solvents, which also affect the folding rates. PPFR incorporates features that are computed based on the sequence and the predicted secondary structure, which do not take into account the position of the residues. We note that some mutations would not affect the secondary structure. Since position specific mutation(s), and especially mutations that do not change the secondary structure, could not be properly identified by our features, our model is not capable of accurate predictions for some proteins in which mutations could change the folding rates by as much as two orders of magnitude.[43] Our method assumes prediction of in-water folding rates, which means that it does not take into account a solvent-induced change in protein stability, which can change the folding rate manifolds.[43,44] This results in a precision of +/− an order of magnitude for our method. However, this is a relatively small error in the context of the 10 orders of magnitude difference in observed protein folding rates. We also note that the same drawbacks are characteristic of all other sequence-based prediction methods.[21–24]

Conclusions

The proposed PPFR method aims at providing accurate high-throughput predictions of protein folding rates from protein sequences. Our method addresses the absence of a methodology that combines multiple factors that can be extracted from the sequence and that could influence the folding rate. The main limitation of the existing prediction methods is that they assume independence of individual factors that can have an impact on folding kinetics. For instance, the Leff method is based solely on the effective length of the protein sequence,[21] the CI method is based on composition of the protein chain,[23] and the QRSM method is based on various properties of the constituent amino acids.[24] In contrast, PPFR assumes that folding kinetics depend on a mutual combination of several factors which include chain length and composition, secondary structure, and physicochemical properties of amino acids. Our design shows which of the above factors are complementary with each other, and how to combine them to improve the quality of the predictions. Additionally, this work reveals the importance of the relationships of the strand content and segment count to the folding rates for two-state proteins, and of the coil content to the folding rates. We also optimize the effective length formula originally proposed in[21] and apply a new secondary structure prediction method, i.e., PROTEUS, to derive the features. Finally, we developed and successfully applied a new set of features, which are based on counts and sizes of secondary structure segments predicted with PSIPRED and PROTEUS.

PPFR incorporates three linear regression models, which are developed for each of the mainstream folding dynamics: two-, multi-, and mixed states. The models show several interesting relations which could provide useful insights into the folding mechanisms. Namely, it suggests that the following factors are complementary to each other in the context of the prediction of folding rates: (1) bigger proteins require more time for folding; (2) high helical content and the presence of Phe, Asn, and Gln may accelerate the folding process; (3) the presence of Ile, Val, Thr, and Ser may slow down the folding process; (4) for the two-state proteins increased β-strand content may slow down the folding process; and (5) increased coil content may accelerate the folding process. We note that the above factors in tandem help to improve the prediction quality, i.e., when used separately in[21–24] they provide lower prediction quality when compared to results obtained by hybridizing them in the proposed method.

The developed method provides predictions characterized by strong correlations of over 0.8 with the actual folding rates. Based on both in-sample and out-of-sample tests, the PPFR's predictions are shown to be better then the majority of competing sequence-only and structure-based predictors. We show that PPFR is capable of accurate prediction for nonredundant sequences, i.e., sequences that share low similarity with the sequences used to develop the prediction model. PPFR is also shown to be complementary to the most recent sequence-based QRSM method.[24] While both PPFR and QRSM provide jackknife test predictions with correlation coefficients above 0.8, the worst predicted chains for one of these methods are predicted with much higher quality by the other method, and vice versa.

Acknowledgments

The authors thank Dr. Ma (Shandong Univ. of Technology, China) for the clarifications of the experimental data and results and Dr. Ivankov (Institute of Protein Research, Pushchino) for providing datasets used in this research.

References

1. Zeeb, M.; Balbach, J. Methods 2004, 34, 65.
2. Fabian, H.; Naumann, D. Methods 2004, 34, 28.
3. Zarrine-Afsar, A.; Davidson, A. R. Methods 2004, 34, 41.
4. Maity, H.; Maity, M.; Krishna, M. M.; Mayne, L.; Englander, S. W. Proc Natl Acad Sci USA 2005, 102, 4741.
5. Xiao, H.; Hoerner, J. K.; Eyles, S. J.; Dobo, A.; Voigtman, E.; Melcuk, A. I.; Kaltashov, I. A. Protein Sci 2005, 14, 543.

6. Maxwell, K. L.; Wildes, D.; Zarrine-Afsar, A.; De Los Rios, M.; Brown, A.; Friel, C.; Hedberg, L.; Horng, J.; Bona, D.; Miller, E.; Vallée-Bélisle, A.; Main, E.; Bemporad, F.; Qiu, L.; Teilum, K.; Vu, N.; Edwards, A.; Ruczinski, I.; Poulsen, F.; Kragelund, B.; Michnick, S.; Chiti, F.; Bai, Y.; Hagen, S.; Serrano, L.; Oliveberg, M.; Raleigh, D.; Wittung-Stafshede, P.; Radford, S.; Jackson, S.; Sosnick, T.; Marqusee, S.; Davidson, A.; Plaxco, K. Protein Sci 2005, 14, 602.
7. Jackson, S. E. Des. 1998, 3, R81–R91.
8. Krantz, B. A.; Mayne, L.; Rumbley, J.; Englander, S. W.; Sosnick, T. R. J Mol Biol 2002, 324, 359.
9. Cranz-Mileva, S.; Friel, C. T.; Radford, S. E. Protein Eng Des Sel 2005, 18, 41.
10. Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. Protein Sci 2003, 12, 2057.
11. 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 (Suppl. 1), D279.
12. Plaxco, K. W.; Simons, K. T.; Baker, D. J Mol Biol 1998, 277, 985.
13. Gromiha, M. M.; Selvaraj, S. J Mol Biol. 2001, 310, 27.
14. Zhou, H. Y.; Zhou, Y. Q. Biophysical J. 2002, 82, 458.
15. Debe, D. A.; Goddard, W. A. J Mol Biol. 1999, 294, 619.
16. Munoz, V.; Eaton, W. A. Proc Natl Acad Sci USA 1999, 96, 11311.
17. Dinner, A. R.; Sali, A.; Smith, L. J.; Dobson, C. M.; Karplus, M. Trends Biochem Sci 2000, 25, 331.
18. Gromiha, M. M. J Chem Inf Model. 2005, 45, 494.
19. Gromiha, M. M.; Thangakani, A. M.; Selvaraj, S. Nucleic Acids Res 2006, 34, W70.
20. Gong, H.; Isom, D. G.; Srinivasan, R.; Rose, G. D. J Mol Biol 2003, 327, 1149.
21. Ivankov, D. N.; Finkelstein, A. V. Proc Natl Acad Sci USA 2004, 101, 8942.
22. Huang, J. T.; Tian, J. Proteins 2006, 63, 551.
23. Ma, B. G.; Guo, J. X.; Zhang, H. Y. Proteins 2006, 65, 362.
24. Huang, L. T.; Gromiha, M. M. J Comput Chem 2008, 29, 1675.
25. Gromiha, M.; Selvaraj, S. Current Bioinformatics 2008, 3, 1.
26. Capriotti, E.; Casadio, R. Bioinformatics 2007, 23, 385.
27. Montgomerie, S.; Sundararaj, S.; Gallin, W. J.; Wishart, D. S. BMC Bioinformatics 2006, 7, 301.
28. Jones, D. T. J Mol Biol. 1999, 292, 195.
29. Bryson, K.; McGuffin, L. J.; Marsden, R. L.; Ward, J. J.; Sodhi, J. S.; Jones, D. T. Nucl. Acids Res. 2005, 33(Web Server issue), W36–38.
30. Garg, A.; Kaur, H.; Raghava, G. P. Proteins 2005, 61(2), 318.
31. Song, J.; Burrage, K. BMC Bioinformatics 2006, 7, 425.
32. Chen, K.; Kurgan, L. A. Bioinformatics 2007, 23, 2843.
33. Witten, I.; Frank, E. Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed; Morgan Kaufmann: San Francisco, 2005.
34. Landwehr, N.; Hall, M.; Frank, E. Machine Learning J 2005, 59(1–2), 161.
35. Kurgan, L.; Chen, K. Bioch Bioph Res Comm 2007, 357, 453.
36. Li, R.; Woodward, C. Protein Sci 1999, 8, 1571.
37. Krantz, B. A.; Moran, L. B.; Kentsis, A.; Sosnick, T. R. Nat Struct Biol 2000, 7, 62.
38. Williams, N. K.; Liepinsh, E.; Watt, S. J.; Prosselkov, P.; Matthews, J. M.; Attard, P.; Beck, J. L.; Dixon, N. E.; Otting, G. J Mol Biol. 2005, 346, 1095.
39. Makarov, D. E.; Plaxco, K. W. Protein Sci 2003, 12, 17.
40. Wallin, S.; Chan, H. S. Protein Sci 2005, 14, 1643.
41. Viguera, A. R.; Vega, C.; Serrano, L. Proc Natl Acad Sci USA 2002, 99, 5349.
42. Calloni, G.; Taddei, N.; Plaxco, K. W.; Ramponi, G.; Stefani, M.; Chiti, F. J Mol Biol. 2003, 330, 577.
43. Fersht, A. Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding; Freeman: New York, 1999, 540.
44. Gunasekaran, K.; Eyles, S. J.; Hagler, A. T.; Gierasch, L. M. Curr Opin Struct Biol 2001, 11, 83.