Research Article
Proteins: Structure, Function and Bioinformatics
DOI 10.1002/prot.22727

Full-title:

Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility

Short-title:

Prediction of protein folding rates

Jianzhao Gao[1], Tuo Zhang[1,2], Hua Zhang[3], Shiyi Shen[1], Jishou Ruan[1,4] and Lukasz Kurgan[5 *]

[1] College of Mathematics and LPMC, Nankai University, Tianjin, PRC

[2] Indiana University School of Informatics, Indiana University – Purdue University, Indianapolis, IN, USA

[3] School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, PRC

[4] Chern Institute for Mathematics, Nankai University, Tianjin, PRC

[5] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, Alberta, CANADA

* author to whom the correspondence should be sent;

Department of Electrical and Computer Engineering, 2nd floor, ECERF (9107 116 Street), University of Alberta, Edmonton, AB, CANADA T6G 2V4, Fax: (780) 492-1811, email: lkurgan@ece.ualberta.ca

Keywords: folding rate; flexibility; solvent accessible surface; B-factor; linear regression.

Received: Jan 06, 2010; Revised: Mar 10, 2010; Accepted: Mar 17, 2010

Abstract

Protein folding rates vary by several orders of magnitude and they depend on the topology of the fold and the size and composition of the sequence. Although recent works show that the rates can be predicted from the sequence, allowing for high-throughput annotations, they consider only the sequence and its predicted secondary structure. We propose a novel sequence-based predictor, PFR-AF, which utilizes solvent accessibility and residue flexibility predicted from the sequence, to improve predictions and provide insights into the folding process. The predictor includes three linear regressions for proteins with two-state, multi-state and unknown (mixed-state) folding kinetics. PFR-AF on average outperforms current methods when tested on three datasets. The proposed approach provides high quality predictions in the absence of similarity between the predicted and the training sequences. The PFR-AF's predictions are characterized by high (between 0.71 and 0.95, depending on the dataset) correlation and the lowest (between 0.75 and 0.9) mean absolute errors with respect to the experimental rates, as measured using out-of-sample tests. Our models reveal that for the two-state chains inclusion of solvent exposed Ala may accelerate the folding, while increased content of Ile may reduce the folding speed. We also demonstrate that increased flexibility of coils facilitates faster folding and that proteins with larger content of solvent exposed strands may fold at a slower pace. The increased flexibility of the solvent exposed residues is shown to elongate folding, which also holds, with a lower correlation, for buried residues. Two case studies are included to support our findings.

Introduction

Protein chains fold, from their initial random coil conformation into their functional three-dimensional structure, with rates that vary between several microseconds and an hour[1]. The two main folding kinetics types include two-state folding in which a given protein folds in an "all-or-none" process and multi-state folding where the protein folds with at least one intermediate state. Although these processes are not yet fully understood, the knowledge of folding kinetics finds useful applications. Misfolding, slow folding, and protein aggregation are responsible for some of the amyloid-related and other "conformational" diseases[2]. For instance, the information concerning the folding kinetics was shown to provide mechanistic and structural insight for formation of amyloid fibrils[3]. On the other hand, ultrafast folding proteins are utilized for benchmarking molecular dynamics simulations and testing protein folding theories since they allow for realistic simulations and direct comparison with experimental observations[4]. The folding kinetics and folding rates are experimentally determined using hydrogen exchange, spectroscopic, laser-induced temperature jumps, mass spectrometry and NMR.[5-10], but the corresponding data are being accumulated at a relatively slow rate. The KineticDB[11] and Protein Folding Database (PFD)[12], the two most comprehensive databases for experimental data on protein folding kinetics, include only 90 and 52 entries, respectively, when compared with close to 9 millions of currently known nonredundant protein chains. A viable alternative to experimental methods is to use the experimental data from these databases to build computational models that estimate/predict the corresponding

kinetic information. This work is concerned with building such model to estimate the protein folding rates.

Prior works reveal that the chain length is one of the key determinants of the folding rate for proteins with the three-state folding kinetics. The standard measurement of the folding rates, which is the logarithm of the folding rate measured (or extrapolated) in water, k_f, is strongly anti-correlated with the chain length L [13]. At the same time, the chain length is shown not to be correlated with the folding rate for two-state folders[13]. Prior works show that the magnitude of the correlation is on average, across both two-state and multi-state folders, at about 0.65[2,14,15]. Other factors, such as the topology of the protein fold, were also shown to affect the folding rates[16]. A wide range of topological characteristics of the protein fold was investigated to build structure-based predictors of the folding rate. Plaxco et al.[16] proposed relative contact order (CO), which is defined as an average sequence separation between contacting residues, to estimate the folding rates of the two-state proteins. Subsequent works explored related residue-contact based characteristics including long-range order (LRO) [17,18], absolute contact order (Abs_CO)[19], total contact distance (TCD)[20], which combines LRO and CO, relative contact order[21], geometric contacts[22], elongation-sensitive contact order[23], and multiple contact index (MCI)[24]. Overall, recent works indicate that the knowledge of short-, medium-, and long-range contacts allows for an accurate discrimination of the slow and fast folding proteins[24]. The folding rates were also investigated using other topological features such as protein

compactness, which is defined as a ratio between the accessible surface area and the ideal sphere of the same volume[25]. A recent study has shown that several related structural descriptors, such as radius of gyration, the radius of cross-section, and the coefficient of compactness, can be used to determine the folding rate[2]. Finally, a few approaches proposed to predict the folding rates using information concerning secondary protein structure[26,27], which was computed with DSSP[28].

The above characteristics are either very simple, i.e., based solely on the chain length, or require the knowledge of the three-dimensional structure of the native folds. The large and growing gap between the number of known protein sequences and known protein structures[29] motivates the development of methods that would rely solely on the knowledge of the protein sequence. Last few years observed development of several sequence-based predictors of folding rates. In one of the first attempts, an effective chain length, Leff[1], which combines the chain length with information concerning secondary structure predicted with PSI-PRED[30] and ALB[31], was shown to correlate with the folding rates. More recently, amino acids composition-based index, CI[32], and Ω value[33], which is based on properties of amino acids including their rigidity and propensity for certain secondary structures, were used to build successful predictors. The most recent methods use more advanced sequence characteristics and different prediction algorithms. The SFoldRate method[22] applies linear regression and encodes the input protein sequence using custom-designed index that quantifies propensity of amino acids for formation of

contacts in the protein fold. The QRSM[34] predictor applies a quadratic response surface model based prediction algorithm which utilizes combination of 49 physicochemical, energetic, and conformational properties of amino acids. The PPFR[35] method combines a wide range of sequence characteristics including the length, effective length, physiochemical properties of residues, and secondary structures predicted by PSI-PRED and PROTEUS[36] as an input to a linear regression model to provide improved prediction quality. Similarly as PPFR, the PredPFR[37,38] predictor hybridized several sequence characteristics such as chain length, properties of amino acids, and secondary structure predicted with PSI-PRED to build a linear regression-based model. The last method has a drawback of not being able to predict folding rates for chains that are shorter than 50 amino acids.

While the above sequence-based methods predict the folding rates that are relatively well correlated with the experimental measurements, they do not consider some of the characteristics that are utilized by the structure-based methods. For instance, surface area of the native structure was implicated to impact the folding rates[2] and changes in kinetic and thermal stabilities were shown to results in up to manifold differences in folding rates[39,40]. Inclusion of additional characteristics could further improve the prediction quality and it also could reveal interesting insights into the folding process. To this end, we consider and analyze the relation between the folding rate and the solvent accessible surface, thermal stability and flexibility which are predicted from the protein sequence. Our work is also motivated by a recent result

that indicates that predicted topological characteristics provide useful input[41]. More specifically, folding rates of small single-domain proteins that fold through two-state kinetics were shown to be predictable using sequence-based predictions of residue-residue contacts in proteins of unknown structure. The authors show that estimates based on relatively inaccurate contact predictions are almost as good as the estimates that utilize the known contacts[41]. We propose three linear regression models, which apply a carefully crafted and selected feature sets to predict folding rates for two-state, multi-state, and mixed-state (unknown folding kinetics type) proteins. These features combine information about the sequence and the predicted secondary structure, residue flexibility, and solvent accessibility.

Materials and Methods

Datasets

Three datasets are used in this study, and they include the D62 and D8 datasets from Jiang et al.[35]. The D62 dataset was originally introduced by Ivankov and Finkelstein et al.[1] and it includes 37 two-state and 25 multi-state proteins. The D8 dataset was extracted from the dataset of 77 proteins (denoted by D77) from Huang et al.[34], by removing sequences that share 35% or larger pairwise sequence identity with the sequences in the D62 dataset.

To accommodate for the remaining experimental data that were not included in the D62 and D77 datasets, we also prepared a new dataset based on depositions in the kineticDB[11] database. We downloaded all 90 sequences with the known folding rates

from this database and removed the proteins that are already included in the D62 and D77 datasets. The remaining sequences were filtered to remove redundancy using BLASTCLUST[42] at http://blast.ncbi.nlm.nih.gov/Blast.cgi with local identity threshold set at 25% and default minimal length coverage of 90%. The resulting set includes 24 proteins. Next, we removed the sequences which share 35% or larger pairwise sequence identity with any sequence in the D62 dataset. The final dataset consists of 16 sequences and is referred to as D16.

The D62 dataset is used to build prediction models and to perform their evaluation. Since evaluation on the D62 dataset is somehow obscured by the fact that these data are used in model building, we perform additional tests on the D8 and D16 datasets, which include sequences that are dissimilar to sequences in the D62 dataset. Experimental folding rates in the three datasets are defined by decimal logarithms of protein folding rates in water in the absence of denaturant, i.e., $\log_{10}(k_f)$. The datasets are available for download from http://biomine.ece.ualberta.ca/PFR-AF/PFR-AF.html.

Experiment Setup

We use three types of tests to evaluate our model. The *resubstitution* (self-consistency) test generates and tests the predictive model on the same dataset; in our case we use the D62 dataset. We apply this test for consistency with prior reports [1,16,17,20,26,32,34,35], although we observe that these results could be overfitted. The *jackknife* test, also called leave-one-out test, uses n-1 chains, where n is the number of proteins in a given dataset, to generate the model which is tested on the remaining

protein chain. This is repeated *n* times, each time choosing a different test chain. This test is geared to utilize as much data as possible to generate the model, which is important in our case due to the limited size of the experimental data, while it still assures that the evaluation is performed for unseen samples. The *independent* test involves testing on a dataset that was not used to generate the model. In our case, we train the model on the D62 dataset and test it on the D8 and D16 datasets, respectively.

Following prior works we use the Pearson correlation coefficient (PCC) between the predicted folding rate and the experimental (actual) folding rate to evaluate predictive models. PCC is defined as

$$PCC = \frac{\sum_{i=1}^{n}(f_i - \overline{f})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(f_i - \overline{f})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where f_i is the predicted folding rate, $\overline{f} = \frac{1}{n}\sum_{i=1}^{n} f_i$ is the average of f_i, y_i is the experimental folding rate, and $\overline{y}$ is the average of y_i.

Since PCC measures only the linear correlation, we also compute the mean absolute error (MAE) to quantify the magnitude of the differences between the predictions and the actual values

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | f_i - y_i |$$

Relative Solvent Accessibility, Flexibility and Thermal Stability

We apply relative solvent accessibility (RSA), which is defined as the ratio of

solvent accessible surface area (ASA) of a residue observed in its three dimensional structure to that observed in an extended (Gly-X-Gly or Ala-X-Ala) tripeptide conformation, to predict the folding rates. The inclusion of the RSA values is supported by their strong correlation with key functional properties of proteins and active amino acid sites[43,44] and the finding that the surface area is one of strong determinants for the folding rates[2]. The RSA values were used to categorize residues as buried or solvent exposed. The residue is considered to be buried if its (predicted) RSA < 25%, otherwise, it is assumed to be exposed. This is consistent with prior works on residue solvent accessibility that often indicate 25% as a suitable threshold[45,46]. We computed the RSA normalized using Ala-X-Ala tripeptide as suggested by Ahmad and colleagues [47, 48]. The ASA values were predicted from the sequence using the Real-Spine 3.0 web server[49], which is motivated by high quality of predictions generated by this method[50].

B-factor describes thermal fluctuations of an atom in the protein structure and is usually used to quantify flexibility or mobility of the corresponding residues. Research indicates that high-B-factor regions in protein sequence are characterized by a higher average flexibility[51]. Flexibility of the residues, expressed using B-factor, is strongly correlated with the solvent exposure and thermal stability[50]. The above combined with the observation that thermal stability impacts folding rates[40] supports inclusion of (predicted) B-factor values in the proposed predictive model. The B-factor values were predicted from the protein sequence using PROFbval web

server[52,53].

We also investigate thermal stability of the protein fold as one of the factors that could impact the folding rates. Structural entropy was shown to be linearly related to thermostability and was used to identify residues involved in thermal stabilization in various protein families[54]. This concept was recently utilized to investigate thermal stability and design stable folds based on optimization of local structural entropy (LSE)[55]. We consider LSE values computed from the protein sequences using procedure developed by Bae et al.[55] as one of the inputs for the proposed predictor.

Secondary Structure

We utilize three web servers to predict the secondary structure, PSI-PRED[30] (version 2.6), PROTEUS[36], and SSPRO (version 4.0)[56,57], since secondary structure predictions are shown to be complementary and to work well in consensus [58]. The selection of PSI-PRED was motivated by its use in numerous protein structure prediction methods[59,60,] as well as its prior successful application in prediction of folding rates [1,35,37,38]. PROTEUS was recently shown to provide favorable prediction accuracies when compared with several other secondary structure predictors[36] and was also previously used in prediction of folding rates [35]. SSPRO is part of the SCRATCH web server[57] and this method, together with PSI-PRED, was ranked as one of the top secondary structure prediction servers in the EVA benchmark [61,62].

Feature Design

We use five sources of input data including protein sequence, predicted secondary structure (SS), predicted solvent accessible surface (ASA), predicted B-factor, and local structure entropy (LSE), to encode the inputs for the proposed folding rate predictor. We also combine information concerning predicted secondary structure and solvent exposure, flexibility and solvent exposure, and flexibility and secondary structure. The following features were considered:

- ***L***: length of the protein chain (1 feature)
- **CV_*i***: composition of 20 amino acid types, where i = 1…20, which is defined as the count of amino acids of a given type divided by L. (20 features)
- **CV_*i*_*x***: composition of 20 amino acid types among buried and exposed residues, where x = {buried, exposed} and RSA was predicted using Real-Spine web server (20*2 = 40 features)
- **CV_*y*_*z***: composition of secondary structure y = {h, e, c}, where h is alpha-helix, e is beta-strand, and c is coil, predicted by web server z = {PSI-PRED, PROTEUS, SSPRO} (3*3 = 9 features)
- **CV_*y*_*x*_*z***: composition of secondary structure y predicted by web server z for residues predicted to be of type x, e.g., CV_h_buried_PSI-PRED denotes the composition of helix residues predicted by PSI-PRED which are buried, as predicted by Real-Spine. (3*2*3 = 18 features)
- **Avg_ASA_*y*_*z***: average solvent accessible surface predicted by Real-Spine for residues predicted by web server z to be in secondary structure of type y. We use

ASA, in contrast to RSA, to define these features. The RSA is used to predict exposed/buried residues. (3*3 = 9 features)

- **Avg_Bfactor_sequence**: average B-factors predicted by PROFbval for the entire protein sequence. (1 feature)
- **Avg_Bfactor_*x***: average B-factors predicted by PROFbval for residues predicted by Real-Spine to be of type *x*. (2 features)
- ***w*_Bfactor_*y*_*z***: maximal, minimal and average B-factor values for secondary structure segments of type *y* predicted by web server *z*, where *w* = {min, max, average}. Using the following predicted secondary structure sequence CCCHHHHHHHHHHHCCHHHHHHHHHCCEECC as an example, we first extract secondary structure segments (for coil CCC, CC, CC, CC; for helix HHHHHHHHHHH and HHHHHHHH; for strand EE), and next we compute average B-factors for each of these segments. Finally, among the average values for segments of each type of the secondary structure we find the minimal, maximal and average values. In case when there is no segment of a given type, we set the min, max and average to 0. (3*3*3 = 27features)
- **LSE**: the local structure entropy estimated for the entire protein sequence. We use the procedure by Bae et al.[55]. We downloaded SCOP-35 database of tetra-peptides from http://sdse.life.nctu.edu.tw/index.cgi?xln=download. This database is used to compute LSE as an average of the *L*-3 local structure entropy values for all tetra-peptides in the input protein chain. (1 feature)

Prediction Model

The folding rate prediction was performed using a linear regression predictor

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

where s = {two-state, multi-state, mixed-state} corresponds to the folding dynamics types, x_{sj} is the j^{th} feature for the s^{th} folding dynamics type, k_s is the number of features for the s^{th} folding dynamics type, and w_{sj} is the j^{th} feature's regression coefficient for the s^{th} folding dynamics type. The values of the regression coefficients were estimated from the data using WEKA (version 3.6.0), which is an open-source library of machine learning methods[63]. The linear regression was also used to develop three other recent folding rate prediction methods[32,33,35].

Feature Selection

The set of 128 features was processed using feature selection to reduce the dimensionality. We apply two different feature selection strategies, a filter-based and a wrapper-based[64].

The filter-based approach was implemented using correlation-based feature selection (CFS) method[65]. This method favors features that are highly correlated with the output (folding rate), and uncorrelated with each other. The selection criterion is defined a ratio between a correlation-based estimate of the predictive value of a given feature set and their estimated redundancy. The CFS method was demonstrated to reduce the dimensionality while maintaining or even improving performance of the

subsequent prediction[65]. For efficiency, we used best-first search with forward feature selection to search through the space of the feature sets. This feature selection method was also used to design the PPFR method[35].

The wrapper-based method[66] was implemented by utilizing linear regression models (which are subsequently used to perform folding rate prediction) built on selected subsets of features. Similarly as for the CFS method, we use best-first search with forward feature selection to generate feature sets; this method is denoted as Wrapper-BF. We also considered greedy stepwise search with forward feature selection; this variant is referred to as Wrapper-GS.

The feature selection was performed for each of the three folding dynamics types using jackknife tests on the D62 dataset to avoid overfitting. The filter-based CFS method generates a different set of features for each of the jackknife folds, while the wrapper-based method generates one feature set for the entire jackknife test. As a result total of five feature sets were generated:

1. Only the features selected using CFS in all 62 folds were accepted; this set is denoted by CFS-100%folds
2. The features selected using CFS in at least 50% of the 62 folds were accepted; this set is denoted by CFS-50%folds
3. The features selected using CFS in at least 1 of the 62 folds were accepted. Since the number of such features is relatively large, they were further processed by using a wrapper-based approach to remove redundant/irrelevant features. We start with a feature from this set that has the highest jackknife-based PCC when used

for prediction of folding rates on the D62 dataset and we incrementally add additional features drawn from this set which further increase the correlation. This is repeated until the inclusion of any of the remaining features does not improve correlation. The final feature set is denoted by CFS-Wrapper-1fold.

4. The features selected using Wrapper-BF method.
5. The features selected using Wrapper-GS method.

Each of these five feature sets was further processed by removing irrelevant/redundant features. This was performed by computing PCC of the predictions generated by a linear regression model computed from a given set of features using jackknife test on the D62 dataset. We start with a given feature set and we remove these features that do not result in decrease of the correlation coefficient. Once the final five feature sets are found, we compute correlations for linear regression models using jackknife tests on D62 and independent test on D8, see Table 1. We do not use the D16 dataset to perform feature selection. This dataset is used exclusively to test the final design of the proposed predictor, which allows verifying whether overfitting occurred. In case of the models for two-state and multi-state chains we use the corresponding 37 and 25 chains from the D62 dataset, respectively.

Results in Table 1 agree with prior works that indicate that wrapper-based feature selection usually results in feature sets that perform better in the subsequent prediction[64]. The three wrapper-based feature sets perform similarly well on the jackknife test on the D62 dataset, but only the Wrapper-GS set performs equally well

on the independent test on the D8 dataset. This suggests that this feature set allows for good quality predictions for sequences that share low identity with the sequences used to derive the model. Therefore, the Wrapper-GS feature set, which is shown in Table 2, was selected to implement the proposed folding rate predictor.

The selected feature sets are compact as they include only 5 to 9 features, depending on the target kinetics type. Although the structural entropy-based LSE feature was not retained, the features based on the other two data sources introduced in this work, namely solvent accessibility and B-factor, are included. Although sequence length was selected for all three models, we observe that its correlation with the folding rates is lower for the two-state proteins, which is consistent with prior reports[13]. We note that for two-state proteins the strongest correlations, which are higher than the correlation for the chain length, were obtained for features that are based on predicted solvent accessibility, B-factor and secondary structure. The predicted solvent accessibility is most frequently used, i.e., it appears in 5 out of 9, 2 out of 5 and 3 out of 6 features for the two-state, multi-state and mixed-state models, respectively. At the same time, the predicted B-factor and secondary structure are also used to compute multiple features in each of the models. The secondary structure used in the mixed-state model comes exclusively from the PSI-PRED, while the predictions from SSPRO were not utilized. The only purely sequence-based features that were found useful are the chain length in all three models and composition of Ile in the two-state model. As our feature selection strives to remove redundant and

irrelevant features, we conclude that the information coming from the considered predicted sequence characteristics are complementary to the sequence length. Detailed discussion of the selected features is included in the Results and Discussion section.

Results and Discussion

Factors Governing Folding Rates

We have built three linear-regression models for prediction of folding rates of two-state, multi-state and mixed-state (unknown folding kinetics) proteins, respectively, using the D62 dataset, see Figure 1. The sign of the coefficients indicates whether a given feature is positively or negatively correlated with the experimental folding rate. We caution the reader that the magnitude of coefficients should not be compared between features (although it could be compared for the same features, such as L, in different models), since the feature values are in different ranges. The regression models not only reveal which features (factors) are related to the folding rate, but most importantly they also indicate which of these factors are complementary with each other, i.e., which could be used in tandem to improve predictions. Our analysis concentrates on features that have high absolute correlation coefficients, >0.28 (see Table 2), for each of the three folding kinetics types.

Figure 1 reveals that the protein length L is negatively correlated with the experimental folding rates in the three models. Since the folding rate is the inverse of the actual folding time, this suggests that larger proteins need more time to fold. The

length is a major determinant for the multi-state chains with PCC = -0.8, is also strongly correlated for the mixed-state sequences with PCC = -0.61, but its PCC equals only -0.33 for the two-state proteins, see Table 2. These correlations are consistent with the corresponding coefficients in the three regression models where the largest magnitude is observed for the multi-state model, followed by mixed-state and two-state models. This agrees with results of Galzitskaya et al.[13] which show that length is a weaker determinant for the two-state proteins. The use of the length in the regression model is also consistent with results by Ivankov et al.[1] and Jiang et al.[35].

The CV_e_exposed_psipred and CV_I are negatively correlated with experimental folding rate for the two-state chains, and they also have negative coefficients in the corresponding prediction model. The first correlation translates into an observation that increased content of solvent exposed beta-strands (as predicted by PSI-PRED and Real-Spine) slows down the folding in the two-state proteins. A similar observation that implicates increased beta-strand content was shown in ref[27, 35], but here we show that this concerns solvent exposed structures. The content of the solvent exposed strands has slightly stronger correlation of -0.66 when compared with the correlation for the content of all predicted strands which equals -0.61. Our model also suggests that increased content of Ile (I) may slow down the folding process for the two-state chains. This is consistent with other works[33,35,67], where this relation is explained by the ability of Ile to form geometric contacts and the fact that Ile has branched side chain, which enlarges the number of potential conformations[68,69,70].

The Min_Bfactor_c_segment_proteus and CV_A_exposed are positively correlated with the experimental folding rates for the two-state proteins and have positive coefficients in the associated predictive model. The first feature quantifies predicted flexibility of the most conserved coil segment and it indicates that increased flexibility of coils results in faster folding. The second feature suggests that increased content of exposed Ala also facilitates faster folding of two-state folders. Although the increased content of Ala was recently implicated in faster folding in ref[22], our work demonstrates that a stronger correlation, 0.37 vs. 0.19, concerns the content of the solvent exposed Ala residues. Since free energy changes during folding are dominated by the changes in the conformational entropy, we hypothesize that the above relation could be explained by a relatively low conformational entropy of Ala[71].

Our model also indicates that the Max_Bfactor_e_segment_proteus, which quantifies the maximal predicted B-factor value for predicted strand segments, is negatively correlated with the experimental folding rate for the multi-state proteins. This suggests that increased flexibility of strand segments results in slower folding. Related works[27,35] show that formation of longer strand segments slows down folding of multi-state folders. Our results indicate that the correlation with the folding rates improves from -0.22 to -0.29 when considering flexibility of these segments rather than their size.

The model for the mixed-state proteins reveals that Min_Bfactor_c_segment_psipred, which quantifies minimal predicted B-factor value for the predicted coil segments, is positively correlated with the experimental folding rate. This is consistent with the model for the two-state folders and shows that flexible coils accelerate folding. On the other hand, Avg_Bfactor_exposed and CV_e_exposed_psipred, which correspond to the average predicted B-factor of the exposed residues and the content of the predicted exposed strands, respectively, are negatively correlated with the experimental folding rate. The latter finding is also consistent with the model for the two-state folders and we observe improved correlation, -0.33 vs. -0.31, when considering the content of the solvent exposed and all strand segments, respectively. We observe that the correlation between the average B-factors of the exposed residues that equals -0.37 is stronger than the correlation for the buried residues which is -0.23. The exposed residues are more flexible than the buried residues, i.e., they have higher B-factors, which is expected. We hypothesize that increased flexibility of residues, and in particular surface residues, would enlarge the number of potential conformations which in turn would elongate the folding process.

The selected sequence composition-based features with $|PCC| \geq 0.2$, see Table 2, include CV_I and CV_P_buried that are negatively correlated with the folding rate, and CV_A_exposed and CV_P_exposed that are positively correlated. Recent results, which do not consider the solvent exposure, confirm that Ile (I) is negatively

correlated while Ala (A) is positively correlated[22]. At the same time, Ouyang and Liang et al.[22] show that Pro (P) is negatively correlated when considering only the protein sequence and positively correlated when considering structure-based residue contacts. Our models could help in resolving this conflicting conclusion since they suggest that exposed Pro is positively correlated while buried Pro is negatively correlated with the folding rate.

Comparative Study

Table 3 lists predictions of the proposed method for the Prediction of Folding Rates based on solvent Accessibility and Flexibility (PFR-AF). The predictions are based on the mixed-state proteins model (assuming no prior knowledge of the kinetics type) using resubstitution and jackknife tests on the D62 dataset, and when testing our model on the D8 and D16 datasets. PCC values achieved by PFR-AF equal 0.88, 0.84, 0.85, and 0.71 for the resubstitution, the jackknife and the tests on D8 and D16 datasets, respectively. We compare these results, as well as results using the models for two-state and multi-state proteins on the D62 dataset, with the existing solutions to demonstrate predictive quality of the proposed method. Since some existing methods predict folding rates expressed using natural logarithm, $\ln(k_f)$, while other methods, like the proposed PFR-AF, use logarithm of base 10, the PCC values were always computed using the same base (PCC between the experimental and the predicted rates in the base 10 are equal to the PCC in the natural base), while the MAE values were computed in base 10 after converting between the bases, if necessary.

Following the prior reports we compare the PCC values between the experimental folding rates and the predicted folding rates computed using the resubstitution test on the D62 dataset, see Table 4. We caution the reader that these predictions may overfit the dataset as the prediction model is designed and tested on the same set of proteins. The comparison includes five structure-based predictors CO[16], Abs_CO[19], LRO[17], TCD[20], and SSC[26], and three sequence-based methods Leff[1], CI[32], and PPFR[35]. The results include predictions with the mixed-state model on the entire D62 dataset, and the predictions for the two-state and multi-state proteins from D62 using the corresponding two-state and multi-state models, respectively. We observe that sequence-based methods provide predictions that are overall comparable or better than the predictions of the structure-based methods. This could be explained by the fact that the sequence-based predictors utilize models that combine multiple features, while structure-based methods are usually based on a single descriptor. The proposed PFR-AF method provides favorable correlations for all three models. This is likely since PFR-AF applies a well designed and complementary set of features that describe not only the sequence, but also sequence-derived characteristics like solvent accessibility and flexibility. The lower correlations obtained by the mixed-state model are consistent with results of other sequence-based methods. They indicate that the folding rates associated with proteins that fold in two-state or multi-state kinetics are governed by different factors which, when put together, may to some extent interfere with each other.

Table 5 compares PCC values from the jackknife test on the D62 dataset. We compare the proposed PFR-AF, a structure-based method K-Fold[21], and five sequence-based methods including PredPFR [37,38], SFoldRate[22], QRSM[34], CI[32], and PPFR[35]. The reason to include a structure-based method that was not considered in Table 4 is that the K-Fold, which is a web server based on a linear kernel SVM predictor that utilizes the relative contact order, was designed and tested using cross-validation test. Most importantly, the more stringent jackknife test results (when compared with resubstitution test) demonstrate that this method provides superior predictions, PCC = 0.74, when compared with other structure-based methods from Table 4, i.e., best performing method has PCC equal -0.61. Among the sequence-based methods, the Leff method cannot be tested using jackknife test (since it was developed using the entire D62 dataset), and we added three most recent methods, PredPFR, SFoldRate and QRSM when compared with Table 4. We observe that PFR-AF obtains comparable results for both tests on the D62 dataset. The proposed method provides equivalent or better results for the two-state and multi-state models when compared with the other two methods, CI and PPFR. When considering the mixed-state model, PFR-AF outperforms K-Fold, PredPFR, SFoldRate and CI, provides similar prediction to the predictions of PPFR, and is outperformed only by QRSM. We observe that the jackknife test results for the QRSM shown in Table 5 are based on a larger D77 dataset[34]. Since the D62 dataset is a subset of the D77 dataset, the results in Table 5 are based on jackknife predictions on the D77 dataset which are

constrained to the proteins from the D62 dataset. We compare the two datasets by computing maximal pairwise sequences identity (MPSI) between a given chain and all other chains in the same dataset using the EMBOSS[72,73] server at http://www.ebi.ac.uk/Tools/emboss/align/index.html. This is motivated by the usage of the jackknife test where all but one sequence are used to derive the predictive model, which means that the most similar sequence to the single test sequence could be used to compute the predictions. Figure 2 shows the distribution of the MPSI values for the D77 and D62 datasets. The distributions show that about 47% of sequences in D62 have MPSI values below 20% and no sequence in D62 has MPSI values larger than 80%. In contrast, only about 29% of sequences in the D77 dataset have MPSI value below 20% and 22% have MPSI values that are larger than 80%. This demonstrates that D62 dataset is characterized by lower pairwise sequence identity than the D77 dataset, which could influence the jackknife-based estimate of the PCC values in Table 5.

We also perform tests on the D8 and D16 datasets, see Table 6, which aim at quantifying the predictive performance on chains that are dissimilar to the chains (from the D62 dataset) used to design and compute the predictive model. The relations between the experimental folding rates and the predictions from the PFR-AF, the PPFR which is the second best method on these datasets, and the structure-based K-Fold method are visualized in Figure 3. The scatter plots show that PFR-AF computes folding rates that are positioned closer to the diagonal line which denotes

perfect predictions. The results demonstrate that PFR-AF outperforms the other considered methods, and they demonstrate a similar level of performance for both the jackknife test on D62 and the tests on both independent datasets. This suggests that the proposed predictor is capable of high quality predictions even in the absence of sequence similarity. A relatively high PCC of 0.81 for QRSM on the D8 dataset is likely since these chains were included in the D77 dataset that was used to design this method. The PPFR, K-Fold, PredPFR and SFoldRate are shown to obtain relatively good correlations of about 0.5 to 0.65 on the D16 dataset.

Furthermore, we computed MAE between the experimental and the predicted rates for the proposed PFR-AF, K-Fold, PredPFR, SFoldRate, QRSM and PPFR, see Table 7. The average errors, which were computed for the jackknife tests on D62 and for the tests on the D8 and D16 datasets, complement the PCC values that only reveal the degree of the linear correlation. The natural logarithm based predictions of PredPFR, SFoldRate and QRSM were converted into base 10 to compute the MAE values. The PFR-AF provides predictions with the lowest MAE on all three datasets. The average absolute errors of the proposed methods are about 0.8 to 0.9 in the base 10 logarithm, which translates into estimates that on average differ by less than one order of magnitude from the experimental rates. This should be considered as relatively accurate considering that the $\log_{10}(k_f)$ values in the three datasets range between -3 and 6, which corresponds to 9 orders of magnitude difference. To compare, the errors of the most recent PredPFR method range between 0.9 and 1.3 where MAE

of 1.3 corresponds to an estimate of k_f that is about 2.5 times worse than the corresponding estimate with MAE of 0.9 provided by the PFR-AF.

Finally, we investigate potential complementarity between the proposed PFR-AF and the two recent well-performing methods, PPFR and QRSM, characterized by high PCC and relatively low MAE values when jackknife tested on the D62 dataset. The average MAE of the PFR-AF for ten chains from D62 for which the proposed method makes the largest errors (1RA9, 1GXT, 256B, 1PIN, 1LOP, 1A6N, 1CBI, 1PNJ, 3CHY, and 1FNF90) equals 1.69, while the MAE values of PPFR and QRSM for these chains equal 1.56 and 0.38, respectively. On the other hand, the average MAE the PFR-AF for ten chains for which our model makes the smallest errors (1BRS, 1CSP, 1URN, 1OPA, 1EAL, 1G6P, 1SRL, 2PDD, 1CEI, and 1MJC) equals 0.12, while the MAEs equal 0.58 and 0.59 for the PPFR and QRSM, respectively. Figure 4 shows a detailed comparison of predictions for the D62 dataset. The *x*-axis on Figure 4 corresponds to the sequences in D62 which are sorted in ascending order by MAE values of predictions by PFR-AF. We observe that the maximal MAE of PFR-AF is lower than the MAE for 5 and 14 predictions by PPFR and QRSM, respectively, and that PFR-AF provides the lowest MAE for 26 sequences. At the same time, predictions of PPFR and QRSM are better (have lower MAE) than the prediction of the proposed method for 25 and 21 out of the 62 sequences, respectively, and these two methods provide the lowest MAE for 19 and 17 chains, respectively. The above suggests that although on average PFR-AF produces predictions with the lowest MAE,

the other two methods outperform it on some sequences supporting the claim that the existing and the proposed methods are complementary.

Regression Models Based on Other Protein Properties

We investigate whether usage of other protein properties could lead to regression models with comparable correlations and MAEs. We use the same design procedure as for PFR-AF, but instead of using the features computed from the sequence, predicted secondary structure, solvent accessibility and B-factor, we consider two scenarios, (1) we use each of the three predicted structural properties separately; and (2) we apply the long range order (LRO) values as suggested in ref[41], the 49 physicochemical, energetic, and conformational properties of amino acids based on ref[34], and the combination of these approaches. The LRO values were predicted from the sequence using PROFcon[74] as explained in ref[41]. The 49 properties were taken from http://www.cbrc.jp/~gromiha/fold_rate/property.html[34]. The first scenario allows quantifying the advantage of combining information coming from these three structural properties, while the second one aims at finding whether multivariate regression based on other protein properties could compete with the proposed method. Table 8 compares correlations and errors obtained with regressions that are based on the above six feature sets. Although the jackknife results on the D62 dataset are comparable for the PFR-AF and the regressions based on the predicted secondary structures and the predicted solvent accessibility, only the proposed model performs similarly well on the D8 and D16 datasets.

Case Studies

A short polypeptide motif BBA5 (PDB id: 1T8J)[75], one of the sequences in the D16 dataset, is an extensively studied 23-residues chain that folds at a microsecond timescale. This motif consists of three structural regions, (1) hairpin region (residues 1-8); (2) alpha-helical region (residues 12-23); and (3) a loop region (residues 9-11) which connects the hairpin with the helix[76], see Figure 5A. The structure is stabilized by a hydrophobic core formed between the helix and the hairpin[77]. The rapid folding of this chain is due to the swift formation of the secondary structures[78]. This chain is characterized by a very low maximal pairwise sequence identity of 14.3% to the sequences in the D62 dataset, which were used to derive the proposed prediction model. The prediction error (predicted folding rate minus the experimental rate) generated by PFR-AF equals -0.2. This prediction, which comes from the mixed-state model, benefits from features that use the predicted secondary structure and the predicted solvent exposure. Except for Phe8, Leu14, and Ala15, all residues are solvent exposed. The actual secondary structure computed with DSSP includes 35% of helical residues and 65% of coil residues, and there are no strands. The mixed-state model from Figure 1 predicts the rate as follows (the curly brackets list the corresponding features)

$$\begin{aligned} Rate_{\text{mixed-state}} = \; & -11.1231*0\ \{\text{CV\_P\_buried}\} - 5.9942*0\ \{\text{CV\_e\_exposed\_psipred}\} \\ & - 2.1851*0.218235\ \{\text{Avg\_Bfactor\_exposed}\} - 0.0106*23\ \{\text{L}\} \\ & + 0.6957*{-0.589}\ \{\text{Min\_Bfactor\_h\_segment\_psipred}\} \\ & + 0.6151*0.262727\ \{\text{Min\_Bfactor\_c\_segment\_psipred}\} + 5.888 \end{aligned}$$

$$= -0.4768652985\ \{\text{Avg\_Bfactor\_exposed}\} - 0.2438\ \{\text{L}\} - 0.4097673\ \{\text{Min\_Bfactor\_h\_segment\_psipred}\} + 0.1616033777\ \{\text{Min\_Bfactor\_c\_segment\_psipred}\} + 5.888$$

$$= 4.92$$

We observe that this chain does not have buried Pro (CV_P_buried = 0) and its predicted secondary structure does not include solvent exposed strands (CV_e_exposed_psipred = 0). The Pro4, which is the only proline in this chain, is solvent exposed, see Figure 5A. The predicted rate is influenced by the predicted B-factors of the exposed residues (with value of 0.22 which suggests that they are relatively flexible) and the helical residues (with value of -0.59 which suggests that they are relatively rigid), which lower the predicted value. At the same time, the relatively high predicted flexibility of the coil segment (value of 0.26) adds to the predicted rate. The final result shows that flexibility of the coil shortens the folding time, while the relative rigidity of the helix and the flexibility of the exposed residues elongate the time.

The GW domain of the Internalin B protein[79] (PDB id: 1M9S, residues 391-466), which is named for the conserved Gly-Trp (GW) dipeptide in the C-terminal of this protein, includes 76 residues. This chain is included in the D16 dataset and its maximal pairwise sequence identity to the sequences in the D62 dataset equals 23.5%. The GW domain resembles SH3 domain[79] and includes several beta-strands and a 3/10 helix, see Figure 5B. About 45% residues are buried and 55% are solvent

exposed. The secondary structure assigned with DSSP includes 4% helices, 22% strands and 74% coils. The folding rate of this domain is lower than that of the BBA5 motif, and equals 1.74. The prediction from PFR-AF based on the mixed-state model is computed as

$$\begin{aligned}
\text{Rate}_{\text{mixed-state}} = & -11.1231 * 0.04\ \{\text{CV_P_buried}\} \\
& - 5.9942 * 0.313725\ \{\text{CV_e_exposed_psipred}\} \\
& - 2.1851 * 0.394510\ \{\text{Avg_Bfactor_exposed}\} - 0.0106 * 76\ \{\text{L}\} \\
& + 0.6957 * -0.53\ \{\text{Min_Bfactor_h_segment_psipred}\} \\
& + 0.6151 * 0.2525\ \{\text{Min_Bfactor_c_segment_psipred}\} + 5.888 \\
= & -0.444924\ \{\text{CV_P_buried}\} \\
& - 1.880530395\ \{\text{CV_e_exposed_psipred}\} \\
& - 0.862043801\ \{\text{Avg_Bfactor_exposed}\} - 0.8056\ \{\text{L}\} \\
& - 0.368721\ \{\text{Min_Bfactor_h_segment_psipred}\} \\
& + 0.15531275\ \{\text{Min_Bfactor_c_segment_psipred}\} + 5.888 \\
= & \ 1.68
\end{aligned}$$

Our prediction slightly underestimates the experimental rate by 0.06. The features employed in the model capture essential information about this domain, which results in the accurate estimate. The CV_P_buried quantifies the impact of Pro that is predicted to be buried, but the largest impact on the prediction comes from the relatively high flexibility of the exposed strands (CV_e_exposed_psipred) and the solvent exposed residues (Avg_Bfactor_exposed), and the fact that this is a longer chain with 76 residues (L). We also note the effects of the predicted rigid helix

(Min_Bfactor_h_segment_psipred) and the predicted flexible coil segment (Min_Bfactor_c_segment_psipred), which are similar to what we show for the BBA5 motif case study.

Conclusions

Protein folding is an open problem with many aspects that require research attention. One of such aspects is the timescale, which may vary substantially between proteins. We have built a simple model for prediction of the folding rate given the knowledge of the protein sequence that improves over the existing solutions. Our work is motivated by the premise that certain structural properties that are predicted from the sequence, such as solvent accessibility, secondary structure and residue flexibility, influence the folding rate. We propose three linear-regression based models that address prediction of the rate for proteins with the two-state, the multi-state and unknown (either two or multi-state) folding kinetics. We also analyze these models to reveal potentially interesting relations between certain topological and structural properties of proteins (that are predicted from the sequences) and the folding rates.

The empirical evaluation that involves three datasets and tests on sequences that share low identity with the sequences used to derive the predictive models demonstrate that the proposed prediction method, referred to as PFR-AF, provides favorable prediction quality when compared with modern methods, including sequence- and structure-based methods. This could be explained by the fact that existing sequence-based methods do not apply information concerning flexibility and solvent accessibility, while the structure-based methods usually use only one

topological descriptor, such as residue contacts, and thus they do not benefits from fusing multiple sources of information. The predictions generated by PFR-AF are characterized by high correlation with the experimental rate, between 0.7 and 0.95 depending on the dataset, and the lowest (among the competitors) mean absolute errors, between 0.75 and 0.9, as measured using out-of-sample tests. Two case studies concerning proteins with low sequences identity are used to support our findings.

We observe that although the solvent exposure- and flexibility-based features used by proposed method are characterized by moderate correlations with the folding rates, they complement each other and the other features based on chain length and secondary structure resulting in an accurate prediction method. Analysis of the proposed models reveals several interesting observations: (1) chain length is one of the key determinants of the folding rate, which is consistent with prior works [2,13,14,15]; (2) Inclusion of exposed Ala may accelerate the folding of two-state proteins, which is likely due to the low conformational entropy of this amino acid; (3) Increased content of Ile in two-state proteins may reduce the folding speed due to the ability of this residue to form geometric contacts [67,68,69,70]; (4) Inclusion of buried Pro may decelerate folding; (5) Increased flexibility of coil segments facilitates faster folding; (6) Proteins with larger content of solvent exposed strands may fold at a slower pace; (7) Increased flexibility of strand segments in multi-state proteins may result in slower folding; and (8) Increased flexibility of the solvent exposed residues elongates the folding, which is likely due to an enlarged number of potential conformation, and this relation also holds, with lower correlation, for buried residues. The above factors may

provide useful clues into the protein folding process.

Acknowledgments

LK was supported in part by NSERC Canada. The authors also acknowledge financial support provided by National Education Committee of China. SS and JR were supported by NSFC (grants 20836005 and 10671100), Liuhui Center for Applied Mathematics, and the joint program of Tianjin and Nankai Universities.

References

1. Ivankov DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci U S A 2004; 101: 8942-8944.
2. Ivankov DN, Bogatyreva NS, Lobanov MY, Galzitskaya OV. Coupling between properties of the protein shape and the rate of protein folding. PLoS One 2009; 4: e6476.
3. Fawzi NL, Chubukov V, Clark LA, Brown S, Head-Gordon T. Influence of denatured and intermediate states of folding on protein aggregation. Protein Sci 2005; 14: 993-1003.
4. Dyer RB. Ultrafast and downhill protein folding. Curr Opin Struct Biol 2007; 17:38-47.
5. Zeeb M, Balbach J. Protein folding studied by real-time NMR spectroscopy. Methods 2004; 34:65-74.

6. Fabian H, Naumann D. Methods to study protein folding by stopped-flow FT-IR. Methods 2004; 34:28-40.
7. Zarrine-Afsar A, Davidson AR. The analysis of protein folding kinetic data produced in protein engineering experiments. Methods 2004; 34:41-50.
8. Maity H, Maity M, Krishna MM, Mayne L, Englander SW. Protein folding: the stepwise assembly of foldon units. Proc Natl Acad Sci U S A 2005; 102:4741-4746.
9. Xiao H, Hoerner JK, Eyles SJ, Dobo A, Voigtman E, Mel'cuk AI, Kaltashov IA. Mapping protein energy landscapes with amide hydrogen exchange and mass spectrometry: I. A generalized model for a two-state protein and comparison with experiment. Protein Sci 2005; 14:543-557.
10. Maxwell KL, Wildes D, Zarrine-Afsar A, De Los Rios MA, Brown AG, Friel CT, Hedberg L, Horng JC, Bona D, Miller EJ, Vallée-Bélisle A, Main ER, Bemporad F, Qiu L, Teilum K, Vu ND, Edwards AM, Ruczinski I, Poulsen FM, Kragelund BB, Michnick SW,Chiti F, Bai Y, Hagen SJ, Serrano L, Oliveberg M, Raleigh DP, Wittung-Stafshede P, Radford SE, Jackson SE, Sosnick TR,Marqusee S, Davidson AR, Plaxco KW. Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. Protein Sci 2005; 14:602-616.
11. Bogatyreva NS, Osypov AA, Ivankov DN. KineticDB: a database of protein folding kinetics. Nucleic Acids Res 2009; 37:D342-D346.

12. Fulton KF, Bate MA, Faux NG, Mahmood K, Betts C, Buckle AM. Protein Folding Database (PFD 2.0): an online environment for the International Foldeomics Consortium. Nucleic Acids Res 2007; 35:D304-D307.

13. Galzitskaya OV, Garbuzynskiy SO, Ivankov DN, Finkelstein AV. Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. Proteins 2003; 51:162-166.

14. Gutin AM, Abkevich VI VI, Shakhnovich EI. Chain Length Scaling of Protein Folding Time. Phys Rev Lett 1996; 77:5433-5436.

15. Galzitskaya OV, Ivankov DN, Finkelstein AV. Folding nuclei in proteins. FEBS Lett 2001; 489:113-118.

16. 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.

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

18. Harihar B, Selvaraj S. Refinement of the long-range order parameter in predicting folding rates of two-state proteins. Biopolymers 2009; 91:928-935.

19. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact order revisited: influence of protein size on the folding rate. Protein Sci 2003; 12:2057-2062.

20. Zhou H, Zhou Y. Folding rate prediction using total contact distance. Biophys J 2002; 82:458-463.

21. Capriotti E, Casadio R. K-Fold: a tool for the prediction of the protein folding kinetic order and rate. Bioinformatics 2007; 23:385-386.

22. Ouyang Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci 2008; 17:1256-1263.

23. Hagai T, Levy Y. Folding of elongated proteins: conventional or anomalous? J Am Chem Soc 2008; 130:14253-14262.

24. Gromiha MM. Multiple contact network is a key determinant to protein folding rates. J Chem Inf Model 2009; 49:1130-1135.

25. Galzitskaya OV, Reifsnyder DC, Bogatyreva NS, Ivankov DN, Garbuzynskiy SO. More compact protein globules exhibit slower folding rates. Proteins 2008; 70:329-332.

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

27. 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.

28. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983; 22:2577-2637.

29. Kryshtafovych A, Fidelis K. Protein structure prediction and model quality assessment. Drug Discov Today 2009; 14:386-393.

30. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT. Protein structure prediction servers at University College London. Nucleic Acids Res 2005; 33:W36-W38.

31. Ptitsyn OB, Finkelstein AV. Theory of protein secondary structure and algorithm of its prediction. Biopolymers 1983; 22:15-25.

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

34. Huang LT, Gromiha MM. Analysis and prediction of protein folding rates using quadratic response surface models. J Comput Chem 2008; 29:1675-1683.

35. 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.

36. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 2006; 7:301.

37. Chou KC, Shen HB. FoldRate: A web-server for predicting protein folding rates from primary sequence. Open Bioinformatics J 2009; 3:31-50.

38. Shen HB, Song JN, Chou, KC. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. Journal of Biomedical Science and Engineering 2009; 2:136-143.

39. Gunasekaran K, Eyles SJ, Hagler AT, Gierasch LM. Keeping it in the family: folding studies of related proteins. Curr Opin Struct Biol 2001; 11:83-93.

40. Takahashi N, Onda M, Hayashi K, Yamasaki M, Mita T, Hirose M. Thermostability of refolded ovalbumin and S-ovalbumin. Biosci Biotechnol Biochem. 2005; 69(5):922-31.

41. Punta M, Rost B. Protein folding rates estimated from contact predictions. J Mol Biol 2005; 348:507-12.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-410.

43. Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. Analysis of catalytic residues in enzyme active sites. J Mol Biol 2002; 324:105-121.

44. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004; 20:477-486.

45. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004; 56:753-767.

46. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002; 47:142-153.

47. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003; 50:629-635.

48. Ahmad S, Gromiha MM, Sarai A: Analysis and Prediction of DNAbinding proteins and their binding residues based on Composition, Sequence and Structural Information. Bioinformatics 2004; 20:477-486.

49. Dor O, Zhou Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. Proteins 2007; 68:76-81.

50. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L. On the relation between residue flexibility and local solvent accessibility in proteins. Proteins 2009; 76: 617-636.

51. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK. Protein flexibility and intrinsic disorder. Protein Sci 2004; 13:71-80.

52. Schlessinger A, Yachdav G, Rost B. PROFbval: predict flexible and rigid residues in proteins. Bioinformatics 2006; 22:891-893.

53. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. Proteins 2005; 61:115-126.

54. Chan CH, Liang HK, Hsiao NW, Ko MT, Lyu PC, Hwang JK. Relationship between local structural entropy and protein thermostability. Proteins 2004; 57:684-691.

55. Bae E, Bannen RM, Phillips GN Jr. Bioinformatic method for protein thermal stabilization by structural entropy optimization. Proc Natl Acad Sci U S A 2008; 105: 9594-9597.

56. Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 2002; 47:228-235.

57. Cheng J, Randall AZ, Sweredoski MJ, Baldi P. SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 2005; 33:W72-W76.

58. Albrecht M, Tosatto SC, Lengauer T, Valle G. Simple consensus procedures are effective and sufficient in secondary structure prediction. Protein Eng 2003; 16:459-462.

59. Garg A, Kaur H, Raghava GP. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 2005; 61: 318-324.

60. Song J, Burrage K, Yuan Z, Huber T. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information.BMC Bioinformatics 2006; 7:124.

61. Eyrich VA, Martí-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B. EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 2001; 17:1242-1243.

62. Rost B, Eyrich VA. EVA: large-scale analysis of secondary structure prediction. Proteins 2001; Suppl 5:192-199.

63. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten Ian H. The WEKA Data Mining Software: An Update. SIGKDD Explorations 2009;11:1.

64. Hall M, Smith L. Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. In Proceedings of the FLAIRS conference 1999; 235-239.

65. Hall M. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In Proceedings of the Machine Learning conference (ICML 2000) 2000; 359-366.

66. Kohavi R ; John GH. Wrappers for feature subset selection, Artif Intell 1997; 97:273–324.

67. Ouyang Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci 2008; 17:1256-1263.

68. Dinner AR, Sali A, Smith LJ, Dobson CM, Karplus M. Understanding protein folding via free-energy surfaces from theory and experiment. Trends Biochem Sci 2000; 25:331-339.

69. Makarov DE, Plaxco KW. The topomer search model: A simple, quantitative theory of two-state protein folding kinetics. Protein Sci 2003; 12:17-26.

70. Wallin S, Chan HS. A critical assessment of the topomer search model of protein folding using a continuum explicit-chain model with extensive conformational sampling. Protein Sci 2005; 14:1643-1660.

71. Pickett SD, Sternberg MJ. Empirical scale of side-chain conformational entropy in protein folding. J Mol Biol 1993; 231:825-839.

72. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 2000; 16:276-277.

73. Harte N, Silventoinen V, Quevillon E, Robinson S, Kallio K, Fustero X, Patel P, Jokinen P, Lopez R. European Bioinformatics Institute. Public web-based services from the European Bioinformatics Institute. Nucleic Acids Res 2004; 32:W3-W9.

74. Punta M, Rost B. PROFcon: novel prediction of long-range contacts. Bioinformatics 2005; 21(13):2960-8.

75. Struthers M, Ottesen JJ, Imperiali B. Design and NMR analyses of compact, independently folded BBA motifs. Fold Des 1998; 3:95-103.

76. Mezo AR, Cheng RP, Imperiali B. Oligomerization of uniquely folded mini-protein motifs: development of a homotrimeric betabetaalpha peptide. J Am Chem Soc 2001; 123:3885-3891.

77. Struthers MD, Cheng RP, Imperiali B. Economy in protein design — evolution of a metal-independent ββα motif based on the zinc finger domains. J Am Chem Soc 1996; 118:3073-3081.

78. Snow CD, Nguyen H, Pande VS, Gruebele M. Absolute comparison of simulated and experimental protein-folding dynamics. Nature 2002; 420:102-106.

79. Marino M, Banerjee M, Jonquières R, Cossart P, Ghosh P. GW domains of the Listeria monocytogenes invasion protein InlB are SH3-like and mediate binding to host ligands. EMBO J 2002; 21:5623-5634.

80. DeLano WL. The PyMOL Molecular Graphics System (2002) DeLano Scientific, Palo Alto, CA, USA. http://www.pymol.org

Figure Legends

Figure 1. Prediction models for two-state, multi-state and mixed-state proteins. The variables are grouped by the sign of the regression coefficients and ordered by the magnitude of the coefficients.

Figure 2. Distribution of maximal pairwise sequence identity (MPSI) values for the D66 and D77 datasets.

Figure 3. Scatter plots of predictions generated by the PFR-AF (panel A), PPFR (panel B), and K-Fold (panel C), which are shown on *y*-axis, against the experimental values of folding rates, which are shown on *x*-axis, for the predictions on the D8 and D16 datasets. Linear regressions are shown using solid lines with the corresponding coefficients of determination R^2 (the squared PCC between a given set of predictions and the actual folding rates).

Figure 4. The MAE (*y*-axis) of the PFR-AF, PPFR and QRSM based on the jackknife test on the D62 dataset where sequences (*x*-axis) are sorted in ascending order by MAE values for predictions by PFR-AF.

Figure 5. (A) The structure of the polypeptide motif BBA5 (PDB id: 1T8J). (B) The structure of the GW domain of the Internalin B proteins (PDB id: 1M9S, residues 391-466). The Pro residues, including Pro4 in 1T8J and Pro416 in 1M9S, are shown using spheres. The structures are shown in cartoon representations with a color gradient that represents the B factor values where blue denotes low values and red denotes high values. Since 18TJ was resolved using NMR, we display the B-factors predicted by PROfbval web server. The figure was plotted using Pymol[80].

Table 1. Pearson correlation coefficients between the experimental and the predicted folding rates using five considered feature sets and linear regression models computed using jackknife tests on D62 and independent test on D8. The two-state and multi-state models were computed using 37 two-state and 25 multi-state chains from the D62 dataset, respectively.

Feature set	Test type	Two-state	Multi-state	Mixed-state
CFS-100%folds	Jackknife D62	0.51	0.76	0.74
	Independent D8	not applicable[1]	not applicable[1]	0.62
CFS-50%folds	Jackknife D62	0.73	0.01	0.40
	Independent D8	not applicable[1]	not applicable[1]	0.36
CFS-Wrapper-1fold	Jackknife D62	0.95	0.97	0.85
	Independent D8	not applicable[1]	not applicable[1]	0.75
Wrapper-BF	Jackknife D62	0.95	0.87	0.90
	Independent D8	not applicable[1]	not applicable[1]	0.58
Wrapper-GS	Jackknife D62	0.94	0.87	0.84
	Independent D8	not applicable[1]	not applicable[1]	0.85

[1]The independent test on the D8 dataset concerns only the mixed-state predictions since this set is too small to be further subdivided to perform tests for the two-state and multi-state chains separately.

Table 2. Features selected using the Wrapper-GS method for the two-state, multi-state and mixed-state kinetics together with their Pearson correlation coefficients with the experimental folding rates. The features are order by the decreasing absolute value of their correlation coefficients.

Folding kinetics	Feature	Input data	PCC
Two-state (9 features)	CV_e_exposed_psipred	RSA[1] and SS[2]	-0.66
	Min_Bfactor_c_segment_proteus	B-factor[3] and SS	0.38
	CV_A_exposed	RSA and sequence	0.37
	CV_I	Sequence	-0.33
	L	Sequence	-0.33
	CV_P_buried	RSA and sequence	-0.27
	CV_L_exposed	RSA and sequence	-0.06
	Min_Bfactor_h_segment_psipred	B-factor and SS	0.04
	CV_c_buried_proteus	RSA and SS	-0.01
Multi-state (5 features)	L	Sequence	-0.80
	Max_Bfactor_e_segment_proteus	B-factor and SS	-0.29
	CV_P_exposed	RSA and sequence	0.20
	Max_Bfactor_h_segment_psipred	B-factor and SS	-0.16
	CV_F_exposed	RSA and sequence	-0.13
Mixed-state (6 features)	L	Sequence	-0.61
	Min_Bfactor_c_segment_psipred	B-factor and SS	0.45
	Avg_Bfactor_exposed	RSA and B-factor	-0.37
	CV_e_exposed_psipred	RSA and SS	-0.33
	CV_P_buried	RSA and sequence	-0.18
	Min_Bfactor_h_segment_psipred	B-factor and SS	0.16

[1] RSA: predicted relative solvent accessible surface
[2] SS: predicted secondary structure
[3] B-factor: predicted flexibility of residues

Table 3. Folding rates predicted using the mixed-state model of the proposed PFR-AF method for the resubstitution and jackknife tests on the D62 dataset, and based on the tests on the low sequence identity datasets D8 and D16 when training the models using the D62 dataset.

Dataset	PDB ID	Kinetics type	Experimental folding rate $\log_{10}(k_f)$	Predicted folding rate $\log_{10}(k_f)$	
				Resubstitution	Jackknife
D62	1PIN	two-state	4.1	2.48	2.237
	2PDD	two-state	4.3	4.121	4.106
	2ABD	two-state	2.9	2.671	2.655
	256B	two-state	5.3	3.594	3.432
	1IMQ	two-state	3.2	2.364	2.257
	1LMB	two-state	3.7	3.427	3.378
	1FNF90	two-state	-0.4	0.956	1.09
	1WIT	two-state	0.2	1.353	1.537
	1TEN	two-state	0.5	1.502	1.565
	1SHG	two-state	0.6	1.406	1.51
	1SRL	two-state	1.7	1.543	1.525
	1PNJ	two-state	-0.5	0.875	1.031
	1SHF	two-state	2	1.731	1.703
	1PSF	two-state	1.4	1.89	1.921
	1CSP	two-state	2.9	2.877	2.872
	1C9O	two-state	3.1	2.58	2.541
	1G6P	two-state	2.7	2.599	2.588
	1MJC	two-state	2.3	2.099	2.09
	1LOP	two-state	2.9	1.422	1.335
	1C8C	two-state	3	2.297	2.266
	1HZ6	two-state	1.8	2.262	2.297
	1PGB57	two-state	2.6	2.427	2.416
	1FKB	two-state	0.7	0.405	0.382
	2CI2	two-state	1.7	2.116	2.179
	1AYE	two-state	3	2.612	2.566
	1URN	two-state	2.5	2.559	2.559
	1APS	two-state	-0.7	0.652	0.77
	1RIS	two-state	2.6	2.009	1.973
	1POH	two-state	1.2	1.593	1.615
	1DIV	two-state	2.6	2.264	2.203
	2VIK	two-state	3	1.589	1.527
	1L2Y	two-state	5.4	4.728	4.648
	1VII	two-state	5	4.182	4.084
	1BDD	two-state	5.1	4.077	3.977
	1ENH	two-state	4.6	4.392	4.367
	2ACY	two-state	0.4	0.68	0.704
	1L8W	two-state	0.7	0.08	-0.322
	1A6N	multi-state	0.5	1.963	2.051
	1CEI	multi-state	2.5	2.326	2.297
	2CRO	multi-state	1.6	2.724	2.903
	2A5E	multi-state	1.5	1.891	1.939
	1TIT	multi-state	1.6	1.345	1.311

1HNG	multi-state	0.8	1.64	1.687
1FNF94	multi-state	2.4	1.88	1.853
1IFC	multi-state	1.5	1.163	1.095
1EAL	multi-state	0.6	0.709	0.706
1OPA	multi-state	0.6	0.689	0.689
1CBI	multi-state	-1.4	0.001	0.133
1QOP268	multi-state	-1.1	-0.879	-0.851
1AON	multi-state	0.3	0.862	0.883
1BRS	multi-state	1.5	1.499	1.493
3CHY	multi-state	0.4	1.843	1.908
2RN2	multi-state	0	1.224	1.287
1RA9	multi-state	2	0.196	-0.011
1QOP396	multi-state	-3	-2.064	-1.849
1PHP175	multi-state	1	0.449	0.395
1PHP219	multi-state	-1.5	-0.838	-0.749
1BNI	multi-state	1.1	2.334	2.43
2LZM	multi-state	1.8	1.246	1.195
1UBQ	multi-state	2.6	2.308	2.296
1SCE	multi-state	1.8	2.065	2.075
1GXT	multi-state	1.9	0.113	-0.076
D8			**Predicted rate $\log_{10}(k_f)$ non-redundant dataset**	
1HRC	two-state	3.8	2.324	
1YCC	two-state	4.18	2.533	
1NYF	two-state	1.97	1.732	
1PKS	two-state	-0.46	1.195	
2AIT	two-state	1.8	2.107	
2HQI	two-state	0.08	1.215	
1PBA	two-state	3	3.196	
1HX5	multi-state	0.32	0.825	
D16			**Predicted rate $\log_{10}(k_f)$ non-redundant dataset**	
1BA5	two-state	2.56	4.195	
1E0L	two-state	4.6	3.334	
1FEX	two-state	3.56	4.039	
1GV2	two-state	3.78	4.366	
1JMQ	two-state	3.65	3.033	
1JO8	two-state	1.09	1.399	
1JYG	two-state	3.95	4.026	
1K0S	two-state	3.21	0.872	
1M9S	two-state	1.74	1.683	
1N88	two-state	0.87	2.086	
1PRB	two-state	5.99	4.383	
1RFA	two-state	3.65	3.118	
1SPR	two-state	3.78	2.14	
1T8J	two-state	5.12	4.92	
1U5P	two-state	4.78	4.426	
2A3D	two-state	5.3	4.999	

Table 4. Comparison of PCC values between the experimental folding rates and the rates predicted by the proposed PFR-AF method, five structure-based methods including CO, Abs_CO, LRO, TCD and SSC, and three sequence-based methods including Leff, CI and PPFR using the resubstitution test on the D62 dataset. Best results are shown in bold.

Folding kinetics	structure-based methods					sequence-based methods			
	CO[a]	Abs_CO[a]	LRO[a]	TCD[a]	SSC[a]	Leff[a]	CI[b]	PPFR[b]	PFR-AF
Two-state	-0.57	-0.64	-0.79	-0.79	0.64	-0.61	0.73	0.92	**0.97**
Multi-state	0.43	-0.44	-0.34	0.23	-0.01	-0.77	0.70	0.92	**0.93**
Mixed-state	0.12	-0.57	-0.61	-0.19	0.42	-0.73	0.72	0.85	**0.88**

[a] Results from ref. 32
[b] Results from ref. 35

Table 5. Comparison of PCC values between the experimental folding rates and the rates predicted by the proposed PFR-AF method, a structure-based method K-Fold, and five sequence-based methods including PredPFR, SFoldRate, QRSM, CI, and PPFR using the jackknife test on the D62 dataset. Best results are shown in bold and "n/a" indicates that a given method does not offer a separate model for prediction of two-state or multi-state chains.

Folding kinetics	K-Fold[a]	PredPFR[b]	SFoldRate[c]	QRSM[d]	CI[e]	PPFR[d]	PFR-AF
Two-state	n/a	n/a	n/a	n/a	0.73	0.87	**0.94**
Multi-state	n/a	n/a	n/a	n/a	0.70	**0.87**	**0.87**
Mixed-state	0.74	0.72	0.27	**0.89**	0.73	0.82	0.84

[a] Results from the K-Fold web server at http://gpcr2.biocomp.unibo.it/cgi/predictors/K-Fold/K-Fold.cgi
[b] Results from the PredPFR web server at http://www.csbio.sjtu.edu.cn/bioinf/FoldingRate/; four sequences were too short to be predicted (<50 amino acids) and were excluded from evaluation
[c] Results from the SFoldRate web server at http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html
[d] Results from ref. 35
[e] Results form ref. 32

Table 6. Comparison of PCC values between the experimental folding rates and the rates predicted by the proposed PFR-AF method, a structure-based method K-Fold, and four sequence-based methods including PredPFR, SFoldRate, QRSM, and PPFR when testing on the D8 and D16 datasets. The PFR-AF method was designed on the D62 dataset, while the predictions of other methods are based on the corresponding web servers. Best results are shown in bold.

Dataset	K-Fold[a]	PredPFR[b]	SFoldRate[c]	QRSM	PPFR	PFR-AF
D8	0.14	0.31	0.03	0.81[d]	0.76[f]	**0.85**
D16	0.46	0.48	0.50	-0.38[e]	0.65	**0.71**

[a] Results from the K-Fold web server at http://gpcr2.biocomp.unibo.it/cgi/predictors/K-Fold/K-Fold.cgi
[b] Results from the PredPFR web server at http://www.csbio.sjtu.edu.cn/bioinf/FoldingRate/; six sequences in D16 were too short to be predicted (<50 amino acids) and were excluded from evaluation
[c] Results from the SFoldRate web server at http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html
[d] Jackknife results from ref. 34, where chains from the D8 dataset were included in the D77 dataset used in the jackknife test
[e] Results from the QRSM web server at http://bioinformatics.myweb.hinet.net/foldrate.htm
[f] Results from ref. 35

Table 7. Comparison of MAE values between the experimental folding rates and the rates predicted by the proposed PFR-AF method, a structure-based method K-Fold, and four sequence-based methods including PredPFR, SFoldRate, QRSM, and PPFR when testing on the D62, D8 and D16 datasets. Best results are shown in bold. Predictions were converted into $\log_{10}(k_f)$, if necessary, and compared against the experimental rate in the same base.

Test method	K-Fold[a]	PredPFR[b]	SFoldRate[c]	QRSM	PPFR	PFR-AF
Jackknife test on D62	0.95	0.91	2.71	1.07[d]	0.93[g]	**0.75**
Independent test on D8	1.38	1.32	2.95	1.12[e]	1.18[g]	**0.89**
Independent test on D16	1.35	1.29	2.28	4.00[f]	1.31	**0.83**

[a] Results from the K-Fold web server at http://gpcr2.biocomp.unibo.it/cgi/predictors/K-Fold/K-Fold.cgi
[b] Results from the PredPFR web server at http://www.csbio.sjtu.edu.cn/bioinf/FoldingRate/; four sequences in D62 and six sequences in D16 were too short to be predicted (<50 amino acids) and were excluded from evaluation
[c] Results from the SFoldRate web server at http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html
[d] Results from ref. 34
[e] Jackknife results from ref. 34, where chains from the D8 dataset were included in the D77 dataset used in the Jackknife test
[f] Results from the QRSM web server http://bioinformatics.myweb.hinet.net/foldrate.htm
[g] Results from ref. 35

Table 8. Comparison of PCC and MAE values between the experimental folding rates and the rates predicted by the proposed PFR-AF method, and results obtained using multivariate regressions based on features generated from predicted secondary structure (SS), solvent accessibility (SA), B-factor (Bf), long range order (LRO), physicochemical, energetic, and conformational properties (PECP), and combination of the long range order and the physicochemical, energetic, and conformational properties (LRO+ PECP). The second column lists features used in each regression model where L is the sequence length, i = 1,…,20 is the amino acid type, k = 1,…,49 denotes the physicochemical, energetic, and conformational property type, x = {buried, exposed}, y = {h,e,c}, and z = {PSI-PRED, PROTEUS, SSPRO}. The last six columns on the right show the PCC and MAE values computed using jackknife test on D62 dataset, and results obtained on the D8 and D16 dataset when using models trained on the D62 dataset. Best results are shown in bold.

Inputs	**Considered features**[1]	**Number of features**		**PCC**			**MAE**		
		all	**selected**	**D62**	**D8**	**D16**	**D62**	**D8**	**D16**
SS	*L*, CV_*i*, CV_*y*_*z*, CV_*i*_*y*_*z*	210	10	**0.87**	0.47	0.22	**0.68**	1.33	2.02
SA	*L*, CV_*i*, CV_*i*_*x*, Avg_ASA_*i*	81	10	0.83	-0.46	0.52	0.76	2.52	1.64
Bf	*L*, CV_*i*, Avg_Bfactor_sequence, Avg_Bfactor_*i*	42	10	0.79	0.83	0.26	0.82	1.07	1.76
LRO	*L*, CV_*i*, Avg_LRO_sequence, Avg_LRO_*i*	42	5	0.66	0.18	0.47	1.15	1.35	1.71
PECP	*L*, CV_*i*, index_*k*	70	4	0.67	0.29	0.50	1.13	1.40	1.74
LRO +PECP	*L*, CV_*i*, Avg_LRO_sequence, Avg_LRO_*i*, index_*k*	91	4[2]	0.67	0.29	0.50	1.13	1.40	1.74
PFR-AF	see ”Feature Design” section	128	6	0.84	**0.85**	**0.71**	0.75	**0.89**	**0.83**

[1]see the “Feature Design” section for the explanation of the acronyms
[2]the same features were selected from both PECP and PECP+LRO feature sets

Prediction model for two-state kinetics

$$Rate_{two\text{-}state} = 23.7625*CV\_I - 12.7472*CV\_L\_exposed - 10.3237*CV\_e\_exposed\_psipred - 8.1658*CV\_P\_buried - 0.0096*L + 6.5696*CV\_A\_exposed + 3.4494*CV\_c\_buried\_psipred + 1.9884*Min\_Bfactor\_h\_segment\_psipred + 0.4501*Min\_Bfactor\_c\_segment\_proteus + 6.6537$$

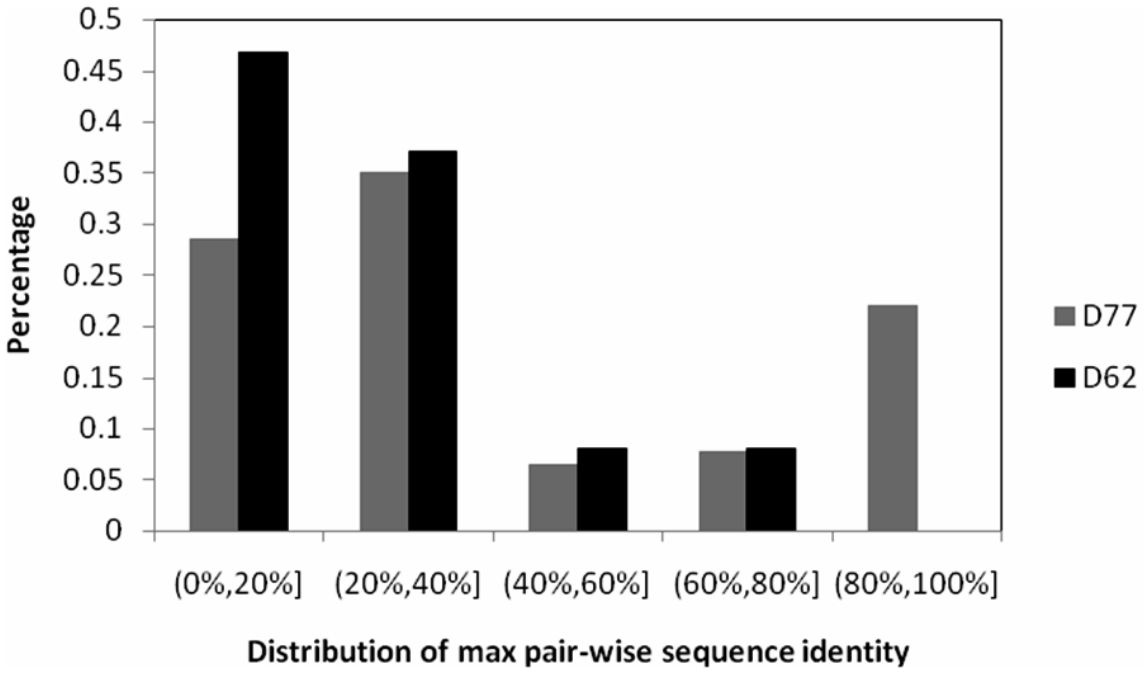
Prediction model for multi-state kinetics

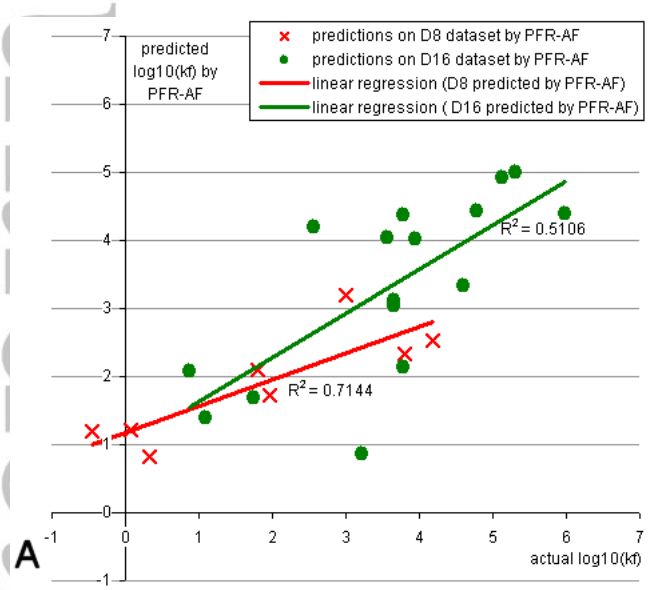
$$Rate_{multi\text{-}state} = -0.7375*Max\_Bfactor\_e\_segment\_proteus - 0.0186*L + 29.2878*CV\_F\_exposed + 13.932*CV\_P\_exposed + 0.5949*Max\_Bfactor\_h\_segment\_psipred + 2.4637$$

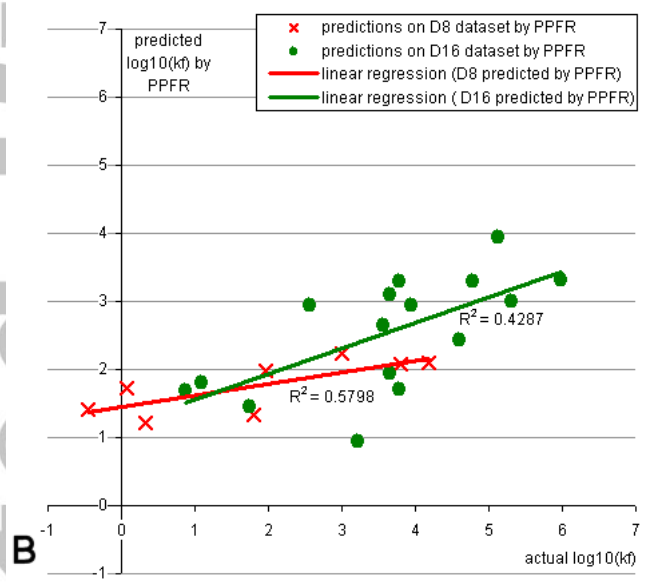
Prediction model for mixed-state kinetics

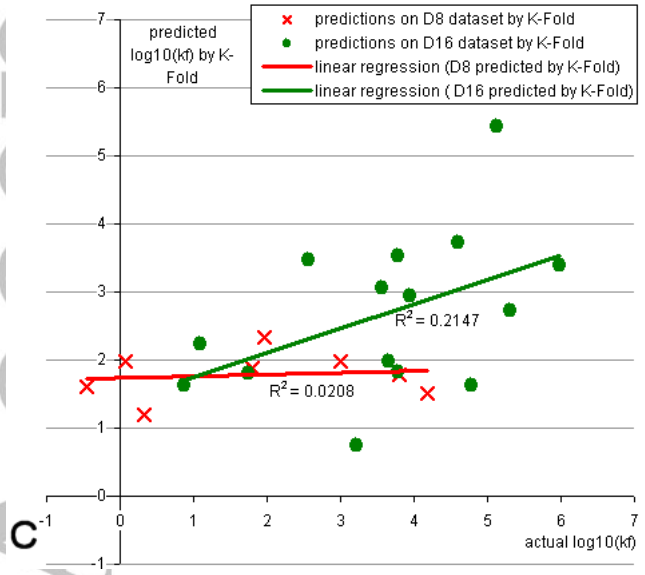
$$Rate_{mixed\text{-}state} = -11.1231*CV\_P\_buried - 5.9942*CV\_e\_exposed\_psipred - 2.1851*Avg\_Bfactor\_exposed - 0.0106*L + 0.6957*Min\_Bfactor\_h\_segment\_psipred + 0.6151*Min\_Bfactor\_c\_segment\_psipred + 5.888$$

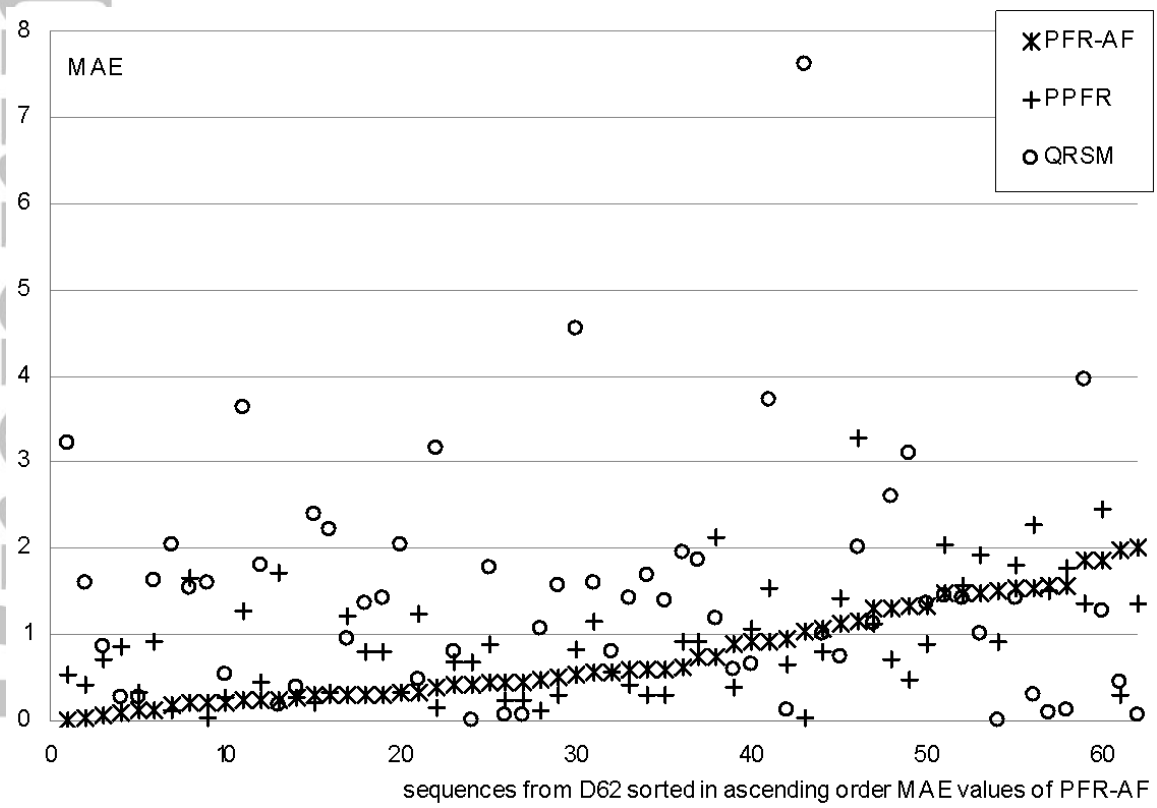
Figure 1. Prediction models for two-state, multi-state and mixed-state proteins. The variables are grouped by the sign of the regression coefficients and ordered by the magnitude of the coefficients.
460x197mm (600 x 600 DPI)

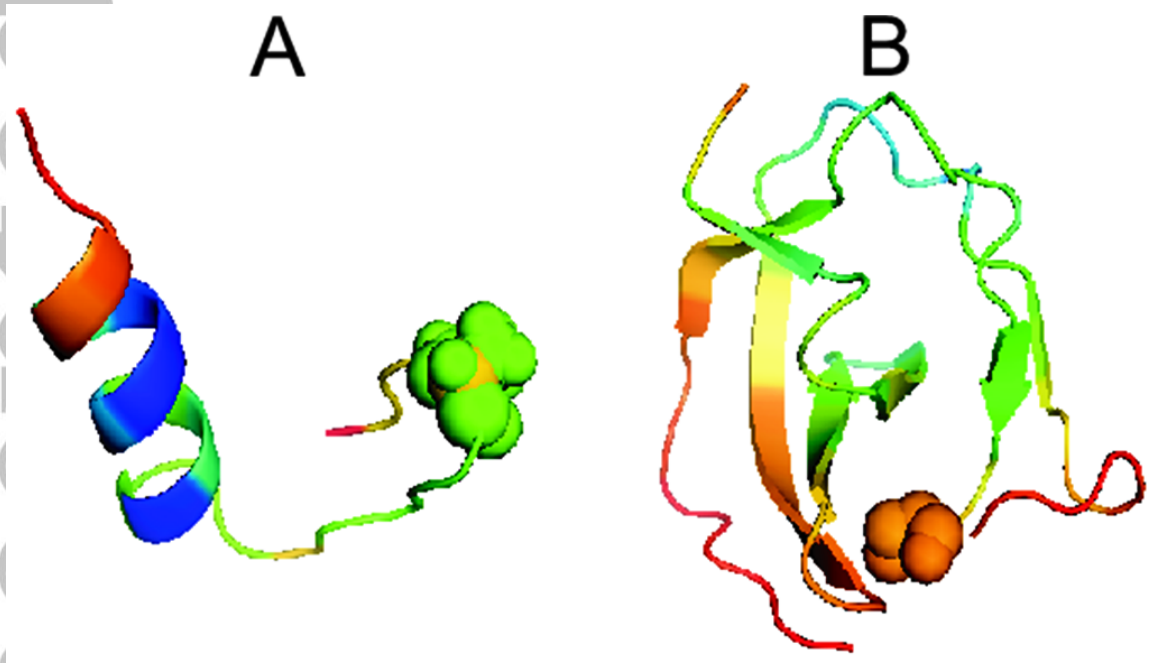



predicted log10(kf) by PFR-AF
predictions on D8 dataset by PFR-AF
predictions on D16 dataset by PFR-AF
linear regression (D8 predicted by PFR-AF)
linear regression (D16 predicted by PFR-AF)
R2 = 0.5106
R2 = 0.7144
actual log10(kf)
A


predicted log10(kf) by PPFR
predictions on D8 dataset by PPFR
predictions on D16 dataset by PPFR
linear regression (D8 predicted by PPFR)
linear regression (D16 predicted by PPFR)
R2 = 0.4287
R2 = 0.5798
actual log10(kf)
B


predicted log10(kf) by K-Fold
predictions on D8 dataset by K-Fold
predictions on D16 dataset by K-Fold
linear regression (D8 predicted by K-Fold)
linear regression (D16 predicted by K-Fold)
R2 = 0.2147
R2 = 0.0208
actual log10(kf)
C


MAE
PFR-AF
PPFR
QRSM
0
1
2
3
4
5
6
7
8
0
10
20
30
40
50
60
sequences from D62 sorted in ascending order MAE values of PFR-AF


A
B