Sequence-Based Prediction of Protein Folding Rates Using Contacts, Secondary Structures and Support Vector Machines

Guan Ning Lin[1], Zheng Wang[2], Dong Xu[1, 2] and Jianlin Cheng[1, 2 *]
[1]*Informatics Institute, University of Missouri, Columbia, Missouri*
[2]*Computer Science Department, University of Missouri, Columbia, Missouri*
[*]*Corresponding author: chengji@missouri.edu*

Abstract

Predicting protein folding rate is useful for understanding protein folding process and guiding protein design. Here we developed a method, SeqRate, to predict both protein folding kinetic type (two-state versus multi-state) and real-value folding rate using features extracted from only protein sequence with support vector machines. On a standard benchmark dataset, the accuracy of folding kinetic type classification is 80%. The Pearson correlation coefficient and the mean absolute difference between predicted and experimental folding rates (sec^{-1}) in the base-10 logarithmic scale are 0.81 and 0.79 for two-state protein folders, and 0.80 and 0.68 for three-state protein folders. SeqRate is the first sequence-based method for protein folding type classification and its accuracy of fold rate prediction is improved over previous sequence-based methods. Both the web server and software of predicting folding rate are publicly available at http://casp.rnet.missouri.edu/fold_rate/index.html.

1. Introduction

Protein folding is one of the most important problems in molecular biology. Two main aspects of the folding process concern the kinetic order and the rate constant. The kinetic order of the protein folding indicates whether the sequence reaches its native structure through intermediate states or not. The folding rate is inversely proportional to the time that the protein needs to collapse into its stable tertiary structure. Proteins have very different rates of folding. Some of them fold within microseconds; some need an hour to fold. Small proteins often (but far from always) fold faster than the larger ones [1]. Many studies have been conducted to estimate protein folding rates based on either experimental protein structural information [2-4] or protein homology sequence searches using databases [5]. However, since only limited amount of experimental folding rates is available for database search and most proteins do not have solved experimental structures, prediction of folding rates based on sequence only is increasing important.

Various theories and simulations suggest a surprising simple linear relation between the number of residues in a protein, its length L, and the rate at which it folds. It is in the form of $\log(k_f) \propto C_1 L^{C_2}$, where k_f is the experimental folding rate, L is the length of the protein, and C_1 and C_2 are simple constants [1]. The correlation between folding rates and protein sizes is stronger for multi-state proteins that have folding intermediates, and weaker for two-state proteins that do not have such intermediates [1]. It implies that protein length does not describe the transition rates of direct folding well.

Baker and coworkers [6] found a strong correlation between the native topological complexity, defined by the parameter contact order (CO), which uses the information about the average sequence separation of all contacting residues in the native state of two-state proteins, and the folding rates of 12 two-state proteins. The correlation between protein-folding rates and their hierarchical structures (secondary structure and structural topology) suggests that hierarchical information could be one of the key features for determining folding rate. Although folding rates of proteins of both two kinetic pathways (i.e. two-state and multi-state folding) can be roughly predicted from the protein secondary structures [7], the prediction scheme should be adjusted to accommodate the differentiation of the two kinetic pathways to improve the accuracy [8].

In the past, various approaches have been designed to estimate the logarithm of the two-state folding rate starting from using structural information. Several

methods based on correlation between the logarithm of the folding rate and structural predictors such as Contact Order (CO) [6], Long-range Contact Order (LRCO) [9] (contact between two residues that are close in space and far in the sequence), total contact distance [10], effective length of folding chain [7] have been developed. These methods require the tertiary structure topologies of a protein as input to predict its folding rate. Since the vast majority of proteins' tertiary structures are still not solved, it is important to design methods that can predict folding rate from protein sequence directly. Toward this goal, in the seminal work [11], Punta and Rost first showed LRCO had better correlation with folding rates than CO. Then they used LRO values predicted from protein sequences for folding rate predictions and achieved 0.61 correlation between the predicted and true folding rates for a set of two-state folding proteins.

Most of folding rate prediction methods are knowledge-based approaches that build a function to map input predictors (e.g. contact order) to folding rates. Traditionally these methods used only a single estimator, either CO, LRCO, or chain length to design linear models between these predictors and protein folding rates. Recently Huang et al. showed that the linear combination of several predictors, such as amino acid rigidity (R), composition vectors (CV), chain length (L), amino acid weight (W), degeneracy (D), and composition index (CI) can increase the correlation between predicted and actual two-state folding rates [8], although the relationship between some of these predictors and the folding rate may not be linear.

Besides folding rate prediction, some studies also have been done to classify the proteins into different folding classes based on their secondary structures. Some classified folders into all-α-class, all-β-class and α/β-class [12], and some even classified folders into 83 different classes. Interestingly, not much has been done to classify the proteins folders based on their binary folding kinetic mechanisms, such as two-state folders or multi-state folders.

A few applications and web servers have been developed for protein rate predictions, such as FOLD-RATE [13], and PPT-DB [5], but only one server K-Fold [4] for fold kinetic classification. K-Fold needs an experimental structure as input to predict fold kinetic type.

Here we developed a non-linear machine learning method (Support Vector Machine classification and regression) that can not only classify proteins into two-state or multi-state folders, but also predict folding rates, using only the information extracted from the amino acid sequence of a protein, without any explicit knowledge of the experimental tertiary or secondary structures. We used a large set of features including protein sequence length, predicted LRCO, predicted long-range contact number (LRCN), predicted α-helical content and β-sheet content and amino acid composition with non-linear SVM models for both protein binary kinetic classification and folding rates prediction. Some features such as secondary structure composition and amino acid composition are new. And our method of deriving LRCO and LRCN are based on predicted residue-residue contact probabilities instead of binary-value contacts used by previous work [11]. Our method performs favorably when compared to other sequence-based methods. We also developed a web server with name 'SeqRate' for the method at our site: http://casp.rnet.missouri.edu/fold_rate/index.html.

2. Material and methods

2.1. Data sets

We used the data set composed by Ivankov in 2004 and also used in [11] that contains experimentally determined 24 multi-states folders and 37 two-state folders, and is referred to as "IvankovData". The folding rate is in the unit of sec^{-1} and transformed in the base-10 logarithmic scale. Sequence and structural information of these proteins is obtained from the Protein Data Bank (PDB).

2.2. Methods

Our method for protein folding rates was developed based on an SVM. In this study we divide our protein rate prediction into two steps. First we use SVM classifier to classify folding types based on binary kinetic mechanism (two-state or multi-state), instead of using structural classes of all-α-class, all-β-class and α/β-class. The second step of protein rate prediction is developing two separate SVM regression prediction models for two-state folders and multi-state folders, considering different folding behaviors between these two types. In this study, multiple input features derived from protein sequences were used in protein folding type classification and folding rate prediction. We also studied the impacts of using different input features, such as protein chain length and several protein topology features, on folding kinetic classification and rate prediction for two-state and multi-state folders.

2.3. Input features

Features, such as protein sequence length, long-range contact order, long-range contact number, α-helical content, β-sheet content and amino acid compositions, used in SVM training models, are defined and discussed as follows.

2.3.1. Protein sequence length. Protein sequence length has been shown to be a relevant factor for evaluating protein folding rates [7, 14], although it is insufficient to just use sequence length to determine the folding type. Smaller sequences usually tend to fold with simpler folding mechanism without any intermediate state like in multi-state proteins.

2.3.2. Contact order (LRCO) and contact number (LRCN). LRCOs and LRCNs used in this study were both calculated based on contact map generated from the SCRATCH suite [3] using protein sequences as inputs. A protein contact map, a two-dimensional matrix, represents the distance information between every two residues' C-alpha atoms of a three-dimensional protein structure. SCRATCH was used to predict the contact probability matrix P for the probabilities of any pair of residues contacting with each other, i.e. the likelihood that their distance is below a threshold. The distance threshold used here is 8 Å and the sequence separation is at least 12 amino acids apart. An element P_{ij} in the matrix is the predicted probability that residues i and j are in contact. As in [9], only long range contacts (i.e. sequence separation of $|i\text{-}j| \geq 12$) were used to derive contact order and contact number.

The LRCN is defined as the expected number of long-rage contacts in a protein. So far, most methods first derive a binary contact map from a probability contact map according to a probability threshold and then count the numbers of contacts [11]. Here, we introduce a modified method to directly calculate contact number from contact probability map and it is further normalized by the power of sequence length. Then the contact number is defined as following

$$\text{LRCN} = \frac{\sum_{|i-j|\geq 12} P_{ij}}{L^c} \quad (1)$$

where P_{ij} is the contact probability of residue i and j, which should be no more then 8 Å away and at least 12 sequence separation apart; L (sequence length) to the power of c is used to normalize contact number. c is set to 1 as in [11].

Different from LRCO (Long-range Contact Order) calculation based on binary contacts in [11], we calculated contact order from contact probabilities as following

$$\text{LRCO} = \frac{\sum_{|i-j|\geq 12} (P_{ij} * |i-j|)}{L^c} \quad (2)$$

where P_{ij} is the probability of residues i and j within 8 Å when at least 12 sequence separation apart; L (sequence length) to the power of c is used to normalize contact order. Just as the calculation in LRCN, probabilistic real values of contacts are used in the formula. c is set to 2 as in [11].

2.3.3. Secondary structure composition. Rose and collaborators [15] observed that folding rates correlate well with the overall secondary structure composition in three states (helix, strand, coil) assigned from 3D coordinates. So we used the predicted percentages of helix, sheet and coil contents of a protein as additional inputs for folding rate prediction. Secondary structures were predicted by SCRATCH [3].

2.3.4. Amino acid composition. Amino acid composition has been shown to be relevant to protein folding types and a good indicator for folding type identification [16]. The basic assumption is that if certain amino acids are optimal for protein structure, natural selection should have acted over evolutionary time to increase the frequency of these amino acids. Therefore, proteins with different amino acid composition would have different folding rates and folding types. In 2007, Ma and his colleagues demonstrated some of contents of amino acids differed between two-state and multi-state folders in a significant level of $p<0.01$ [17]. Here we use the each amino acid occurrence frequency in the protein sequence as amino acid composition. Then, each of 20 amino acid compositions is used as one input feature for SVM.

3. Results and discussion

3.1. Effectiveness of each feature in folding rate prediction

In order to test the effectiveness of each individual feature, we used each feature as input to predict folding rate separately through SVM regression. The default parameter settings were used for SVM regression since comparisons were within feature set. Two different measures were applied to evaluate the performance of the results. One is Pearson correlation coefficient between predicted rates and experimental rates. The other measure is mean absolute difference (MAD), which measures how much predicted values deviate from real values. The correlation coefficient and MAD

were calculated for two-state and multi-state proteins separately. Each feature-specific SVM prediction model was trained using leave-one-out procedure and used to predict the folding rate on the left-out protein. **Table 1** demonstrates the general trend of understanding, which is protein sequence length has more than two times higher correlation values with multi-state folders than two-state folders, and protein topologies (e.g. secondary structure information) have almost twice correlation values with two-state folders as with multi-state folders. These strongly kinetically biased features support the need of separate prediction models for different folding kinetic.

Table 1. Correlation between predicted folding rates and experimental folding rates using sequence length and other estimated predictors on IvankovData training data.
L = protein sequence length, LRCO = estimated long-range contact order, CO = estimated contact order in [15], LRCN = estimated long-range contact number.
First line of values is from two-state folding rates and second line of values is from multi-state folding rates.

L	LRCO	CO	LRCN	α-helical content	β-sheet content	Coil content
-0.32	0.72	0.61	0.68	-0.51	0.57	0.13
-0.80	0.46	0.33	0.55	-0.18	0.11	0.05

LRCO yields the best performance with correlation 0.72 for two-state proteins while protein sequence length demonstrates the best negative correlation of 0.8 for multi-state proteins. For both two-state and multi-state folders, LRCO was preferred over CO since it has higher correlations in both folding kinetics. On multi-state proteins contact number performs the second best with correlation 0.55. Note that the correlation using estimated LRCO on two-state proteins is 0.72, higher than CO has, which is 0.61 reported in [11] on the same data set, indicating that LRCO calculated from contact probability map in our method may be more informative than that derived from binary contact map used in [11].

Coil content has low correlations, 0.13 and 0.05, with both two-state folders and multi-state folders respectively; therefore it is not used in building either folding rate prediction model. Also α-helical content and β-sheet content have low correlation values of -0.18 and 0.11, respectively in multi-state folders, therefore both features are not included for the multi-state folding rate prediction model. Actually by including α-helical content and β-sheet content as features, the prediction results have shown no changes, neither increasing nor decreasing accuracies.

One feature needed to be mentioned here and is not shown on Table 1 is amino acid composition, which is a set of 20 amino acid frequency values. It has shown to be a relevant feature for deciding folding kinetic [16]. It was included as one of our classification features, but it has shown weak correlations with folding rates of both folding kinetic orders in our results. Our tests have indicated the overall correlations of amino acid compositions with the folding rates are only around 0.3. Therefore, this feature is not used for SVM regression rate prediction model in order to avoid over-fitting.

3.2. Sequence-based folding kinetic type classification

Protein sequence length and protein topologies are both favorable folding rate determination factors for two folding types. Protein sequence length is a good predictor in multi-state folder rate prediction, but not in two-state folders. And protein topologies have better correlations with two-state folding rates than multi-state folding rates. We built an SVM classification model based on sequence length, estimated LRCO and CN, α-helical content, β-sheet content and 20 frequency values of amino acid compositions. As in other multivariate statistical models, the performances of the SVM for classification depend on the combination of several parameters. In general, the SVM classification involves two classes of parameters: the parameter C for the tradeoff between training error and margin size and kernel function parameters such as inverse of variance γ for Gaussian kernel. To maximize the performance, we performed the parameter optimization using a grid search approach within a limited range. The classification model is trained and tested using leave-one-out cross-validation (LOOCV). **Figure 1** shows the profile of classification accuracy vs. the variations of parameters C and γ. The prediction accuracy profile peaked at (C, γ) = (1, 0.25). The best classification accuracy of using Gaussian kernel function is 80%, which is higher than any of other classifiers in the literature.

We have used other kernel functions, namely linear, sigmoid and polynomial functions on SVM model for the same data set. The accuracy of those three kernels were 62%, 50% and 72%, respectively.

3.3. Sequence-based fold rate prediction using multiple features and non-linear SVM regression

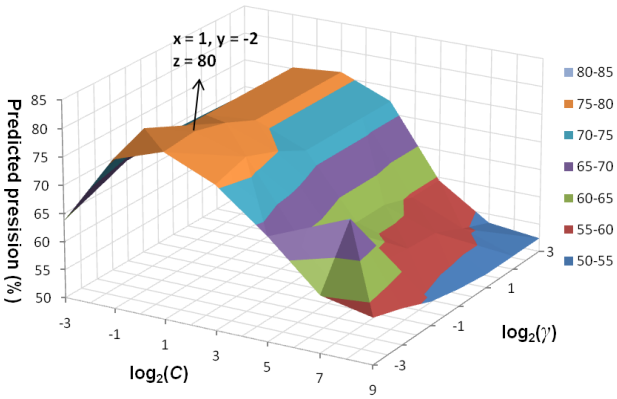

Figure 1. Classification accuracy surface VS. variations of parameters *C* and γ.

We selected four input features including LRCO, CN, α-helical content, and β-sheet content with SVM regression to predict two-state folding rates. Besides two parameters C and γ used for SVM classification, SVM regression requires additional important parameter ε (regulate regression tube width) for performance optimization. Due to the intensive computational nature of grid search algorithm in high dimensions, we performed the tuning of parameter set (C, γ, ε) heuristically. We first obtained the optimal parameter values for C and γ with the fixed value of ε = 0.1 (default SVM value), then searched for the best value for ε. With the same procedure we did for SVM classification, we obtained the optimal parameter set of (C, γ, ε) = (8, 0.125, 0.1) for constructing prediction model.

Five different sets of training and testing data were generated. Each one was generated by randomly selecting 10% for testing and 90% for training from IvankovData. Then five different SVM prediction models using optimal parameter set was trained using leave-one-out cross-validation (LOOCV) and predicted on the test data sets. The average correlation and MAD are 0.81 and 0.78, respectively, from five test sets. The results are substantially better than the linear combination of multiple features, indicating the relationship between the features and folding rates is probably non-linear.

For multi-state folder rate prediction, one extra feature, protein sequence length, besides four other features used for two-state folders, was included for the SVM regression to predict multi-state folder's rate.

Our multi-feature SVM-regression method is compared with or better than other sequence-based methods in **Table 2**. Our method not only has better correlation between predicted rates and experimental rates than all the sequence-based method except FOLD-RATE, but also has smaller MAD (mean average difference) values between predicted and real rates than all the sequence-based methods. FOLD-RATE has obtained the highest 0.91 correlation between predicted and experimental rates, but its mean absolute difference between predicted and experimental values is around 1.1, which is much higher than our method. The reason could be due to FOLD-RATE breaks proteins into structural classes for individual training, which largely decrease the number of proteins per structural class, resulting in high correlation but high variance between predicted and real values. K-Fold uses experimental structure information to predict folding rates and classify folding types. Its accuracy for folding type classification is 0.81. Our novel sequence-based method has the fold type classification accuracy of 0.80, which is very close to that of K-fold.

To study how the integration of fold type classification and fold rate prediction would affect the results, we investigated a few cases. Chromosomal protein Ubiquitin (PDB ID: 1UBQ) has a sequence length of 76 amino acids and experimental folding rate of 7.3 (in natural-base logarithm scale) in the unit of sec^{-1}. It has been used by many researchers as multi-state folder [7,11], but later it was shown experimentally to be a two-state folder [5]. Assuming 1UBQ as multi-state folder, we used the multi-state prediction model and obtained fold rate of 3.97. But after being correctly classified into two-state using our SVM classification model, a value of 6.21 was obtained, which is much close to the experimental rate. DNA-binding protein Engrailed Homeodomain (PDB ID: 1ENH) is another example of such a case. It has a sequence length of 16 and folding rate of 10.5 (in natural-base logarithm scale) in the unit of sec^{-1}. Assuming it was as multi-state [18], then the predicted folding rate would be 2.55. However, our classification model has classified 1ENH as a two-state folder and we used two-state prediction model to predict the folding rate as 10.05. 1ENH has been shown and used as two-state folder in later literatures [7,11]. These examples demonstrated that our folding type classifier can help correct errors in manual folding type classifications.

4. Conclusion

We have developed a new protein fold rate prediction method (SeqRate) using Support Vector Machine regression with a set of features derived from protein sequences only. As the first method that can predict protein folding kinetic types from protein sequences, it achieved the accuracy comparable to the

methods based on experimental structures. The accuracy of fold rate prediction of the method was also improved over previous sequence-based prediction methods. SeqRate is a fast and robust method suitable for large-scale protein folding rate prediction.

Table 2. Comparison among different folding rate prediction methods.
Method 1: Effective length method [7]
Method 2: LRCO method [11]
Method 3: FOLD-RATE [13]
Method 4: K-Fold [4]
Method 5: Our multi-predictor SVM (two-state)
Method 6: Our multi-predictor SVM (multi-state)
Method-Type means if the method is using experimental structural data (structure) or using only sequence data (sequence). Correlation here means the correlation value between predicted rates and experimental rates. MAD is mean absolute difference between predicted rates and experimental rates.

Method	Method Type	Fold kinetic Classify Accuracy	Correlation	MAD
1	sequence	NA	0.70	0.96
2	sequence	NA	0.61	0.81
3	sequence	NA	0.91	1.1
4	structure	81%	0.74	0.75
5	sequence	80%	0.81	0.79
6	sequence	80%	0.80	0.68

5. References

[1] O.V. Galzitskaya, D.N. Ivankov, A.V. Finkelstein. Folding nuclei in proteins. FEBS Lett 2001;489:113–118.

[2] G.D. Fasman. Prediction of Protein Structure and the Principles of Protein Conformation. New York, NY: Plenum Press 1998.

[3] J. Cheng, A. Randall, M. Sweredoski, P. Baldi. SCRATCH: a Protein Structure and Structural Feature Prediction Server. Nucleic Acids Research 2005;33(Web server issue):W72-W76.

[4] E. Capriotti, R. Casadio. K-Fold: a tool for the prediction of the protein folding kinetic order and rate. Bioinformatics 2007;23(3):385-386.

[5] D.S. Wishart, D. Arndt, M. Berjanskii, A.C. Guo, Y. Shi, S. Shrivastava, J. Zhou, Y. Zhou, G. Lin. PPT-DB: the protein property prediction and testing database. Nucleic Acids Research 2008;36(Database issue):D222-D229.

[6] K.W. Plaxco, K.T. Simons, D. Baker. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;227:985-994.

[7] D.N. Ivankov, A.V. Finkelstein. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci 2004;101:8942-8944.

[8] J.T. Huang, J.P. Cheng, H. Chen. Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics. Proteins 2007;67:12-17.

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

[10] H. Zhou, Y. Zhou. Folding rate prediction using total contact distance. Biophys J 2002;82:458-462.

[11] M. Punta, B. Rost. Protein Folding Rates Estimated from Contact prediction. J Mol Biol 2005;348:507-512.

[12] V. Di Francesco, P.J. Munson, J. Garnier. FORESST: fold recognition from secondary structure predictions of proteins. Bioinformatics 1999;15(2):131-140.

[13] M.M. Gromiha, A.M. Thangakani, S. Selvaraj. FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res 2006;34:W70-74.

[14] A.V. Finkelshtein, O.V. Galzitskaya. Physics of protein folding. Phys Life Rev 2004;1:23-56.

[15] H. Gong, D.G. Isom, R. Srinivasan, G.D. Rose. Local secondary structure content predicts folding rates for simple, two-state protein. J Mol Biol 2003;327:1149-1154.

[16] B. Mao, K.C. Chou, C.T. Zhang. Protein folding classes: a geometric interpretation of the amino acid composition of globular proteins. Protein Engineering 1994;7(3):319-330.

[17] B.G. Ma, L.L. Chen, H.Y. Zhang. What determines protein folding type? An investigation of intrinsic structural properties and its implications for understanding folding mechanisms. J Mol Biol 2007;370:439-488.

[18] S. Gianni, N.R. Guydosh, F. Khan, T.D. Caldas, U. Mayor, G.W.N. White, M.L. DeMarco, V. Daggett, A.R. Fersht. Unifying features in protein-folding mechanism. Proc Natl Acad Sci 2003;100:13286-13291.