IPSJ Transactions on Bioinformatics Vol. 3 40–53 (May 2010)

Original Paper

Prediction of Protein Folding Rates from Structural Topology and Complex Network Properties

Jiangning Song,[†1,†2] Kazuhiro Takemoto,[†3]
Hongbin Shen,[†4] Hao Tan,[†2] M. Michael Gromiha[†5]
and Tatsuya Akutsu[†1]

As a fundamental biological problem, revealing the protein folding mechanism remains to be one of the most challenging problems in structural bioinformatics. Prediction of protein folding rate is an important step towards our further understanding of the protein folding mechanism and the complex sequence-structure-function relationship. In this article, we develop a novel approach to predict protein folding rates for two-state and multi-state protein folding kinetics, which combines a variety of structural topology and complex network properties that are calculated from protein three-dimensional structures. To take into account the specific correlations between network properties and protein folding rates, we define two different protein residue contact networks, based on two different scales Protein Contact Network (PCN) and Long-range Interaction Network (LIN) to characterize the corresponding network features. The leave-one-out cross-validation (LOOCV) tests indicate that this integrative strategy is more powerful in predicting the folding rates from 3D structures, with the Pearson's Correlation Coefficient (CC) of 0.88, 0.90 and 0.90 for two-state, multi-state and combined protein folding kinetics, which provides an improved performance compared with other prediction work. This study provides useful insights which shed light on the network organization of interacting residues underlying protein folding process for both two-state and multi-state folding kinetics. Moreover, our method also provides a complementary approach to the current folding rate prediction algorithms and can be used as a powerful tool for the characterization of the foldomics protein data. The implemented webserver (termed PRORATE) is freely accessible at http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/folding/.

1. Introduction

A major issue in molecular biology today is to understand how a protein folds into its characteristic three-dimensional (3D) structure and how to gain its biological function as a linear string of amino acid sequence[1]. Unraveling the protein folding mechanisms remains to be one of the most challenging problems and has been considered as deciphering the second half of genetic code[2]. Protein folding rate is a measure for evaluating how slow or fast the folding of proteins from the unfolded state to native three-dimensional structure[3], which is usually described by the folding rate constant K_f, whose unit is $\sec^{-1}$. On one hand, proteins can fold into their native structures at very different folding rates, varying from several microseconds to even an hour[4]. On the other hand, subtle changes in the solvent environment or protein sequence can dramatically alter the protein folding kinetics, accounting for the distinct kinetic behaviors under different experimental conditions[5]. Further, the misfolding of proteins into non-native states altered by the folding kinetics could lead to several degenerative disorders, such as prion and Alzheimer's disease[6]. Numerous previous studies of protein folding kinetics as well as its association with protein structure and function, either from the perspectives of experimentalists or theoreticians, have led to our improved understanding of the physical processes of protein folding and the fundamental rules governing protein folding behaviors[7)–12].

Prediction of protein folding rate from its amino acid sequence is an important step towards our understanding of the protein folding mechanism and the complex sequence-structure-function relationship[4]. Previous studies have indicated that protein folding kinetics can be categorized into two kinetic orders: simple two-state (TS) folding behaviors without the visible intermediates, and three-state (or multi-state, MS) folding kinetics that exhibits the obvious intermediate state during folding process under experimental conditions[4]. Furthermore, some proteins can undergo the switching process from two-state to multi-state or vice versa, by single point mutations or simply changing the experimental conditions such as the temperature or solvent concentration[13),14]. With the increasing availability of protein folding data deposited in public databases as the consequence of structural genomics projects[9),10], efficient computational tools are desired to

[†1] Bioinformatics Center, Institute for Chemical Research, Kyoto University
[†2] Department of Biochemistry and Molecular Biology, Monash University
[†3] PRESTO, Japan Science and Technology Agency
[†4] Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University
[†5] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology

be developed to predict protein folding rates, which will not only provide important complementary information for annotating protein folding data, but also contribute to the deep understanding of protein folding mechanisms.

In the past two decades, a number of prediction studies have been performed to infer protein folding rates using different topological parameters from three-dimensional structures [5),7),8),12)–25)]. The majority of these analyses mainly focused on inferring the statistical significance of the correlations between protein folding rate and different topological parameters, including contact order (CO)[5)], absolute contact order (Abs_CO)[19)], total contact distance (TCD)[16)], long range order (LRO)[15)], long range contact order (LR_CO)[14)], effective secondary structure length (L_{eff})[4)], the fraction of local contact (FLC)[7),8)] and chain topology parameter (CTP)[21)]. Gromiha, et al. implemented a web server FOLD-RATE, which used multiple regression equations based on 49 physical-chemical, energetic and conformational properties of amino acid residues to predict protein folding rates from amino acid sequences[3)]. More recently, Capriotti and Casadio developed K-Fold server to predict the protein folding kinetic order and folding rate using support vector machine based on a dataset with 63 proteins[26)].

Graph theoretic approaches that model protein structures as connecting networks of interacting residues, from the perspective of complex networks[27),28)], provide new insights into the importance of the key residues that are characterized by a relatively small number of vertices with large connectivity and play an essential role in the protein folding process[29)–33)]. Moreover, a most-recent study indicates that both Protein Contact Network (PCN) and Long-range Interaction Network (LIN) exhibit the "assortative mixing" phenomena and their corresponding assortative coefficients show positive correlations with the folding rates of thirty singled-domain two-state proteins[11)]. Based on the these views, it would be interesting to investigate whether protein folding rates can be more accurately predicted on the basis of the integration of various structural topology parameters and the general complex network properties calculated from protein 3D structures.

In the present study, we propose a novel approach to predict the folding kinetic orders and folding rates for two-state and multi-state protein folders using support vector regression (SVR) approach. We combine a variety of structural topology parameters with complex network properties with respect to the PCN and LIN networks as the input features into the SVR models, which allows accurate quantification of the relationships between protein folding rates and these structure and network properties. We construct the SVR models by mapping these input feature vectors into a high-dimensional feature space using the non-linear polynomial kernel functions. The rigorous leave-one-out cross-validation (LOOCV) tests show that the generic complex network properties coupled with structural topology parameters can significantly improve the prediction accuracy, suggesting that this approach can be effectively utilized for reliable inference of protein folding rates and folding kinetic orders, which could provide important complementary information for the annotation of the foldomics data.

2. Methods

2.1 Datasets

A larger dataset that has been recently constructed by Ouyang and Liang[12)] was used as the benchmark dataset in this study. It consists of 80 protein folders with their folding rates experimentally determined. Among them, 45 proteins exhibit two-state (TS) folding behaviors without the visible intermediates, while the other 35 proteins belong to the three-state or multi-state (MS) folding kinetics that exhibit the obvious intermediate state during the folding process under experimental conditions. They belong to different structural classes: 18 are all-α proteins, 32 are all-β proteins, and the remaining 30 are $\alpha\beta$ proteins. The folding rates of these 80 proteins range from $\ln K_f = -6.9$ to $\ln K_f = 12.9$ (where K_f is the experimental folding rate), which is more than eight orders of magnitude. The detailed PDB codes for these folders with the TS or MS folding kinetics, as well as the experimentally determined folding rates can be found in their work [12)].

2.2 Structural Topology Measures

Previous studies have indicated that several structural topology properties of a protein have significant correlations with protein folding rates. In this study, we selected and calculated eight topology measures to investigate their specific correlations with the folding rates both for two-state and multi-state proteins, by defining four different sphere radii R_d centered on three respective C_α, C_β or non-hydrogen atoms of the target residue, i.e., R_d=5, 6, 7 and 8Å [5)]. These eight parameters are defined as follows.

2.2.1 Contact Order (CO)

Contact order is given by:

$$CO = \frac{1}{n_c n_r} \sum_{\substack{k=1 \\ |i-j|>2}}^{n_c} \Delta L_{ij}$$

where unless otherwise stated, n_r is the number of amino acid residues in a protein, n_c is the total number of contacts, and ΔL_{ij} is the sequence separation between contacting residues i and j in the protein sequence [5].

2.2.2 Absolute Contact Order (Abs_CO)

Absolute contact order (Abs_CO) is defined by Ivankov et al. [19]:

$$Abs\_CO = \frac{1}{n_c n_r} \sum_{\substack{k=1 \\ |i-j|>2}}^{n_c} \Delta L_{ij}$$

2.2.3 Total Contact Distance (TCD)

The total contact distance (TCD) [16] is defined as:

$$TCD = \frac{1}{n_r^2} \sum_{\substack{k=1 \\ |i-j|>0}}^{n_c} |i-j|$$

2.2.4 Long Range Order (LRO)

Long range order (LRO) denotes the residue contacts between two residues that are close in space but far from each other in the sequence [15]. It is defined as:

$$LRO = \sum \frac{n_{i,j}}{n_r}, \begin{cases} n_{i,j} = 1 \text{ if } |i-j| > 12 \\ n_{i,j} = 0 \text{ otherwise} \end{cases}$$

where i and j are two residues whose C_α–C_α distance is ≤ 8Å.

2.2.5 Long Range Contact Order (LR_CO)

Long range contact order (LR_CO) denotes the residue contacts between two sequentially distant residues with the C_α–C_α distance less than a cutoff of R_d ($R_d = 5, 6, 7$ and 8Å) [14]. It is defined as:

$$LR\_CO = \frac{1}{n_r^2} \sum_{\substack{k=1 \\ |i-j|>L_{\text{cut}}}}^{n_c} |i-j|$$

where ΔL_{ij} is the sequence separation between contacting residues i and j in the protein sequence. In this study, we set $L_{\text{cut}} = 12$.

2.2.6 Effect Protein Chain Length (L_{eff})

The effect length of protein chain is defined by

$$L_{\text{eff}} = L - LH + l \times NH$$

where L is the protein chain length, LH is the number of residues in helical conformation, NH is the number of helices, and l denotes the chain length of the α-helix turn ($l \leq 4$) [4]. We set $l = 3$ in this study.

2.2.7 The Fraction of Local Contact (FLC)

The fraction of local contact (FLC), i.e. contacts between residues is defined by

$$FLC = \frac{\sum_{|i-j|\leq 4} \delta(i,j)}{\sum_{i,j} \delta(i,j)}$$

where $\delta(i,j) = 1$ if residues i and j are in contact, and 0 otherwise [7],[8].

2.2.8 Chain Topology Parameter (CTP)

The chain topology parameter (CTP) is defined by

$$CTP = \frac{1}{n_r n_c} \sum \Delta L_{ij}^2$$

where ΔL_{ij} is the sequence separation between the contacting residues i and j in the protein sequence [21].

2.3 Complex Network Properties

In recent years, graph-theoretic approaches have well established that protein structures can be modeled as complex networks of interacting residues that are characterized by large values of the clustering coefficient C and small values of the characteristic path length [27)–31),33)]. The representation of protein structures as complex networks of interacting amino acid residues has been applied as a powerful tool to study a variety of problems in structural bioinformatics in regards to protein structure-function relationship, such as the identification of key residues involved in protein folding [29],[30], the correlations between network properties and the determinants of protein folding[31], the identification of functional residues in protein structures[34], the prediction of central residues at protein-protein interfaces[33],[35], the prediction of viable circular permutants[36] and the assortative mixing in protein contact networks and protein folding kinetics[11].

It is of particular interest to investigate whether the complex network properties derived from protein structures can be used to accurately predict folding rates for both two-state and multi-state protein folders. Here, we consider building two kinds of network models based on two length scales[11]: Protein Contact Network (PCN) that takes into consideration all inter-residue contacts and Long-range Interaction Network (LIN) that considers only the long-range interactions formed by the residues whose sequential distance $L_{cut} \geq 12$ and excludes the short-range interactions[15]. For a better understanding of the difference between the PCN and LIN, the graphical representations of the three-dimensional structure of the Actin-binding protein (PDB ID: 2VIK) were provided as an illustration (**Fig. 1**). Figure 1 A and B are its all-atom model and backbone representations, respectively. Its LIN representation (Fig. 1 C) is actually a subset of its corresponding PCN (Fig. 1 D) with the same numbers of nodes but lesser numbers of edges due to the removal of the short-range interactions.

In this study, the C_α (or C_β or the non-hydrogen atom) of each residue in a protein structure is considered as the node of the network, two residues will be regarded as in contact if their node atoms locate within a sphere of the threshold radius R_d. We defined and calculated several types of complex network properties on the two length scales of PCN and LIN:

2.3.1 Clustering Coefficient (CC)

The clustering coefficient for residue i is defined as $C_i = 2M_i/[k_i(k_i-1)]$, where M_i denotes the number of contacts among neighbors of residue i, and k_i is the number of contacts of residue i[27]. Then we defined the average value of C_i as the clustering coefficient:

$$CC = \frac{1}{n_r}\sum_{i=1}^{n_r} C_i$$

where n_r is the number of residues in a protein.

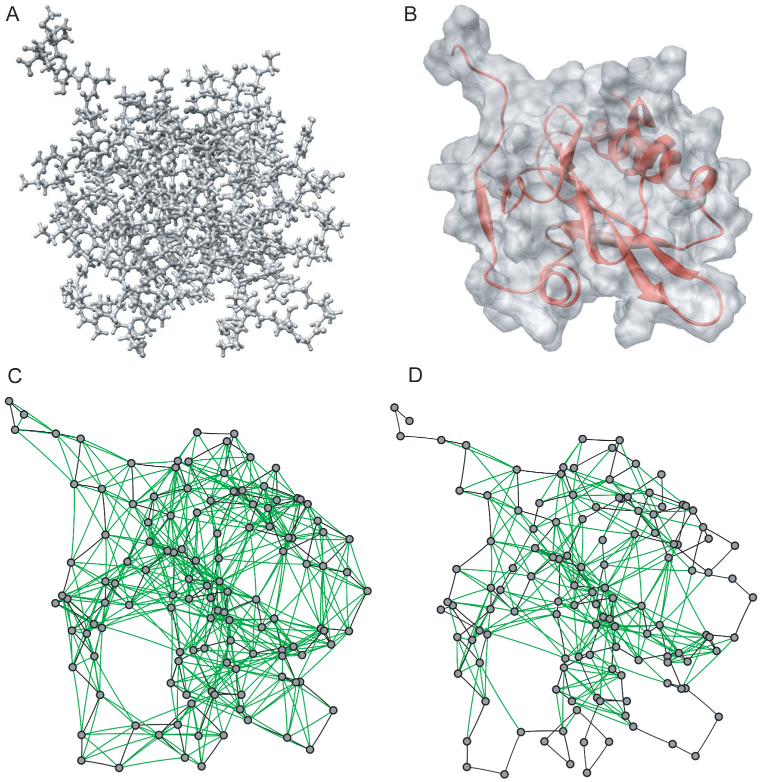


Fig. 1 Graphical representations of the three-dimensional structure of the Actin-binding protein (PDB ID: 2VIK) in four different manners. (A) The all-atom model representation; (B) The backbone model representation highlighted in red color while the surface was shown in gray; (C) The Protein Contact Network (PCN) representation; (D) The Long-range Interaction Network (LIN) representation. The radius cutoff was set at $R_d = 8$Å to construct the PCN and LIN. The black edges denote the main chain, while the green edges represent the inter-residue contact. A and B were rendered using UCSF Chimera package[40], and C and D were generated using network software Pajek[41].

2.3.2 Cyclic Coefficient (CYC)

The cyclic coefficient of residue i is defined as

$$CYC_i = \frac{2}{k_i(k_i-1)}\sum_{\langle jh\rangle}(S^i_{jh}-2)^{-1}$$

where k_i is the number of contacts of residue i, and $\langle jh\rangle$ is for all the pairs of neighbors of the residue i. S^i_{jh} is the smallest size of the closed path that passes through residue i and its two neighbor residues j and h[37].

Then we define the average value of CYC_i as the cyclic coefficient:

$$CYC = \frac{1}{n_r}\sum_{i=1}^{n_r} CYC_i$$

2.3.3 Triangle Density (TD)

We define the triangle density as,

$$TD = \frac{3T}{n_r}$$

where T is the number of triangles in a network[38].

2.3.4 Characteristic Path Length (CPL)

The characteristic path length is defined as follows

$$CPL = \frac{1}{n_r(n_r-1)}\sum_{ij} d(i,j)$$

where $d(i,j)$ is the shortest path length between residues i and j[27].

2.3.5 Assortative Coefficient (AC)

The assortative coefficient is defined as

$$AC = \frac{4\langle k_i k_j\rangle - \langle k_i + k_j\rangle^2}{2\langle k_i^2 + k_j^2\rangle - \langle k_i + k_j\rangle^2}$$

where k_i and k_j are the number of contacts of two residues at the ends of a contact (edge), and $\langle\ldots\rangle$ denotes the average overall contacts[39].

2.4 Support Vector Regression

Support vector machine (SVM) is an efficient machine learning technique based on Statistical Learning Theory[42]. This algorithm separates the positive from the negative samples by mapping the data into a higher dimensional feature space and constructing an optimal separating hyperplane (OSH) that maximizes its distance from the closest training samples. SVM can usually perform better than other machine learning algorithms due to its excellent capacity and ability to control error without causing overfitting to the data. It has been widely used in many classification problems of bioinformatics and computational biology, such as microarray data analysis[43], protein subcellular location[44], single nucleotide polymorphisms (SNPs) prediction[45], proline cis/trans isomerization prediction[46], protein fold recognition[47], residue contact prediction[48], disulfide connectivity prediction[49], protein-protein interaction[50],[51] and DNA-repair protein prediction[52].

SVM has two practical modes: support vector classification (SVC) and support vector regression (SVR). Compared with SVC, SVR has outstanding ability in predicting the property values of testing samples. SVR has attracted increasing attention in recent years and has been successfully applied in the real-value prediction studies of accessible surface area[53], residue contact number[54]–[56], residue-wise contact orders[57],[58], missing value estimation in microarray data[59], disulfide connectivity pattern[60], half-sphere exposure[56], residue depth[61]–[63] and caspase substrate cleavage sites[64].

As an implementation of the SVR approach, the SVM_light package[65] was used in this study to train and build the SVR classifiers for the two-state and multi-state protein folding kinetics. The regularization parameter C and the polynomial kernel degree d need to be determined in advance. The selection of the kernel function parameters is an important step for SVR training and testing, because it implicitly determines the structure of the high dimensional feature space when constructing the OSH. In the present study, we selected different combinations of optimal parameters of polynomial kernel functions to build the different SVR models with respect to the TS and MS protein folders.

2.5 Performance Evaluation

In order to objectively evaluate the performance and avoid the over-fitting, we performed the back-check and the leave-one-out cross-validation (LOOCV) tests. In the LOOCV test, one sample was singled out in turn as the testing dataset, while all other samples in the dataset were merged into the training dataset to infer prediction rules and build SVR models.

For the classification task of predicting protein's folding kinetic orders, we evaluated the performance by calculating the overall accuracy (ACC), Sensitivity, Specificity and the Matthew's correlation coefficient (MCC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

$$Specificity = \frac{TN}{TN + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

For the regression task of predicting protein folding rates, the Pearson's correlation coefficient (CC) between the predicted and observed folding rates and the root mean square error (RMSE) are used to evaluate the prediction performance.

3. Results

3.1 The Difference between the TS and MS Folders Indicated by Topology and Network Properties

We calculated topology measures and network properties for the TS and MS protein folders (Appendix **Table 4**). The results indicate that there are two significant topology measures that show distinguishable preferences for the TS and MS proteins. One significant measure is CO which was originally proposed by Plaxco et al.[5] to describe the complexity of protein topology and has been shown to be strongly correlated with the folding rates of the TS proteins. The average value of CO measure of TS proteins is significantly larger than that of the MS proteins (*p*-value=0.005 by the *t*-test), indicating that the topological structure of TS proteins are more complex than that MS folders. The other significant topology parameter is LR_CO (*p*-value = 0.002), which is a measure of long-range contact order. The TS proteins have much higher LR_CO values in contrast to the MS proteins, which is in a good agreement with previous work by Ma, et al.[14]. On the other hand, five out of ten different complex network properties exhibit significant statistical significance between the TS and MS protein folding kinetics, including four properties of PCN (CC_PCN, CYC_PCN, CPL_PCN and AC_PCN) and one property of LIN (AC_LIN). These results also indicate the differences between the topological characteristics of PCNs and LINs. The PCNs have larger clustering coefficients, cyclic coefficients and triangle densities than the corresponding LINs, which on the contrary have larger characteristic path lengths and larger assortative coefficients than the PCNs (due to the reduced number of short-range residue contacts in LINs[11]).

3.2 Specific Correlations Between Topology Parameters, Network Properties and Protein Folding Rates

We next computed the Pearson's correlation coefficients between topological parameters/network properties and the corresponding protein folding rates in our dataset (**Table 1**). We observed that five topology parameters (CO, Abs_CO, TCD, LRO and CTP) show significant negative correlations, and FLC has significant positive correlation with the folding rates of TS proteins. However, in the case of the MS protein folders, CO and LR_CO exhibit positive correlations with their folding rates. This correlation differentiation between the same topology measures with the folding rates might imply the difference of folding mechanisms

Table 1 Correlation coefficients between topology parameters, network properties and the corresponding folding rate lnK_f values. The results are computed with the traditional threshold R_d = 8Å using the C$_\alpha$ atom for the TS proteins as the node and R_d = 8Åusing the non-hydrogen atom for the MS proteins as the node, respectively.

Measures		Two-state	Multi-state	Overall
Topology	CO	− 0.725	0.406	−0.191
	Abs_CO	−0.512	−0.845	−0.583
	TCD	−0.746	0.095	−0.291
	LRO	−0.733	–	−0.585
	LR_CO	−0.020	0.572	0.297
	FLC	0.678	0.587	0.498
	CTP	−0.567	−0.771	−0.570
	Prolength	−0.108	−0.838	−0.428
Network	CC_PCN	0.321	0.803	0.516
	CC_LIN	−0.753	−0.041	−0.494
	CYC_PCN	0.278	0.810	0.504
	CYC_LIN	−0.708	−0.227	−0.512
	TD_PCN	−0.411	−0.600	−0.401
	TD_LIN	−0.756	−0.637	−0.555
	CPL_PCN	0.048	−0.656	−0.230
	CPL_LIN	0.398	−0.175	0.129
	AC_PCN	0.186	−0.351	−0.137
	AC_LIN	0.353	−0.353	−0.062

Table 2 Prediction performances in terms of CC and RMSE using different SVR models based on topology, network and the combined features.

SVR models		Topology		Network		Combined	
		Back-check	Jack-knife	Back-check	Jack-knife	Back-check	Jack-knife
Two-state	CC	0.810	0.780	0.856	0.791	0.933	0.853
	RMSE	2.20	2.34	1.93	2.29	1.34	1.95
Multi-state	CC	0.872	0.821	0.831	0.813	0.882	0.824
	RMSE	1.84	2.14	2.08	2.18	1.77	2.12

of the TS and MS proteins. The overall correlations between folding rates and topology parameters based on the whole dataset without TS and MS classification are also presented as a comparison in Table 1.

With respect to the complex network properties, we also observed that there are significant correlations between three network properties (CC_LIN, CYC_LIN and TD_LIN) and the corresponding folding rates of the TS proteins, whose CCs are -0.753, -0.708 and -0.756, respectively. All these network parameters have strong negative correlations with the folding rates of TS proteins. It is particularly interesting to notice that all LINs' properties exhibit stronger correlations with protein folding rates in contrast to the corresponding PCNs (Table 1). Nevertheless, when it comes to the MS proteins, four PCN properties have significant correlations with the folding rates. For example, CC_PCN and CYC_PCN have significant positive correlations with MS folding rates, whereas TD_PCN and CPL_PCN have strong negative correlations. Only one LIN property TD_LIN exhibits significant correlation with the MS folding rate. Based on these observations, we conclude that PCN parameters have better correlations with the folding rates of the MS proteins, while LIN measures have stronger correlations with the folding rates of the TS proteins. All these findings suggest that distinctive folding mechanisms hold for the TS and MS protein folding kinetics.

3.3 Improving Folding Rate Prediction by Integrating Topology Parameters, Network Properties and Combined Features

To explore the possibility of improving the prediction of protein folding rates, we further encoded these topology and/or network parameters as the input features into SVR classifiers. Feature selection was performed using a recursive elimination strategy: the SVR models were initially trained and built using all the feature vectors based on eight topology parameters, ten network properties, and as well as these combined features as the input. Then a feature that is considered as making no contribution to the predictive performance will be removed from the input feature sets, if the subsequent performance of SVR remains steady or even increases after its removal. The aim of this procedure is to improve the prediction accuracy, due to the fact that using all the features might not lead to the best prediction performance [56),57),60),63),64)]. When no further performance improvements were observed, we fixed up the selected feature sets, re-trained our SVR models and predicted the folding rates. The final feature sets used for building the SVR models are described in Appendix **Table 5** and the resulting prediction performances are summarized in **Table 2**.

We assume that the uniquely networked structures (if there are any) can be reflected and captured by these topology and network properties. We also compared the performance based on different network sizes (network sizes depend on the R_d thresholds. Assigning lower R_d values results to more densely connected networks while higher R_d values lead to sparsely connected networks) and found that the resulting performance differences have fluctuations but are not significant. The preliminary results show that single-property SVR model (using only one property at a time to build the SVR model to predict folding rate) can provide relative success to predict folding rate, however, the performance of using single-property SVR models is worse than that of using multi-property SVR models. This might imply that incorporating multiple properties in terms of topology and network parameters that can provide complementary information have potential advantages in improving the predictive performance.

The SVR classifier based on multiple topology parameters could predict the folding rates with the CC of 0.780 and RMSE of 2.34 for the TS proteins and with the CC of 0.821 and RMSE of 2.14 for the MS proteins, respectively, when

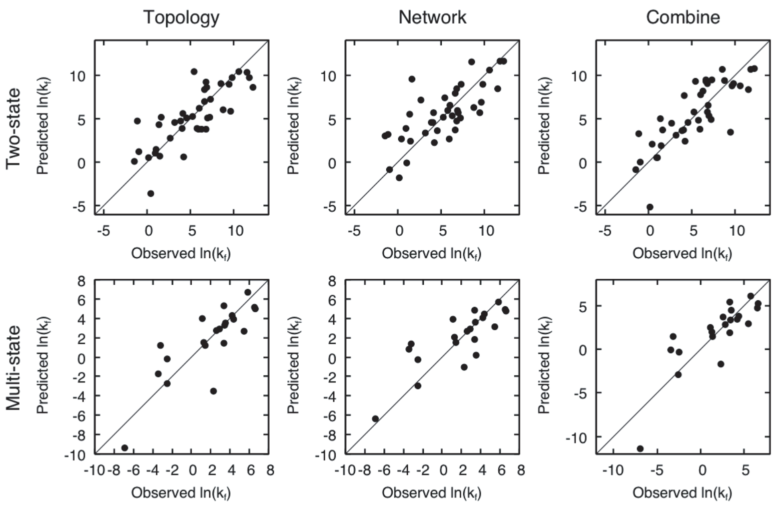


Fig. 2 The scatter-plots of the observed and predicted folding rates of the TS and MS protein folders by the jack-knife cross-validation test.

evaluated by the leave-one-out cross-validation (LOOCV) tests. The back-check prediction results are also included here (Table 2). As a comparison, the SVR classifier based on multiple network properties could provide prediction accuracy with the CC of 0.791 and RMSE of 2.29 for the TS proteins, and with the CC of 0.813 and RMSE of 2.18 for the MS proteins, respectively. In the case of two-state protein folding kinetics, the SVR classifier based on network properties performed better than that based on topology parameters. In contrast, for the multi-state protein folding, the SVR classifier based on topology parameter provides better performance compared with that based on network properties (**Fig. 2**). We argue that these results might be a reflection of the difference of folding mechanisms between the two-state and multi-state protein folders.

Moreover, after combining the topology and network properties, the resulting SVR classifier further improves the prediction accuracy: for the TS proteins, CC is equal to 0.853 and RMSE is 1.95, while for the MS proteins, CC is equal to 0.824 and RMSE is 2.12. The prediction accuracy achieved by integrating multiple topology parameters suggest that this prediction strategy is successful in improving the performance compared with linear regression equation using single-parameter as input. This observation is also consistent with previous studies that using multiple topology parameters of protein structure can improve the prediction performance of protein folding rates[16),26)].

3.4 Formulating as a Two-class Prediction Problem and Comparing Prediction Performance with Two Recent Studies

Since previous studies examined the predictive performance by a conventional two-class classification, namely, to predict whether a protein folds via two-state or multi-state kinetics, we also examined and compared our SVR classifier with two recent methods, including the binary logistic regression (BLR) which uses chain length as the feature[66)] and the composition-based predictor which is based on the differentiation of amino acid contents between the TS and MS folders[14)]. To make an objective comparison, these methods are measured on the same training and test datasets. The result comparison is presented in **Table 3**. As can be seen, the SVR classifier performs much better than the BLR method, with the ACC and MCC scores by 1.1% and 0.024 higher than those of the BLR, respectively. The SVR classifier also compares favorably with the composition-based predictor with the same accuracy of ACC of 80.8%. These results suggest that this SVR classifier is at least competitive with, if not better than, the two recently developed methods.

4. Discussion

Prediction of protein folding rates is an important step towards our deep understanding of the protein folding mechanism and remains to be one of most challenging tasks in structural bioinformatics today. One of the main contributions of this paper is that we comprehensively integrate the complex network properties along with a variety of structural topology features of protein structures as the input features to build the SVR classifiers in order to improve the prediction performance. In particular, for the TS proteins, the predictive power of network properties is stronger than that of structural topology parameters, suggesting that network properties can be used to better describe the underlying mechanism that dominates the TS protein folding process. On the other hand, topology parameters are more indicative of the MS protein folders than the net-

Table 3 Two-class prediction accuracy in terms of the Sensitivity, Specificity, ACC and MCC for the prediction of two-state and multi-state protein folders by different approaches.

Methods		Prediction accuracy (%)			
		Sensitivity	Specificity	ACC	MCC
Comparison with Ma, et al.	Composition-based predictor	79.7%	82.0%	80.8%	–
	SVR	76.0%	85.7%	80.8%	0.607
Comparison with Huang and Cheng	Binary logistic regression (BLR)	98.3%	72.0%	90.6%	0.774
	SVR	90.6%	95.0%	91.7%	0.798

work measures, which might imply that the topology parameters are the most important determinants in the case of the MS folding kinetics.

To improve the prediction performance of protein folding rates, we adopted the recursive elimination strategy to optimize the feature selection of the SVR by comparing the performance using different combinations of topology and network parameters. The primary goal here was to improve the prediction accuracy, due to the fact that using all the features together might not lead to the best prediction performance. However, several ways may help to further improve the prediction performance in the future. The first method is to use more accurately determined PDB structure data with better resolutions, as it is well-known that SVR has better performance when trained on larger dataset with adequate training samples. The second strategy can be focused on improvement of feature selection and SVR parameter selection procedures that have important effect on the final prediction accuracy. The third way is to use high-quality folding rate dataset that has refined data representation, which can ensure better representation particularly for the MS protein folders when fed into the SVR classifiers. This might be applicable when more protein foldomics data are available[9),10)].

It is likely that the improvement in prediction accuracy for both the TS and MS protein folders is a reflection of the fact that the folding mechanism of a protein is largely determined by its global structural topology and network organization rather than its local inter-atomic interactions, as previously discussed by Bagler and Sinha[11)]. The specific correlations between various network properties and protein folding rates found in this study may further enhance our understanding of the protein folding process from the perspective of complex network organization. Our method provides useful insights by utilizing as many as ten different properties of the complex networks in the form of the PCNs and LINs, which could shed light on the network organization underlying the complex protein folding process that applies not only to the two-state but also to the multi-state protein folding kinetics.

5. Conclusion

In this work, we attempted to address the important problem of predicting protein folding rates of proteins with two-state and multi-state folding kinetics, by developing a multiple-feature framework based on support vector regression (SVR) approach. Our method integrated a variety of structural topology and complex network properties as the input features into the SVR models. We comprehensively investigated the specific correlations between topology parameters/network properties and protein folding rates, based on short-range and long-range contact scales: Protein Contact Network (PCN) and Long-range Interaction Network (LIN). Statistical analyses indicate that LINs show much stronger correlations with protein folding rates in compassion with the corresponding PCNs. Moreover, our approach could yield favorable or at least comparable prediction performance in contrast to two recently published methods. The results highlighted that our integrative approach is computationally competitive and can be used as a powerful tool for the characterization of the foldomics protein data.

Acknowledgments JS would like to thank the National Health and Medical Research Council of Australia (NHMRC) and the Japan Society for the Promotion of Science (JSPS) for supporting this research via the NHMRC Peter Doherty and JSPS Postdoctoral Fellowships. TA was partially supported by Grant-in-Aid #19200022 from MEXT, Japan. HS was supported by the National

Natural Science Foundation of China (60704047), Science and Technology Commission of Shanghai Municipality (08ZR1410600, 08JC1410600), and sponsored by Shanghai Pujiang Program and Innovation Program of Shanghai Municipal Education Commission (10ZZ17). The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

1) Anfinsen, C.B.: Principles that govern the folding of protein chains, *Science*, Vol.181, No.96, pp.223–230 (1973).

2) Gierasch, L.M. and King, J.: *Protein Folding: Deciphering the Second Half of the Genetic Code*, American Association for the Advancement of Science, Washington DC (1990).

3) Gromiha, M.M., et al.: FOLD-RATE: prediction of protein folding rates from amino acid sequence, *Nucleic Acids Res.*, Vol.34, pp.W70–74 (2006).

4) Ivankov, D.N. and Finkelstein, A.V.: Prediction of protein folding rates from the amino acid sequence-predicted secondary structure, *Proc. Natl. Acad. Sci. USA*, Vol.101, No.24, pp.8942–8944 (2004).

5) Plaxco, K.W., et al.: Contact order, transition state placement and the refolding rates of single domain proteins, *J. Mol. Biol.*, Vol.277, No.4, pp.985–994 (1998).

6) Taubes, G.: Protein Chemistry: Misfolding the way to Disease, *Science*, Vol.271, No.5255, pp.1493–1495 (1996).

7) Mirny, L. and Shakhnovich, E: Protein folding theory: from lattice to all-atom models, *Annu. Rev. Biophys. Biomol. Struct.*, Vol.30, pp.361–396 (2001).

8) Kuznetsov, I.B. and Rackovsky, S.: Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors, *Proteins*, Vol.54, No.2, pp.333–341 (2004).

9) Fulton, K.F., et al.: PFD: A database for the investigation of protein folding kinetics and stability, *Nucleic Acids Res.*, Vol.33, pp.D279–283 (2005).

10) Fulton, K.F., et al.: Protein Folding Database (PFD 2.0): An online environment for the International Foldeomics Consortium, *Nucleic Acids Res.*, Vol.35, pp.D304–307 (2007).

11) Bagler, G. and Sinha, S.: Assortative mixing in Protein Contact Networks and protein folding kinetics, *Bioinformatics*, Vol.23, No.14, pp.1760–1767 (2007).

12) Ouyang, Z. and Liang, J.: Predicting protein folding rates from geometric contact and amino acid sequence, *Protein Sci.*, Vol.17, No.7, pp.1256–1263 (2008).

13) Huang, J.T., et al.: Secondary structure length as a determinant of folding rate of proteins with two- and three-state kinetics, *Proteins*, Vol.67, No.1, pp.12–17 (2007).

14) Ma, B.G., et al.: What Determines Protein Folding Type? An Investigation of Intrinsic Structural Properties and its Implications for Understanding Folding Mechanisms, *J. Mol. Biol.*, Vol.370, No.3, pp.439–448 (2007).

15) Gromiha, M.M. and 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.*, Vol.310, No.1, pp.27–32 (2001).

16) Zhou, H. and Zhou, Y.: Folding rate prediction using total contact distance, *Biophys. J.*, Vol.82, No.1, pp.458–463 (2002).

17) Galzitskaya, O.V., et al.: Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics, *Proteins*, Vol.51, No.2, pp.162–166 (2003).

18) Gong, H., et al.: Local secondary structure content predicts folding rates for simple, two-state proteins, *J. Mol. Biol.*, Vol.327, No.5, pp.1149–1154 (2003).

19) Ivankov, D.N., et al.: Contact order revisited: influence of protein size on the folding rate, *Protein Sci.*, Vol.12, No.9, pp.2057–2062 (2003).

20) Micheletti, C.: Prediction of folding rates and transition-state placement from native-state geometry, *Proteins*, Vol.51, No.1, pp.74–84 (2003).

21) Nölting, B., et al.: Structural determinants of the rate of protein folding, *J. Theor. Biol.*, Vol.223, No.3, pp.299–307 (2003).

22) Punta, M. and Rost, B.: Protein folding rates estimated from contact predictions, *J. Mol. Biol.*, Vol.348, No.3, pp.507–512 (2005).

23) Dixit, P.D. and Weikl, T.R.: A simple measure of native-state topology and chain connectivity predicts the folding rates of two-state proteins with and without crosslinks, *Proteins*, Vol.64, No.1, pp.193–197 (2006).

24) Ma, B.G., et al.: Direct correlation between proteins' folding rates and their amino acid compositions: An ab initio folding rate prediction, *Proteins*, Vol.65, No.2, pp.362–372 (2006).

25) Prabhu, N.P. and Bhuyan, A.K.: Prediction of folding rates of small proteins: empirical relations based on length, secondary structure content, residue type, and stability, *Biochemistry*, Vol.45, No.11, pp.3805–3812 (2006).

26) Capriotti, E. and Casadio, R.: K-Fold: A tool for the prediction of the protein folding kinetic order and rate, *Bioinformatics*, Vol.23, No.3, pp.385–386 (2007).

27) Watts, D.J. and Strogatz, S.H.: Collective dynamics of 'small-world' networks, *Nature*, Vol.393, No.6684, pp.440–442 (1998).

28) Vázquez, A., et al.: The topological relationship between the large-scale attributes and local interaction patterns of complex networks, *Proc. Natl. Acad. Sci. USA*, Vol.101, No.52, pp.17940–17945 (2004).

29) Vendruscolo, M., et al.: Three key residues form a critical network in a protein folding transition state, *Nature*, Vol.409, No.6820, pp.641–645 (2001).

30) Vendruscolo, M., et al.: Small-world view of the amino acids that play a key role in protein folding, *Phys. Rev. E*, Vol.65, 061910 (2002).

31) Greene, L.H. and Higman, V.A.: Uncovering network systems within protein struc-

tures, *J. Mol. Biol.*, Vol.334, No.4, pp.781–791 (2003).

32) Atillgan, A.R., et al.: Small-world communication of residues and significance for protein dynamics, *Biophys. J.*, Vol.86, No.1, pp.85–91 (2004).

33) Del Sol, A., et al.: Topology of small-world networks of protein-protein com-plex structures, *Bioinformatics*, Vol.21, No.8, pp.1311–1315 (2005).

34) Amitai, G., et al.: Network analysis of protein structures identifies functional residues, *J. Mol. Biol.*, Vol.344, No.4, pp.1135–1146 (2004).

35) Del Sol, A. and O'Meara, P.: Small-world network approach to identify key residues in protein-protein interaction, *Proteins*, Vol.58, No.3, pp.672–682.

36) Paszkiewicz, K.H. et al.: Prediction of viable circular permutants using a graph theoretic approach, *Bioinformatics*, Vol.22, No.11, pp.1353–1358 (2006).

37) Kim, H.-J. and Kim, J. M.: Cyclic topology in complex networks, *Phys. Rev. E*, Vol.72, 036109 (2005).

38) Takemoto, K., et al.: Correlation between structure and temperature in prokaryotic metabolic networks, *BMC Bioinformatics*, Vol.8, 303 (2007).

39) Newman, M.E.: Assortative mixing in networks, *Phys. Rev. Lett.*, Vol.89, No.20, 208701 (2005).

40) Pettersen, E.F., et al.: UCSF Chimera- A Visualization System for Exploratory Research and Analysis, *J. Comput. Chem.*, Vol.25, No.13, pp.1605–1612.

41) Batagelj, V. and Mrvar, A.: Pajek — Analysis and Visualization of Large Networks, Lecture Notes in Computer Science, Vol.2265, pp.8–11 (2002).

42) Vapnik, V.: *The nature of statistical learning theory*, Springer, New York (2000).

43) Brown, M.P.S., et al.: Knowledge-based analysis of microarray gene expression data by using support vector machines, *Proc. Natl. Acad. Sci. USA*, Vol.97, No.1, pp.262–267 (2000).

44) Tamura, T. and Akutsu, T.: Subcellular location prediction of proteins using support vector machines with alignment of block sequences utilizing amino acid composition, *BMC Bioinformatics*, Vol.8, 466 (2007).

45) Bao, L. and Cui, Y.: Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information, *Bioinformatics*, Vol.21, No.10, pp.2185–2190 (2005).

46) Song, J. et al.: Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information, *BMC Bioinformatics*, Vol.7, 124 (2006).

47) Cheng, J. and Baldi, P.: A machine learning information retrieval approach to protein fold recognition, *Bioinformatics*, Vol.22, No.12, 1456–1463 (2006).

48) Cheng, J. and Baldi, P.: Improved residue contact prediction using support vector machines and a large feature set, *BMC Bioinformatics*, Vol.8, 113 (2007).

49) Vullo, A. and Frasconi, P.: Disulfide connectivity prediction using recursive neural networks and evolutionary information, *Bioinformatics*, Vol.20, No.5, pp.653–659 (2004).

50) Bradford, J.R. and Westhead, D.R.: Improved prediction of protein-protein binding sites using a support vector machines approach, *Bioinformatics*, Vol.21, No.8, pp.1487–1494 (2005).

51) Shen, J., et al.: Predicting protein-protein interactions based only on sequences information, *Proc. Natl. Acad. Sci. USA*, Vol.104, No.11, pp.4337–4441 (2007).

52) Brown, J.B. and Akutsu, T.: Identification of novel DNA repair proteins via primary sequence, secondary structure, and homology, *BMC Bioinformatics*, Vol.10, 25 (2005).

53) Yuan, Z. and Huang, B.: Prediction of protein accessible surface areas by support vector regression, *Proteins*, Vol.57, No.3, pp.558–564.

54) Yuan, Z.: Better prediction of protein contact number using a support vector regression analysis of amino acid sequence, *BMC Bioinformatics*, Vol.6, 248 (2005).

55) Ishida, T., Nakamura, S. and Shimizu, K.: Potential for assessing quality of protein structure based on contact number prediction, *Proteins*, Vol.64, No.4, pp.940–947 (2006).

56) Song, J., et al.: HSEpred: predict half-sphere exposure from protein sequences, *Bioinformatics*, Vol.24, No.13, pp.1489–1497 (2008).

57) Song, J. and Burrage, K.: Predicting residue-wise contact orders in proteins by support vector regression, *BMC Bioinformatics*, Vol.7, 425 (2006).

58) Rangwala, H., et al.: svmPRAT: SVM-based Protein Residue Annotation Toolkit, *BMC Bioinformatics*, Vol.10, 439 (2009).

59) Wang, X., et al.: Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme, *BMC Bioinformatics*, Vol.7, 32 (2006).

60) Song, J., et al.: Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure, *Bioinformatics*, Vol.23, No.23, pp.3147–3154 (2007).

61) Yuan, Z. and Wang, Z.X.: Quantifying the relationship of protein burying depth and sequence, *Proteins*, Vol.70, No.2, pp.509–516 (2008).

62) Zhang, H., et al.: Sequence based residue depth prediction using evolutionary information and predicted secondary structure, *BMC Bioinformatics*, Vol.9, 388 (2008).

63) Song, J., et al.: Prodepth: predict residue depth by support vector regression approach from protein sequences only, *PLoS ONE*, Vol.4, No.9, e7072 (2009).

64) Song, J., et al.: Cascleave: towards more accurate prediction of caspase substrate cleavage sites, *Bioinformatics*, Vol.26, No.6, pp.752–760 (2010).

65) Joachims, T.: Making large-Scale SVM Learning Practical, *Advances in Kernel Methods: Support Vector Learning* (Schölkopf, B., Burges, C. and Smola, A. (ed.)), MIT Press, Cambridge MA, pp.169–184 (1999).

66) Huang, J.T. and Cheng, J.P.: Differentiation between two-state and multi-state folding proteins based on sequence, *Proteins*, Vol.72, No.1, pp.44–49 (2008).

Appendix

Table 4 The average values of structural topology and complex network measures. The results are expressed as Mean±Standard Deviation.

Measures		Two-state	Multi-state	p-value
Topology	CO	0.261±0.076	0.216±0.076	0.005
	Abs_CO	22.94±10.04	27.40±9.60	0.036
	TCD	0.743±0.250	0.607±0.216	0.060
	LRO	1.379±0.573	1.400±0.440	0.862
	LR_CO	0.455±0.097	0.386±0.113	0.002
	FLC	0.596±0.115	0.599±0.096	0.122
	CTP	6.66±3.53	8.04±3.50	0.049
	L_{eff}	76.8±48.0	113.9±57.3	0.016
Network	CC_PCN	0.589±0.028	0.569±0.022	< 0.0001
	CC_LIN	0.297±0.108	0.305±0.084	0.174
	CYC_PCN	0.784±0.015	0.772±0.012	< 0.0001
	CYC_LIN	0.497±0.127	0.508±0.092	0.772
	TD_PCN	22.41±3.68	22.08±.3.31	0.229
	TD_LIN	5.37±2.69	5.21±2.17	0.480
	CPL_PCN	3.19±0.78	3.74±0.56	0.004
	CPL_LIN	8.90±1.44	5.51±1.21	0.159
	AC_PCN	0.247±0.099	0.274±0.095	0.005
	AC_LIN	0.367±0.105	0.449±0.106	0.002

(Received December 25, 2009)
(Accepted February 27, 2010)
(Released May 12, 2010)

(Communicated by *Tetsuo Shibuya*)

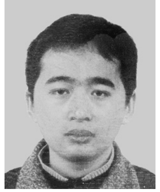
Jiangning Song was born in 1978. He received his B.Eng. in Biotechnology and his Ph.D. in Bioinformatics from Jiangnan University China in 2000 and 2005, respectively. From 2007 to 2009, he worked as a JSPS Research Fellow at Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan. He is currently an NHMRC Peter Doherty Fellow at the Department of Biochemistry and Molecular Biology, Faculty of Medicine, Monash University, Melbourne, Australia. His research interests include structural bioinformatics, machine learning, data mining and complex networks. He is a member of Japan Society for Bioinformatics (JSBi) and International Proteolysis Society (IPS).

Kazuhiro Takemoto was born in 1981. He received his Doctoral degree in Informatics from Kyoto University in 2008 after earned his Bachelor's and Master's degrees in Computer Science from Kyushu Institute of Technology in 2004 and 2006, respectively. He belonged to University of Tokyo as a Postdoctoral Fellow after served as a JSPS Research Fellow for two years from 2007. He is currently a JST PRESTO Researcher. His research interests are network science including theory and bioinformatics. In particular, he recently focuses on extremophilic metabolism. He is a member of JSBi.

Table 5 The selected feature sets and the corresponding parameter options for building SVR models.

Feature sets and SVR parameters		Topology	Network	Combined
Two-state	SVR option	-t 1 -d 2 -c 0.1	-t 1 -d 2 -c 1.4	-t 1 -d 2 -c 1.5
	R_d cutoff	C_α atom, $R_d = 8$Å	C_α atom, $R_d = 8$Å	C_α atom, $R_d = 8$Å
	Features	CO, TCD, LRO, FCL, CTP	CC_LIN, CYC_LIN, TD_PCN, TD_LIN, CPL_LIN	CO, TCD, LRO, FCL, CTP, CC_LIN, CYC_LIN, TD_PCN, TD_LIN
Multi-state	SVR option	-t 1 -d 2 -c 0.02	-t 1 -d 2 -c 0.02	-t 1 -d 2 -c 0.002
	R_d cutoff	Heavy atom, $R_d = 8$Å	Heavy atom, $R_d = 8$Å	Heavy atom, $R_d = 8$Å
	Features	Abs_CO, CTP, $\log_{10}$(Prolength), $\log_{10}(L_{\rm eff})$	CC_PCN, CYC_PCN, TD_LIN, CPL_PCN	Abs_CO, CTP, $\log_{10}$(Prolength), $\log_{10}(L_{\rm eff})$, CC_PCN, CYC_PCN, CPL_PCN
Overall	SVR option	-t 1 -d 2 -c 3.8	-t 1 -d 2 -c 0.7	-t 1 -d 2 -c 2.7
	R_d cutoff	C_α atom, $R_d = 8$Å	C_α atom, $R_d = 8$Å	C_α atom, $R_d = 8$Å
	Features	Abs_CO, LRO, FLC, CTP, $\log_{10}(L_{\rm eff})$	CC_PCN, CC_LIN, CYC_PCN, CYC_LIN, TD_PCN, TD_LIN	Abs_CO, LRO, CTP, $\log_{10}(L_{\rm eff})$, CC_PCN, CYC_PCN, CYC_LIN, TD_LIN

Hongbin Shen was born in 1979. He received his Doctoral degree in 2007 from Shanghai Jiaotong University. He was a postdoc research fellow of Harvard Medical School from 2007 to 2008. Currently, he is a professor of Institute of Image Processing and Pattern Recognition, Shanghai JiaoTong University. His current research interests include data mining and bioinformatics. He is particularly interested in the researches of predicting protein structure and functions as well as intelligent modeling for complex biological networks. Dr. Shen has published more than 60 papers and constructed 20 bioinformatics servers in these areas and he serves the editor members of several international journals.



Hao Tan was born in 1982. He received his B.Eng in Communication Engineering in Southeast University China in 2004 and his Master degree in Applied Information Technology from Monash University Australia in 2008. He is a research postgraduate student in the Faculty of Information Technology, Monash University Melbourne, Australia. His research interests are computer software design and bioinformatics.

M. Michael Gromiha received his Ph.D. in Physics from Bharathidasan University, Tiruchirappalli, India in 1994. He pursued his post doctoral research at the International Center for Genetic Engineering and Biotechnology (ICGEB), Trieste, Italy and The Institute of Physical and Chemical Research (RIKEN), Tsukuba, Japan. At present he is working as a Senior Research Scientist at the Computational Biology Research Center (CBRC) of the National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan. His main research interests are structural analysis, prediction, folding and stability of globular and membrane proteins, protein interactions and development of databases and online tools in Bioinformatics. He has published over 130 research articles, 30 reviews and written a book on "Protein Bioinformatics: From Sequence to Function". He is an Associate Editor of BMC Bioinformatics, Editor-in-Chief of Open Structural Biology Journal and a member of Nature Reader Panel. He is also a program committee member of ISMB and ECCB.

Tatsuya Akutsu received his M.Eng. degree in Aeronautics in 1996 and his Dr. Eng. degree in Information Engineering in 1989 both from University of Tokyo, Japan. From 1989 to 1994, he was with Mechanical Engineering Laboratory, Japan. He was an associate professor in Gunma University from 1994 to 1996 and in Human Genome Center, University of Tokyo from 1996 to 2001 respectively. He joined Bioinformatics Center, Institute for Chemical Research, Kyoto University, Japan as a professor in Oct. 2001. His research interests include bioinformatics and discrete algorithms. He is a member of ACM, IEICE, JSAI, and JSBi.