Vol.2010-BIO-21 No.31
2010/6/19

構造トポロジーと複雑ネットワーク特徴量からのタンパク質フォールディング速度予測

宋　江寧[†1,†2]　竹本和広[†3]　沈　紅斌[†4]
檀　浩[†2]　マイケル・グロミハ[†5]
阿久津達也[†1]

タンパク質のフォールディング速度を予測することはそのフォールディングを理解に向けて重要なステップのひとつである。私たちはタンパク質の三次元構造から得られる様々な構造トポロジーと複雑ネットワーク特徴量を併用した新規のフォールディング速度予測法を提案する。この提案手法は二状態と多状態タンパク質のフォールディング速度の予測において既存手法より高い精度を示す。この予測モデル（PRORATE）は http://sunflower.kuicr.kyoto-u.ac.jp/~sjn/folding/において利用可能である。

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]

Prediction of protein folding rate is an important step towards our further understanding of the protein folding mechanism. We develop a novel approach to predict protein folding rates, which combines a variety of structural topology and complex network properties calculated from protein three-dimensional (3D) structures. 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. 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 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. 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 have led to our improved understanding of the physical processes of protein folding and the fundamental rules governing protein folding behaviors.

Prediction of protein folding rate from its amino acid sequence is an important step towards our understanding of the protein folding mechanism[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)]. With the increasing availability

†1 京都大学 化学研究所 バイオインフォマティクスセンター
Bioinformatics Center, Institute for Chemical Research, Kyoto University
†2 Department of Biochemistry and Molecular Biology, Monash University
†3 科学技術振興機構 さきがけ
PRESTO, Japan Science and Technology Agency
†4 Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University
†5 産総研 生命情報工学研究センター
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology

of protein folding data deposited in public databases as the consequence of structural genomics projects, efficient computational tools are desired to 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. 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)[7], total contact distance (TCD)[8], long range order (LRO)[9], long range contact order (LR_CO)[10], effective secondary structure length (L_{eff})[4], the fraction of local contact (FLC)[11),12] and chain topology parameter (CTP)[13].

Graph theoretic approaches that model protein structures as connecting networks of interacting residues, from the perspective of complex networks[14),15], provide new insights into protein folding mechanism[?]. Moreover, a most-recent study indicates that complex network properties correlate with folding rates[16]. 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 this 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 approach. We combine a variety of structural topology parameters with complex network properties as the input features into the SVR models. 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. In this article, some details are omitted, which are given in the journal version of this article[17].

2. Methods

2.1 Datasets

We used a larger dataset that has been recently constructed by Ouyang and Liang[18]. It consists of 80 protein folders with their folding rates experimentally determined. Among them, 45 proteins exhibit two-state (TS) folding behaviors, while the other 35 proteins belong to the three-state or multi-state (MS) folding kinetics. 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 logarithmic folding rates of these proteins range from -6.9 to 12.9.

2.2 Structural Topology Measures

We utilized 8 structural topology properties (CO, Abs_CO, TCD, LRO, LR_CO, L_{eff}, FLC and CTP), mentioned in Sec. 1.

2.3 Complex Network Properties

We constructed PCNs and LINs based on protein 3D structures, and calculated the following graph-theoretic metrics: Clustering Coefficient (CC)[14], Cyclic Coefficient (CYC)[19], Triangle Density (TD)[20], Characteristic Path Length (CPL)[14] and Assortative Coefficient (AC)[21]. Henceforth, in addition, the complex network properties X obtained from PCNs and LINs are represented as X_PCN and X_LIN, respectively.

2.4 Support Vector Regression

To predict folding rates, the support vector regression (SVR) in SVM_light package[22] was utilized. We selected different combinations of optimal parameters of polynomial kernel functions to build the different SVR models.

2.5 Performance Evaluation

To evaluate the performance and avoid the over-fitting, we performed the back-check and the leave-one-out cross-validation (LOOCV) tests.

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).

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

There are two significant topology measures that show distinguishable preferences for the TS and MS proteins: CO and LR_CO. 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).

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 (**表 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 of the TS and MS proteins.

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. 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 (表 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.

表 1 Correlation coefficients between topology parameters, network properties and the corresponding folding rate $\ln K_f$ values. The results are computed with the traditional threshold $R_d = 8$Åusing the C_α atom for the TS proteins as the node and $R_d = 8$Åusing 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

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 resulting prediction performances are summarized in **表 2**.

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

表 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

the multi-state protein folding, the SVR classifier based on topology parameter provides better performance compared with that based on network properties (**図 1**). We argue that these results might be a reflection of the difference of folding mechanisms between the TS and MS protein folders. Moreover, after combining the topology and network properties, the resulting SVR classifier further improves the prediction accuracy.

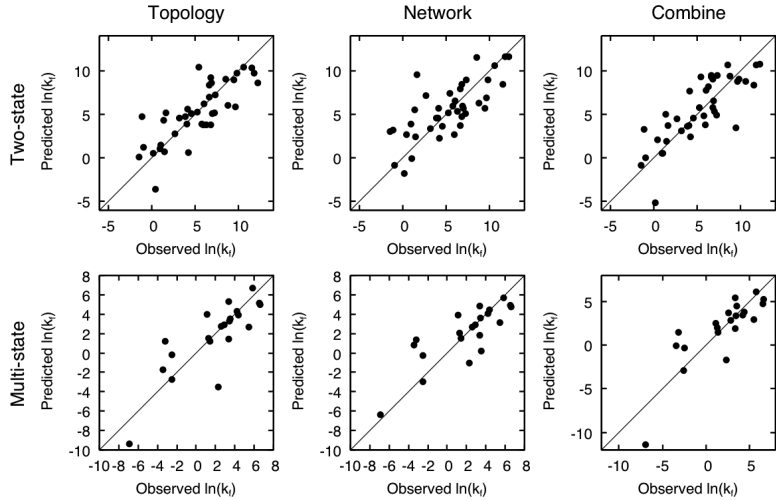

図 1 The scatter-plots of the observed and predicted folding rates of the TS and MS protein folders by the jack-knife cross-validation test.

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 TS or MS 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[23] and the composition-based predictor which is based on the differentiation of amino acid contents between the TS and MS folders[10]. To make an objective comparison, these methods are measured on the same training and test datasets. The result comparison is presented in **表 3**. As can be seen, the SVR classifier performs much better than the BLR method. The SVR classifier also compares favorably with the composition-based predictor. 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

表 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

the other hand, topology parameters are more indicative of the MS protein folders than the network 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[24].

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[16]. 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

We attempted to predict protein folding rates of proteins with TS and MS folding kinetics, by developing a multiple-feature framework based on 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. 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.

参 考 文 献

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) 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).
8) Zhou, H. and Zhou, Y.: Folding rate prediction using total contact distance, *Biophys. J.*, Vol.82, No.1, pp.458–463 (2002).
9) 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).
10) 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).
11) 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).
12) 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).
13) Nölting, B. et al.: Structural determinants of the rate of protein folding, *J. Theor. Biol.*, Vol.223, No.3, pp.299–307 (2003).
14) Watts, D.J. and Strogatz, S.H.: Collective dynamics of 'small-world' networks, *Nature*, Vol.393, No.6684, pp.440–442.
15) 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.
16) Bagler, G. and Sinha, S.: Assortative mixing in Protein Contact Networks and protein folding kinetics, *Bioinformatics*, Vol.23, No.14, pp.1760–1767 (2007).
17) Song, J. et al.: Prediction of protein folding rates from structural topology and complex network properties, *IPSJ Transactions on Bioinformatics*, in press.
18) 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).
19) Kim, H.-J. and Kim, J. M.: Cyclic topology in complex networks, *Phys. Rev. E*, Vol.72, 036109 (2005).
20) Takemoto, K. et al.: Correlation between structure and temperature in prokaryotic metabolic networks, *BMC Bioinformatics*, Vol.8, 303 (2007).
21) Newman, M.E.: Assortative mixing in networks, *Phys. Rev. Lett.*, Vol.89, No.20, 208701 (2005).
22) 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).
23) 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).
24) 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).