

國立交通大學

生物資訊研究所

碩士論文

蛋白質摺疊率的研究  
Study on Protein Folding Rates



研究生：葉書瑋

指導教授：黃鎮剛 教授

中華民國九十五年七月

# 蛋白質摺疊率的研究

學生：葉書瑋

指導教授：黃鎮剛

國立交通大學生物資訊研究所碩士班

## 摘要

蛋白質序列是一條由 20 種胺基酸所組成的線性結構，而這每一條蛋白質序列都可以對應到其特定的三維結構。蛋白質由序列到三維結構的過程稱之為「蛋白質摺疊」。蛋白質是如何摺疊為其特定的三維結構？序列和結構之間又有著什麼樣的關係存在？在生物科學的領域裡，研究尋找這關係的現象與作用一直以來都是相當重要的議題。我們試著藉由研究蛋白質摺疊率與序列結構的關係，來了解蛋白質的摺疊。不同蛋白質的蛋白質有著相當不一樣的摺疊率。通常比較小的蛋白質其摺疊所需花的時間往往比較大的蛋白質所需花的時間來要少。在本研究中，我們利用向量支持回歸 (Support Vector Regression) 作為主要的研究工具。在只使用序列資訊的情況下，結果和蛋白質摺疊率的相關性達 80% 左右。

# Study on Protein Folding Rates

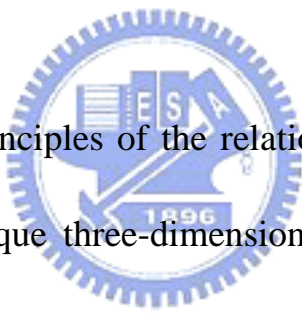
Student : So-Wei Yeh

Advisor : Jenn-Kang Hwang

Institute of Bioinformatics

National Chiao Tung University

## ABSTRACT

The logo of National Chiao Tung University is a circular emblem with a gear-like border. Inside the circle, there are stylized letters 'NCTU' and the year '1896' at the bottom.

Understanding the principles of the relationship between a primary amino acid sequence and its unique three-dimensional structures is one of the most important issues in biology science. A related and challenging task is to understand the relationship between sequences and folding rates of proteins. Proteins have different rates of folding. Small proteins usually fold faster than larger ones. We currently use amino acid sequences (which predicts properties such as protein secondary structure) as feature vectors to predict protein folding rates, using support vector regression in machine learning tool. Preliminary results show 80% correlation between the predicted and experimental folding rates.

## 誌謝

在生命的旅程上，不斷地求知與學習是每個階段必經的過程。相當幸運的可以在碩士班的階段，來到交通大學生物資訊所這個良好的環境來學習。在這裡不僅僅開拓了我的視野，也使我成長了不少。

首先要感謝的是我的指導老師黃鎮剛老師，提供一個自由獨立的研究環境。在這兩年的時間裡，老師不厭其煩地給予我許多寶貴的意見、指導與叮嚀。老師帶給我的除了科學研究的知識外，更重要的是在科學研究上的態度與精神。在此向老師致上最崇高的敬意。

接下來要感謝的是尤禎祥學長為我們建立一個穩定的實驗計算環境，使得我們可以安心無慮的在這環境中著手每一個實驗。也要感謝這兩年來實驗室的夥伴，簡思樸、張世瑜以及盧慧同學們在生活上還有研究上提供相當寶貴的意見彼此切磋。有大家在研究學問的路上扶持與鼓勵，才得以完成本研究。另外也感謝實驗室的所有學長姐和學弟妹，讓實驗室的生活能夠如此多彩多姿。

最後要感謝關心我的所有人，無時無刻地付出你們的關懷與鼓勵，使我得以順利地完成碩士班的學業。在此僅獻上此書，以表達我最深的謝意。

# CONTENTS

摘要 .....	i
ABSTRACT .....	ii
誌謝 .....	iii
Contents .....	iv
Table Contents .....	v
Figure Contents .....	vi
1 Introduction .....	1
2 Material and Method .....	4
2.1 Definition of folding rate .....	4
2.2 Dataset .....	4
2.2.1 Classification of two states and multi-sates proteins .....	4
2.2.2 Classification of protein secondary structure .....	5
2.3 Support vector regression .....	5
2.4 Cross validation method .....	6
2.5 Feature vectors .....	7
2.5.1 Sequence information .....	7
2.5.2 Structure information .....	7
2.6 Performance measure .....	8
3 Result and Discussion .....	9
3.1 Comparison with previous work .....	9
3.2 Comparison among different secondary structure coding features .....	9
3.3 Comparison among different protein classification with secondary structure coding features .....	9
3.4 Comparison among different protein classification with contact order ....	10
References .....	11

## TABLE CONTENTS

Table 1 List of 64 proteins .....	18
Table 2 Classification of 64 proteins .....	21
Table 3 List of feature vectors .....	22
Table 4 Comparison with Ivankov's results .....	23
Table 5 Comparison different predicted secondary structure features .....	24
Table 6 Comparison with different classification .....	25
Table 7 Comparison with different classification with contact order .....	26



## FIGURE CONTENTS

Figure 1	System flowchart .....	27
Figure 2	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (PSIPRED) with all dataset .....	28
Figure 3	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (ALB) with all dataset .....	29
Figure 4	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (DSSP) with all dataset .....	30
Figure 5	Correlation between experimental and predicted folding rates using $L+L_S+N_S$ (PSIPRED) with all dataset .....	31
Figure 6	Correlation between experimental and predicted folding rates using $L+L_C+N_C$ (PSIPRED) with all dataset .....	32
Figure 7	Correlation between experimental and predicted folding rates using $L+L_S+N_S$ (ALB) with all dataset .....	33
Figure 8	Correlation between experimental and predicted folding rates using $L+L_C+N_C$ (ALB) with all dataset .....	34
Figure 9	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (PSIPRED) with all- $\alpha$ protein .....	35
Figure 10	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (PSIPRED) with all- $\beta$ protein .....	36
Figure 11	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (PSIPRED) with mixed-class protein .....	37
Figure 12	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (PSIPRED) with two state protein .....	38
Figure 13	Correlation between experimental and predicted folding rates using $L+L_H+N_H$ (PSIPRED) with multi-state protein .....	39
Figure 14	Correlation between experimental and predicted folding rates using $L+L_S+N_S$ (PSIPRED) with all- $\alpha$ protein .....	40
Figure 15	Correlation between experimental and predicted folding rates using $L+L_S+N_S$ (PSIPRED) with all- $\beta$ protein .....	41
Figure 16	Correlation between experimental and predicted folding rates using $L+L_S+N_S$ (PSIPRED) with mixed-class protein .....	42
Figure 17	Correlation between experimental and predicted folding rates using $L+L_S+N_S$ (PSIPRED) with two state protein .....	43
Figure 18	Correlation between experimental and predicted folding rates using $L+L_S+N_S$ (PSIPRED) with multi-state protein .....	44
Figure 19	Correlation between experimental and predicted folding rates using $L+L_C+N_C$ (PSIPRED) with all- $\alpha$ protein .....	45
Figure 20	Correlation between experimental and predicted folding rates using $L+L_C+N_C$ (PSIPRED) with all- $\beta$ protein .....	46

Figure 21 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (PSIPRED) with mixed-class protein ..... 47

Figure 22 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (PSIPRED) with two state protein ..... 48

Figure 23 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (PSIPRED) with multi-state protein ..... 49





# 1 Introduction

All the proteins begin their existence on a ribosome as a linear polypeptide chain. It is known that a protein sequence can fold into its unique three-dimensional conformation to achieve the biologically active native state. A protein structure evolves to have function only in a particular cellular environment. This kind of results is due to the evolution. But how does a sequence can find its most stable structure exactly? Proteins may lose their own function when only a simple mistake happens during the folding process.

In recent years, to understand the principles of the relationship between a primary amino acid sequence and its unique three-dimensional structures become one of the most important challenges in biology science. A folded protein is stabilized by many specific interactions, as seen in an X-ray or NMR structure. When the protein is unfolded, the interactions and functions are lost.<sup>1</sup> An understanding of this fundamental process would help in attempts to predict structure from sequence, in the rational design of proteins de novo, and in understanding how and why proteins misfold.<sup>2</sup> Besides, to realize the issue of protein folding may be the molecular basis for a wide range of human genetic disorders. For example, cystic fibrosis is caused by defects in a membrane-bound protein called cystic fibrosis trans-membrane conductance regulator (CFTR), which acts as a channel for chloride ions.<sup>3</sup> It could lead to new therapies for this kind of diseases by an improvement of comprehending

protein folding. A related and challenging task is to understand the relationship between sequences and folding rates of proteins.<sup>4</sup>

Proteins have very different rates of folding. In general, small proteins usually fold faster than the large ones.<sup>5</sup> Many proteins with differing structures, stabilities and sequences, have been shown to fold with two-state kinetics or multi-state kinetics.<sup>2</sup> To be very similar, small proteins usually fold with two-state kinetics which are called two-state proteins; large proteins usually fold with multi-state kinetics which are called multi-state proteins. Two-state protein follows the simplest rule of protein folding. In this rule, it only contains two states, unfolded state and folded state. The rule of multi-state protein folding is more complicated than the two-state protein. Besides the unfolded state and the folded state, it contains various intermediate states. The intermediate states are between the unfolded state and the folded state. The number of intermediate states is based on the chemical and physical properties of a protein. If there is only one intermediate state between the unfolded and folded state, the protein is called three-stated protein.

There are a lot of factors which would affect the process of protein folding. One of them, which is relatively well understood, is the dependence of folding rate on temperature.<sup>36</sup> At very high temperature, protein conformations are usually tend to fold faster. On the other hand, proteins are tend to fold slower from the denature state to the native state at very low

temperature. Therefore, in order to reduce the influence from the temperature, all the folding rates of protein sequences we choose in this work are observed at the experimental environment of around 25°C. Many theoretical studies have found some important factors which correlated with the protein folding rates. Plexco et al. proposed the relative contact order (CO), which is the average sequence distance between all pairs of contacting residues normalized by the total sequence length.<sup>7</sup> Gromiha et al. define a parameter, long-range order (LRO) for a protein from the knowledge of long-range contacts (contacts between two residues that are close in space and far in the sequence) in protein structure.<sup>8</sup> Ivankov et al. emphasize the importance of effective chain length ( $L_{\text{eff}}$ ) for protein folding.<sup>5</sup>  $L_{\text{eff}}$  is a specific number of the chain residues. By this kind of studies, we can realize the complex process of protein folding more.



However, it comes an interesting issue which factors is the main determinant that affects the time of protein folding most. In this work, we build a method to predict the folding rate from the protein sequence and structure information. The first goal of this work is to compare the importance between each factor to understand the rules that govern protein folding. The second one is to develop a useful tool for predicting the protein folding rates from their sequence and structure information using support vector regression which is a novel machine learning skill.

## 2 Material and Method

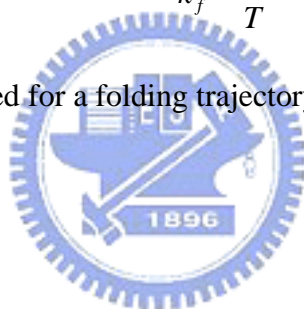
### 2.1 Definition of folding rate

The folding rate ( $k_f$ ) is a parameter which measures and describes how much time it takes in a protein folding process. It is an inverse of the time required for a folding process. The bigger folding rate is; the shorter time it takes. In this work, we use the logarithm of folding rate to do the experiments. The reason is based on the suggestion of both analytical theory<sup>9,10</sup> and off-lattice computer simulations<sup>11</sup> of folding.

The folding rate is given by

$$k_f = \frac{1}{T} \quad (1)$$

where  $T$  is the time required for a folding trajectory to reach the native conformation.



### 2.2 Dataset

Our dataset includes 64 proteins<sup>12-69</sup> that fold with two-state or multi-state kinetics which are shown in Table 1. The list including single-domain proteins and peptides that lack both disulfide bonds and covalent bonds to ligands is taken from Ivankov et al.<sup>5</sup> All chemical and physical properties including in-water folding rates are collected from the experimental literatures. If folding of some protein is investigated at different temperatures, we use the closest to 25°C. Structural properties of the proteins are obtained from the literatures and Protein Data Bank<sup>70</sup>.

#### 2.2.1 Classification of two states and multi-states proteins

The primary criterion for the classification of a given protein as either a two state or a multi-state protein is considered to be whether the folding kinetics is single-exponential or

multi-exponential<sup>71</sup>. However changing the environment condition (e.g. temperature or pH), a protein may switch from two state folding to multi-state folding and contrariwise. The classification of folding mechanism of each protein in our dataset is based on the experiment.

The 64 proteins in our dataset includes 37 two state proteins, 25 multi-state proteins and two small artificial peptides which are shown in Table 2.

## 2.2.2 Classification of protein secondary structure

The rule of protein secondary structure classification is based on the SCOP Classification (version 1.69). We obtain from the experimental literatures and the website of Protein Data Bank. (<http://www.rcsb.org/pdb/Welcome.do>) The 64 proteins in our dataset includes 15 all alpha proteins, 18 all beta proteins and 31 mixed-class proteins which includes  $\alpha + \beta$  and  $\alpha / \beta$  proteins. Details are shown in Table 2.

## 2.3 Support vector regression

Support Vector Machines (SVM) was developed by Vapnik<sup>72</sup> to solve the classification problem, but recently, SVM have been successfully extended to regression and density estimation problem<sup>73</sup>. The support vector regression (SVR) is a powerful regression method that has become popular in computational biology. The original idea of SVR like the traditional linear regression is to solve a linear function given training data  $(x_1, y_1), \dots, (x_n, y_n)$ . SVR maps the data to a high dimensional space by a function  $\Phi(x)$  and avoids the under-fitting and over-fitting problems of the training data by minimizing the training error.

Support vector regression proceeds two modifications to avoid over-fitting problems. The first one is to give a threshold  $\varepsilon$  so that if the  $i$ th data satisfies the followed equation:  $-\varepsilon \leq y_i - (w^T \phi(x_i) + b) \leq \varepsilon$ , it is considered a correct approximation. Then  $\xi_i = \xi_i^* = 0$ . The second one is to smooth the function  $w^T \phi(x_i) + b$ , an additional term  $w^T w$  is added to the objective function. Clearly,  $\xi_i$  is called the upper training error ( $\xi_i^*$  is the lower) subject to

the  $\varepsilon$ -insensitive tube  $|y_i - (w^T \phi(x_i) + b)| \leq \varepsilon$ . If  $x_i$  is not in the tube, there is an error  $\xi_i$  or  $\xi_i^*$  which we would like to minimize in the objective function. SVR avoids underfitting and overfitting the training data by minimizing the training error  $C \sum_{i=1}^l \xi_i + \xi_i^*$  as well as the regularization term  $\frac{1}{2} w^T w$ . The addition of the term  $w^T w$  can be explained by a similar way to that for classification problems.

This article uses LIBSVM<sup>74</sup> as computing tools to perform all the calculations. The version of LIBSVM is 2.8. In the SVR training procedure, it is necessary to use cross-validation to find the best parameter  $C$ ,  $\gamma$  and  $p$  for RBF kernel. The LIBSVM provides a tool called gridregression to find the best parameters. The prediction result is correlated with these three parameters. Using the wrong parameters may generate worst prediction result. We use leave-one-out cross-validation to do the experiment. The system flowchart is illustrated in Figure 1.



## 2.4 Cross validation method

In order to check the performance and the efficiency of prediction methods, the method is often developed by cross-validation method or jack-knife method. In the cross-validation method, the datasets are divided into  $N$  groups for  $N$  fold cross-validation. One of each group would be the testing set and the other  $N-1$  groups would be the training set. This process is repeated by  $N$  times. Every group would be the testing set by turns. The final prediction results would be averaged over  $N$  testing sets. If the number of groups  $N$  equals the size of the whole dataset, it is called jack-knife method or leave-one-out cross-validation method. In this study, a leave-one-out cross-validation technique is used. One protein is removed from the whole dataset. The training is done on the remaining 63 proteins and the testing is done on the removed protein. This process is repeated 64 times by removing each protein in turn.

## 2.5 Feature vectors

Several different input features for SVR are considered in our experiments. They are including sequence and structure information. After combining and comparing these feature vectors by different protein classification, we use SVR to generate the prediction results. Details are shown in Table 3.

### 2.5.1 Sequence information

In this part, 10 major feature vectors are selected to do the experiments. They are sequence length ( $L$ ), number of residues in helical conformation ( $L_H$ ), number of helices ( $N_H$ ), number of residues in strand conformation ( $L_S$ ), number of strands ( $N_S$ ), number of residues in coil conformation ( $L_C$ ), and number of coil ( $N_C$ ). The information of secondary structure is predicted by ALB<sup>75</sup> (<http://i2o.protres.ru/alb>) and PSIPRED<sup>76</sup> (<http://www.psipred.net>).

The residues predicted as helical are marked by H by PSIPRED and by H and & by ALB, and those predicted as strand are marked by E by PSIPRED and by S and B by ALB, and those predicted as coil are marked by C by PSIPRED.

### 2.5.2 Structure information

Relative contact order<sup>4</sup> (RCO) and absolute contact order<sup>77</sup> (ACO) are collected as feature vectors in this work. They are define by Plaxco et al. and Ivankov et al respectively.

The RCO is given by

$$RCO = \frac{1}{L \times N} \sum^N \Delta L_{ij} \quad (2)$$

Where  $N$  is the number of contacts within a cutoff of 6 angstrom between non-hydrogen atoms in the protein,  $L$  is the length of protein in amino acid residues, and  $\Delta L_{ij}$  is the number of residues separating the interacting pair of non-hydrogen atoms. For example, adjacent residues are assumed to be separated by one residue.

The ACO is given by

$$ACO = \frac{1}{N} \sum^N \Delta L_{ij} = RCO \times L \quad (3)$$

## 2.6 Performance measure

The performance is measure by Correlation Coefficient (CC). The value of CC is between -1 and 1. When the value is closer to 1 or -1, it means there is a stronger positive or negative correlation between two variables respectively.

The CC is given by

$$CC = \frac{Cov(X_1, X_2)}{\sigma_1 \sigma_2} \quad (4)$$

Where  $Cov(X_1, X_2)$  is the covariance of two random variables,  $X_1$  and  $X_2$ ,  $\sigma_1$  and  $\sigma_2$  are the standard deviations of sample 1 and 2.





## 3 Result and Discussion

### 3.1 Comparison with previous work

We use the same coding scheme as Ivankov's. Instead of using their mathematical formula which is based on knowledge, we use SVR to perform the prediction to verify whether there is a correlation between the feature vectors and the folding rates. Our results achieve a 0.86 correlation coefficient compared to 0.82 in the feature vector of PSIPRED and achieve a 0.86 correlation coefficient compared to 0.78 in the feature vector of ALB and achieve a 0.86 correlation coefficient compared to 0.81 in the feature vector of DSSP. Our results are all better than the Ivankov's. The comparison is listed in Table 4. The correlation coefficient plots are showed as Figure 2 to Figure 4.

### 3.2 Comparison among different secondary structure coding features

The features used to encode include the helical, strand, and coil conformation predicted by PSIPRED and ALB respectively. The best result is given by the helical conformation. The features predicted by PSIPRED and ALB are almost have the same performance. The results are listed in Table 5. The correlation coefficient plots are showed as Figure 2 to Figure 3 and Figure 5 to Figure 8.

### 3.3 Comparison among different protein classification with different secondary structure coding features

Besides the classify the protein into two state protein and multi-state protein, we also do the classification by the secondary structure which are All- $\alpha$  protein, All- $\beta$  protein and Mixed-class protein. Different features play the different abilities. The performances of helical conformation divided into several groups are not as good as training and predicting using the whole dataset. It seems that the coil conformance play an important role in the dataset

grouping by secondary structure. And the strand conformation get the better performance when the dataset classified by the mechanism of folding. The results are listed in Table 6. The correlation coefficient plots are showed as Figure 2 to Figure 3 and Figure 5 to Figure 23.

### **3.4 Comparison among different protein classification with contact order**

Comparing the results between RCO and ACO, we can tell RCO has the better performance in multi-state proteins and ACO has the better result in two state proteins. The results are listed in Table 7.



## References

1. Zhou HX. How do Biomolecular Systems Speed Up and Regulate Rate? *Phys. Biol.* 2005;2:R1-R25.
2. Jackson SE. How do Small Single-domain Proteins Fold? *Folding & Design* 1998;3:R81-R91.
3. Nelson DL, Cox MM. *Lehninger Principles of Biochemistry*, 4<sup>th</sup> Edition, W.H. Freeman and Company, New York, 2005.
4. Zhang L, Li J, Jiang Z, Xia A. Folding Rate Prediction Based on Neural Network Model. *Polymer* 2003;44:1751-1756.
5. Ivankov DN, Finkelstein AV. Prediction of Protein Folding Rates From the Amino Acid Sequence-predicted Secondary Structure. *PNAS* 2004;101(24):8942-8944.
6. Gutin AM, Abkevich VI, Shakhnovich EI. Chain Length Scaling of Protein Folding Time. *Phys. Rev. Lett.* 1996;77:5433.
7. Plaxco KW, Simons KT, Baker D. Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins. *JMB* 1998;277:985-994.
8. Gromiha MM, Selvaraj S. Comparison between Long-range Interactions and Contact Order in Determining the Folding Rate of Two-state Proteins: Application of Long-range Order to Folding Rate Prediction. *JMB* 2001;310:27-32.
9. Thirumalai D. From Minimal Models to Real Proteins: Time Scales for Protein Folding Kinetics. *J. Phys.* 1995;5:1457-1469.
10. Finkelstein AV, Badretdinov AY. Rate of Protein Folding near the Point of Thermodynamic Equilibrium between the Coil and the Most Stable Chain Fold. *Folding & Design* 1997;2:115.
11. Koga N, Takada S. Roles of Native Topology and Chain-length Scaling in Protein Folding: A Simulation Study with a Go-like Model. *JMB* 2001;313:171.
12. Munoz, V., Thompson, P. A., Hofrichter, J. & Eaton, W. A. Folding Dynamics and Mechanism of  $\beta$ -hairpin Formation. *Nature* 1997;390:196-199.

13. Qui, L., Pabit, S. A., Roitberg, A. E. & Hagen, S. J. Smaller and Faster: The 20-Residue Trp-Cage Protein Folds in 4  $\mu$ s. *JACS* 2002;124:12952-12953.
14. Thompson, P. A., Eaton, W. A. & Hofrichter, Laser Temperature Jump Study of the Helix $\leftrightarrow$ Coil Kinetics of an Alanine Peptide Interpreted with a 'Kinetic Zipper' Model. *J. Biochemistry* 1997;36:9200-9210.
15. Jager, M., Nguyen, H., Crane, J. C., Kelly, J. W. & Gruebele, M. The Folding Mechanism of a  $\beta$ -Sheet: The WW Domain. *JMB* 2001;311:373-393.
16. Islam, S. A., Karplus, M. & Weaver, D.L. Application of the Diffusion-Collision Model to the Folding of Three-helix Bundle Proteins. *JMB* 2002;318:199-215.
17. Spector S, Raleigh DP. Submillisecond Folding of the Peripheral Subunit-binding Domain. *JMB* 1999;293:763-768.
18. Kuhlman, B, Luisi DL, Evans PA, Raleigh DP. Global Analysis of the Effects of Temperature and Denaturant on the Folding and Unfolding Kinetics of the N-terminal Domain of the Protein L9. *JMB* 1998;284:1661-1670.
19. McCallister EL, Alm E, Baker D. Critical Role of  $\beta$ -hairpin Formation in Protein G Folding. *Nat. Struct. Biol.* 2000;7:669-673.
20. Myers JK, Oas TG. Preorganized Secondary Structure as an Important Determinant of Fast Protein Folding. *Nat. Struct. Biol.* 2001;8:552-558.
21. Mayor U, Johnson CM, Daggett V, Fersht AR. Protein Folding and Unfolding in Microseconds to Nanoseconds by Experiment and Simulation. *PNAS.* 2000;97:13518-13522.
22. Viguera AR, Serrano L, Wilmanns M, Different Folding Transition States May Result in the Same Native Structure. *Nat. Struct. Biol.* 1996;3:874-880.
23. Kim DE, Fisher C, Baker D. A Breakdown of Symmetry in the Folding Transition State of Protein L. *JMB* 2000;298:971-984.
24. Guerois R, Serrano L. The SH3-fold Family: Experimental Evidence and Prediction of Variations in the Folding Pathways. *JMB* 2000;304:967-982.
25. Grantcharova VP, Baker D, Folding Dynamics of the src SH3 Domain. *Biochemistry*

- 1997;36:15685-15692.
26. Jackson SE, Fersht AR, Folding of Chymotrypsin Inhibitor 2. 1. Evidence for a Two-State Transition. *Biochemistry* 1991;30:10428-10435.
  27. Perl D, Welker C, Schindler T, Schroder K, Marahiel MA, Jaenicke R, Schmid FX. Conservation of Rapid Two-state Folding in Mesophilic, Thermophilic and Hyperthermophilic Cold Shock Proteins. *Nat. Struct. Biol.* 1998 5:229-235.
  28. Plaxco KW, Guijarro JI, Morton CJ, Pitkeathly M, Campbell ID, Dobson CM. The Folding Kinetics and Thermodynamics of the Fyn-SH3 Domain. *Biochemistry* 1998 37:2529-2537.
  29. Schindler T, Herrler M, Marahiel MA, Schmid FX. Extremely Rapid Protein Folding in the Absence of Intermediates. *Nat. Struct. Biol.* 1995;2: 663-673.
  30. Reid KL, Rodriguez HM, Hillier BJ, Gregoret LM. Stability and Folding Properties of a Model P-sheet Protein, *Escherichia coli* CspA. *Protein Sci.* 1998;7: 470-479.
  31. Laurents DV, Corrales S, Elias-Arnanz M, Sevilla P, Rico M, Padmanabhan S. Folding Kinetics of Phage 434 Cro Protein. *Biochemistry* 2000;39:13963-13973.
  32. Khorasanizadeh S, Peters ID, Roder H. Evidence for a Three-state Model of Protein Folding from Kinetic Analysis of Ubiquitin Variants with Altered Core Residues. *Nat. Struct. Biol.* 1996;3:193-205.
  33. Burton RE, Huang GS, Daugherty MA, Fullbright PW, Oas TG. Microsecond Protein Folding Through a Compact Transition State. *JMB* 1996;263:311-322.
  34. Villegas V, Azuaga A, Catusas L, Reverter D, Mateo PL, Aviles FX, Serrano L. Evidence for a Two-state Transition in the Folding Process of the Activation Domain of Human Procarboxypeptidase A2. *Biochemistry* 1995;34:15105-15110.
  35. Van Nuland NA, Meijberg W, Warner J, Forge V, Scheek RM, Robillard GT, Dobson C M. Slow Cooperative Folding of a Small Globular Protein HPr. *Biochemistry* 1998;37:622-637.
  36. Kragelund BB, Robinson CV, Knudsen J, Dobson CM, Poulsen FM. Folding of a Four-helix Bundle: Studies of Acyl-Coenzyme A Binding Protein. *Biochemistry*

- 1995;34:7217-7224.
37. Ferguson N, Capaldi AP, James R, Kleantous C, Radford SE. Rapid Folding with and without Populated Intermediates in the Homologous Four-helix Proteins Im7 and Im9. *JMB* 1999;286:1597-1608.
  38. Fowler SB, Clarke J. Mapping the Folding Pathway of an Immunoglobulin Domain: Structural Detail from Phi Value Analysis and Movement of the Transition State. *Structure* 2001;9:355-366.
  39. Schreiber G, Fersht AR. The Refolding of cis- and trans-Peptidylprolyl Isomers of Barstar. *Biochemistry* 1993;32:11195-11203.
  40. Plaxco KW, Spitzfaden C, Campbell ID, Dobson CM. A Comparison of the Folding Kinetics and Thermodynamics of Two Homologous Fibronectin Type III Modules. *JMB* 1997;270:763-770.
  41. Clarke J, Hamill SJ, Johnson CM. Folding and Stability of a Fibronectin Type III Domain of Human Tenascin. *JMB* 1997;270:771-778.
  42. Guijarro JI, Morton CJ, Plaxco KW, Campbell ID, Dobson CM. Folding Kinetics of the SH3 Domain of PI3 Kinase by Real-time NMR Combined with Optical Spectroscopy. *JMB* 1998;276:657-667.
  43. Calloni G, Taddei N, Plaxco KW, Ramponi G, Stefani M, Chiti F. Comparison of the Folding Processes of Distantly Related Proteins. Importance of Hydrophobic Content in Folding *JMB* 2003;330:577-591.
  44. Clarke J, Cota E, Fowler SB, Hamill SJ. Folding Studies of Immunoglobulin-like  $\beta$ -sandwich Proteins Suggest that They Share a Common Folding Pathway. *Struct. Folding Des.* 1999;7:1145-1153.
  45. Cota E, Clarke J. Folding of Beta-sandwich proteins: Three-state Transition of a Fibronectin Type III Module. *Protein Sci.* 2000;9:112-120.
  46. Van Nuland NA, Chiti F, Taddei N, Raugei G, Ramponi G, Dobson CM. Slow Folding of Muscle Acylphosphatase in the Absence of Intermediates. *JMB* 1998;283:883-891.
  47. Taddei N, Chiti F, Paoli P, Fiaschi T, Bucciantini M, Stefani M, Dobson CM, Ramponi G.

- Thermodynamics and Kinetics of Folding of Common-Type Acylphosphatase: Comparison to the Highly Homologous Muscle Isoenzyme. *Biochemistry* 1999;38:2135-2142.
48. Parker MJ, Dempsey CE, Lorch M, Clarke AR. Acquisition of Native  $\beta$ -Strand Topology During the Rapid Collapse Phase of Protein Folding. *Biochemistry* 1997;36:13396-13405.
49. Otzen DE, Oliveberg M. Salt-induced Detour through Compact Regions of the Protein Folding Landscape. *PNAS* 1999;96:11746-11751.
50. Silow M, Oliveberg M. High-Energy Channeling in Protein Folding. *Biochemistry* 1997;36:7633-7637.
51. Wittung-Stafshede P, Lee JC, Winkler JR, Gray HB. Cytochrome b562 folding triggered by electron transfer: Approaching the Speed Limit for Formation of a Four-helix-bundle Protein. *PNAS* 1999;96:6587-6590.
52. Main ER, Fulton KF, Jackson SE. Folding Pathway of FKBP12 and Characterisation of the Transition State. *JMB* 1999 291:429-444.
53. Matouschek A, Kellis JT Jr, Serrano L, Bycroft M, Fersht AR. Transient Folding Intermediates Characterized by Protein Engineering. *Nature* 1990;346:440-445.
54. Schymkowitz JW, Rousseau F, Irvine LR, Itzhaki LS. The Folding Pathway of the Cell-cycle Regulatory Protein p13suc1: Clues for the Mechanism of Domain Swapping. *Struct. Folding Des.* 2000;8:89-100.
55. Choe SE, Matsudaira PT, Osterhout J, Wagner G, Shakhnovich EI. Folding Kinetics of Villin 14T, a Protein Domain with a Central  $\beta$ -Sheet and Two Hydrophobic Cores. *Biochemistry* 1998;37:14508-14518.
56. Dalessio PM, Ropson IJ.  $\beta$ -Sheet Proteins with Nearly Identical Structures Have Different Folding Intermediates. *Biochemistry* 2000;39:860-871.
57. Munoz V, Lopez EM, Jager M, Serrano L. Kinetic Characterization of the Chemotactic Protein from *Escherichia coli*, CheY. Kinetic Analysis of the Inverse Hydrophobic Effect. *Biochemistry* 1994;33:5858-5866.

58. Burns LL, Dalessio PM, Ropson IJ. Folding Mechanism of Three Structurally Similar  $\beta$ -Sheet Proteins. *Proteins* 1998;33:107-118.
59. Cavagnero S, Dyson HJ, Wright PE. Effect of H Helix Destabilizing Mutations on the Kinetic and Equilibrium Folding of Apomyoglobin. *JMB* 1999;285:269-282.
60. Golbik R, Zahn R, Harding SE, Fersht AR. Thermodynamic Stability and Folding of GroEL Minichaperones. *JMB* 1998;276:505-515.
61. Parker MJ, Marqusee S. The Cooperativity of Burst Phase Reactions Explored. *JMB* 1999;293:1195-1210.
62. Tang KS, Guralnick BJ, Wang WK, Fersht AR, Itzhaki LS. Stability and Folding of the Tumour Suppressor Protein p16. *JMB* 1999;285:1869-1886.
63. Jennings PA, Finn BE, Jones BE, Matthews CR. A Reexamination of the Folding Mechanism of Dihydrofolate Reductase from *Escherichia coli*: Verification and Refinement of a Four-channel Model. *Biochemistry* 1993;32:3783-3789.
64. Ikura T, Hayano T, Takahashi N, Kuwajima K. Fast Folding of *Escherichia coli* Cyclophilin A: A Hypothesis of a Unique Hydrophobic Core with a Phenylalanine Cluster. *JMB* 2000;297:791-802.
65. Parker MJ, Spencer J, Clarke AR. An Integrated Kinetic Analysis of Intermediates and Transition States in Protein Folding Reactions. *JMB* 1995;253:771-786.
66. Parker MJ, Sessions RB, Badcoe IG, Clarke AR. The Development of Tertiary Interactions during the Folding of a Large Protein. *Folding Des.* 1996;1:145-156.
67. Ogasahara K, Yutani K. Unfolding-refolding Kinetics of Tryptophan Synthase  $\alpha$  Subunit by CD and Fluorescence Measurements. *JMB* 1994;236:1227-1240.
68. Jones K, Wittung-Stafshede P. The Largest Protein Observed To Fold by Two-State Kinetic Mechanism Does Not Obey Contact-Order Correlation. *J. Am. Chem. Soc.* 2003;125:9606-9607.
69. Goldberg ME, Semisotnov GV, Friguet B, Kuwajima K, Ptitsyn OB, Sugai S. An Early Immunoreactive Folding Intermediate of the Tryptophan Synthase  $\beta_2$  Subunit is a 'Molten Globule'. *FEBS Lett.* 1990;263:51-56.



70. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research*. 2000;28:235-242.
71. Kamagata Kiyoto, Arai Munehito, Kuwajima Kunihiro. Unification of the Folding Mechanisms of Non-two-state and Two-state Proteins. *JMB* 2004;339:951-965.
72. Vapnik V, *The Nature of Statistical Learning Theory*, Second Edition, Springer, New York, 2001.
73. Farag A, Mohamed R, *Regression Using Support Vector Machine: Basic Foundations*. Technical Report. 2004;1-17.
74. Chang CC, Lin CJ, *LIBSVM: a Library for Support Vector Machines*, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
75. Ptitsyn OB, Finkelstein AV. Theory of Protein Secondary Structure and Algorithm of Its Prediction. *Biopolymers*. 1983;22:15-25.
76. Jones DT. Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *JMB* 1999;292:195-202.
77. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact Order Revisited: Influence of Protein Size on the Folding Rate. *Protein Science* 2003;12:2057-2062.

## Tables

**Table 1 List of 64 proteins**

PDB_id	Length	Log( $k_f$ )	State	SS	ref
1PGB	16	5.2	peptide	All- $\beta$	12
1L2Y	20	5.4	two-state protein	All- $\alpha$	13
-	21	6.7	peptide	All- $\alpha$	14
1PIN	34	4.1	two-state protein	All- $\beta$	15
1VII	36	5	two-state protein	All- $\alpha$	16
2PDD	41	4.3	two-state protein	All- $\alpha$	17
1DIV	56	2.6	two-state protein	Mix	18
1PGB	57	2.6	two-state protein	Mix	19
1BDD	58	5.1	two-state protein	All- $\alpha$	20
1ENH	61(54)	4.6	two-state protein	All- $\alpha$	21
1SHG	62(58)	0.6	two-state protein	All- $\beta$	22
1HZ6	62	1.8	two-state protein	Mix	23
1C8C	64	3	two-state protein	Mix	24
1SRL	64(55)	1.7	two-state protein	All- $\beta$	25
2CI2	64	1.7	two-state protein	Mix	26
1C9O	66	3.1	two-state protein	All- $\beta$	27
1G6P	66	2.7	two-state protein	All- $\beta$	27
1SHF	67	2	two-state protein	All- $\beta$	28
1CSP	67	2.9	two-state protein	All- $\beta$	27
1PSF	69	1.4	two-state protein	All- $\beta$	29
1MJC	69	2.3	two-state protein	All- $\beta$	30
2CRO	71(64)	1.6	multi-state protein	All- $\alpha$	31

**Table 1 List of 64 proteins (Continued)**

PDB_id	Length	Log( $k_f$ )	State	SS	ref
1UBQ	76	2.6	multi-state protein	Mix	32
1LMB	80	3.7	two-state protein	All- $\alpha$	33
1AYE	80	3	two-state protein	Mix	34
1POH	85	1.2	two-state protein	Mix	35
2ABD	86	2.9	two-state protein	All- $\alpha$	36
1IMQ	86	3.2	two-state protein	All- $\alpha$	37
1CEI	87	2.5	multi-state protein	All- $\alpha$	37
1TIT	89	1.6	multi-state protein	All- $\beta$	38
1BRS	89	1.5	multi-state protein	All- $\beta$	39
1FNF	90	-0.4	two-state protein	All- $\beta$	40
1TEN	90	0.5	two-state protein	All- $\beta$	41
1PNJ	90(84)	-0.5	two-state protein	All- $\beta$	42
1GXT	91(88)	1.9	multi-state protein	Mix	43
1WIT	93	0.2	two-state protein	All- $\beta$	44
1FNF	94	2.4	multi-state protein	All- $\beta$	45
1APS	98	-0.7	two-state protein	Mix	46
2ACY	98	0.4	two-state protein	Mix	47
1HNG	98(97)	0.8	multi-state protein	All- $\beta$	48
1RIS	101(97)	2.6	two-state protein	Mix	49
1URN	102(96)	2.5	two-state protein	Mix	50
256B	106	5.3	two-state protein	All- $\alpha$	51
1FKB	107	0.7	two-state protein	Mix	52
1BNI	110	1.1	multi-state protein	Mix	53
1SCE	113	1.8	multi-state protein	Mix	54
2VIK	126	3	two-state protein	Mix	55

**Table 1 List of 64 proteins (Continued)**

PDB_id	Length	Log(k <sub>f</sub> )	State	SS	ref
1EAL	127	0.6	multi-state protein	Mix	56
3CHY	129(128)	0.4	multi-state protein	Mix	57
1IFC	131	1.5	multi-state protein	Mix	58
1OPA	133	0.6	multi-state protein	Mix	58
1CBI	136	-1.4	multi-state protein	Mix	58
1A6N	151	0.5	multi-state protein	All- $\alpha$	59
1AON	155	0.3	multi-state protein	Mix	60
2RN2	155	0	multi-state protein	Mix	61
2A5E	156	1.5	multi-state protein	All- $\alpha$	62
1RA9	159	2	multi-state protein	Mix	63
1LOP	164	2.9	two-state protein	Mix	64
2LZM	164	1.8	multi-state protein	Mix	61
1PHP	175	1	multi-state protein	Mix	65
1PHP	219	-1.5	multi-state protein	Mix	66
1QOP	268	-1.1	multi-state protein	Mix	67
1L8W	341	0.7	two-state protein	All- $\alpha$	68
1QOP	396	-3	multi-state protein	Mix	69

**Table 2 Classification of 64 proteins**

Classification	Two-state	Multi-state	Peptides <sup>a</sup>	Total
All- $\alpha$	10	4	1	15
All- $\beta$	13	4	1	18
Mixed-class	14	17	-	31
Total	37	25	2	64

<sup>a</sup> Artificial peptides  $\alpha$ -helix and  $\beta$ -hairpin

<sup>b</sup> Structural properties of the proteins are obtained from Protein Data Bank

<sup>c</sup> All chemical and physical properties are collected from the literatures<sup>12-69</sup>



**Table 3 List of feature vectors**

<b>Feature vectors</b>	<b>Description</b>
L+L <sub>H</sub> +N <sub>H</sub> ( PSIPRED )	Helical conformation predicted by PSIPRED
L+L <sub>H</sub> +N <sub>H</sub> ( ALB )	Helical conformation predicted by ALB
L+L <sub>H</sub> +N <sub>H</sub> ( DSSP )	Helical conformation predicted by DSSP
L+L <sub>S</sub> +N <sub>S</sub> ( PSIPRED )	Strand conformation predicted by PSIPRED
L+L <sub>S</sub> +N <sub>S</sub> ( ALB )	Strand conformation predicted by ALB
L+L <sub>C</sub> +N <sub>C</sub> ( PSIPRED )	Coil conformation predicted by PSIPRED
L+L <sub>C</sub> +N <sub>C</sub> ( ALB )	Coil conformation predicted by ALB
RCO	Relative contact order
ACO	Absolute contact order



**Table 4 Comparison with Ivankov's results**

Input Feature Vectors	Ivankov et al	This Work
L+L <sub>H</sub> +N <sub>H</sub> ( PSIPRED )	0.82	0.86
L+L <sub>H</sub> +N <sub>H</sub> ( ALB )	0.78	0.86
L+L <sub>H</sub> +N <sub>H</sub> ( DSSP )	0.81	0.86



**Table 5 Comparison different predicted secondary structure features**

Input Features Vectors	Ivankov et al	This Work
L+L <sub>H</sub> +N <sub>H</sub> ( PSIPRED )	0.82	0.86
L+L <sub>S</sub> +N <sub>S</sub> ( PSIPRED )	-	0.71
L+L <sub>C</sub> +N <sub>C</sub> ( PSIPRED )	-	0.74
L+L <sub>H</sub> +N <sub>H</sub> ( ALB )	0.78	0.86
L+L <sub>S</sub> +N <sub>S</sub> ( ALB )	-	0.71
L+L <sub>C</sub> +N <sub>C</sub> ( ALB )	-	0.72





**Table 6 Comparison with different classification**

Class (number)	L+L <sub>H</sub> +N <sub>H</sub> (PSIPRED)	L+L <sub>S</sub> +N <sub>S</sub> (PSIPRED)	L+L <sub>C</sub> +N <sub>C</sub> (PSIPRED)
All- $\alpha$ proteins(15)	0.64	0.72	0.75
All- $\beta$ proteins(18)	0.76	0.77	0.82
Mixed-class proteins(31)	0.51	0.52	0.81
Two-state proteins(37)	0.77	0.87	0.61
Multi-state proteins(25)	0.68	0.68	0.70
All(64)	0.86	0.71	0.74



**Table 7 Comparison with different classification**

Class (number)	RCO	ACO
All- $\alpha$ proteins(15)	-0.87	0.75
All- $\beta$ proteins(18)	-0.83	0.80
Mixed-class proteins(31)	0.56	0.52
Two-state proteins(37)	0.56	0.78
Multi-state proteins(25)	0.75	0.64
All(64)	-0.78	0.72



# Figures

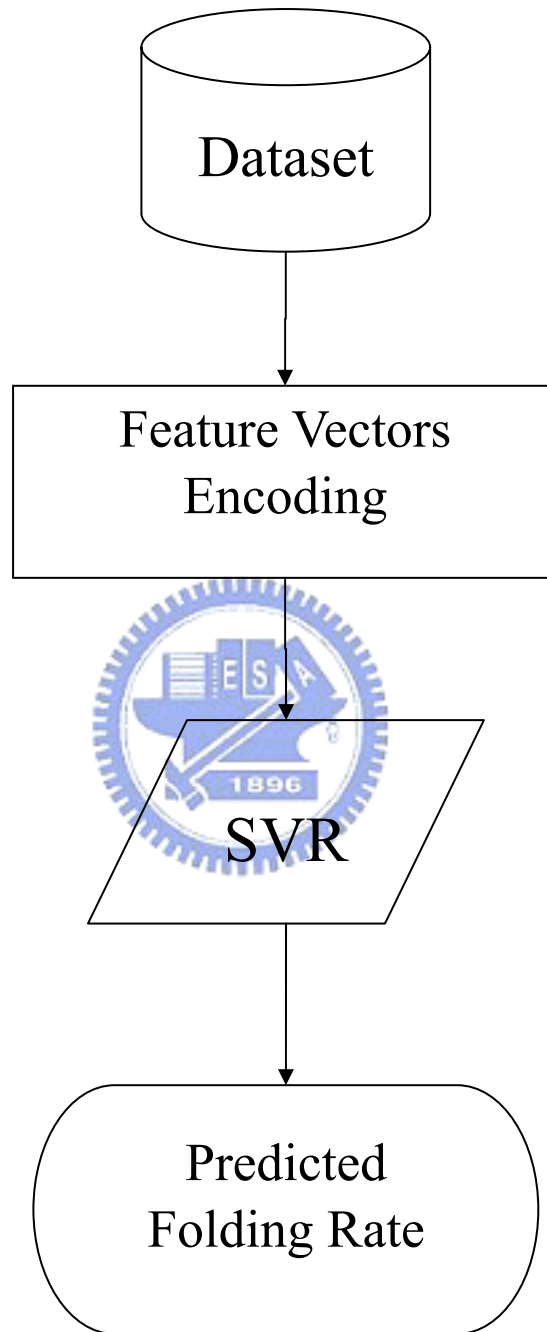


Figure 1 System flowchart

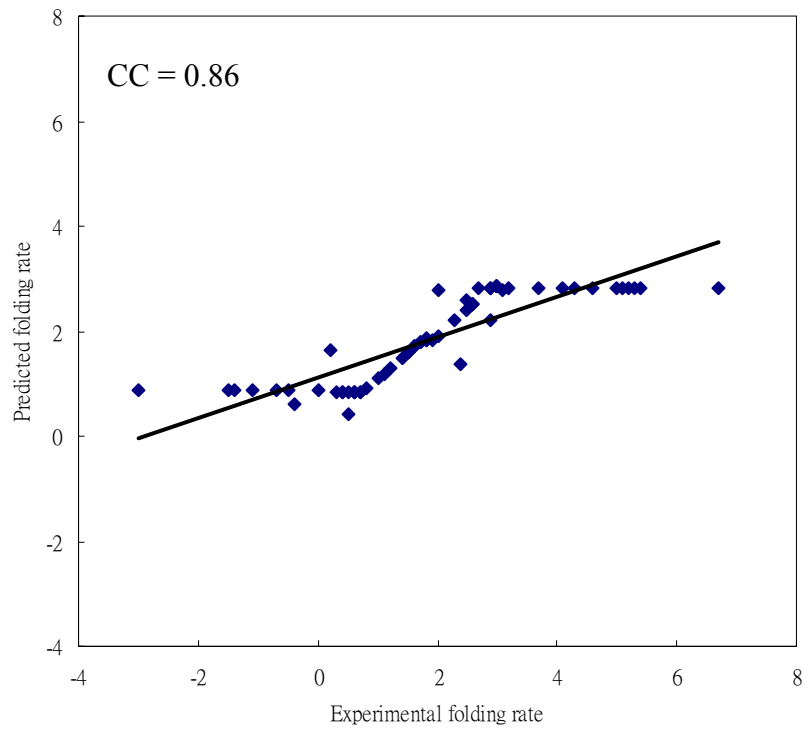


Figure 2 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (PSIPRED) with all dataset



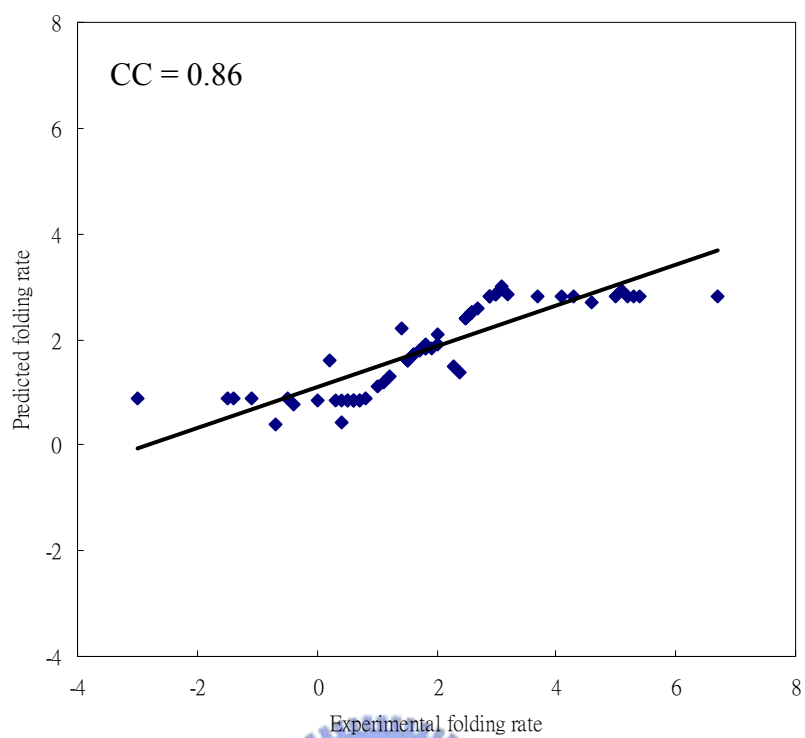
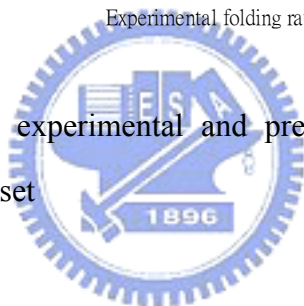


Figure 3 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (ALB) with all dataset



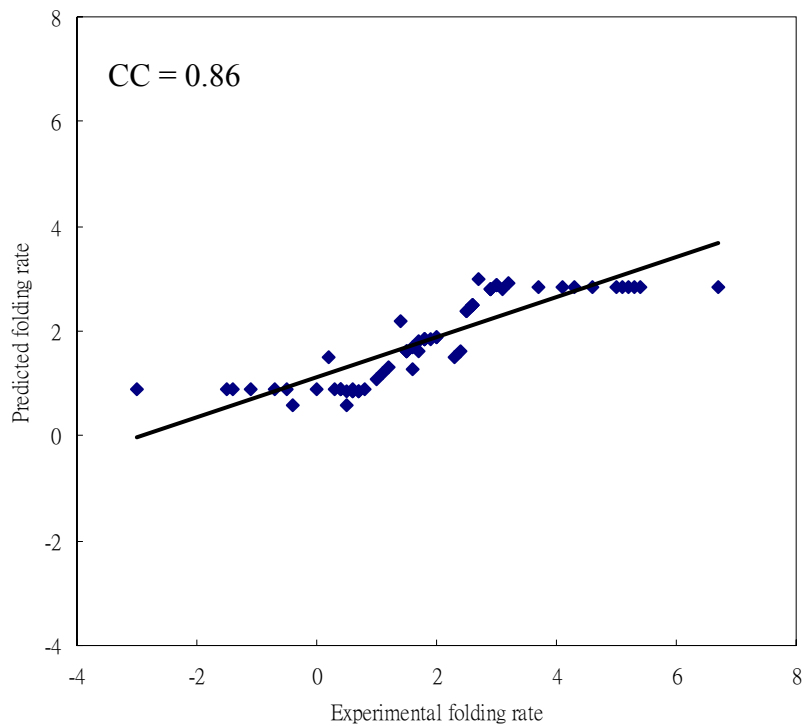


Figure 4 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (DSSP) with all dataset



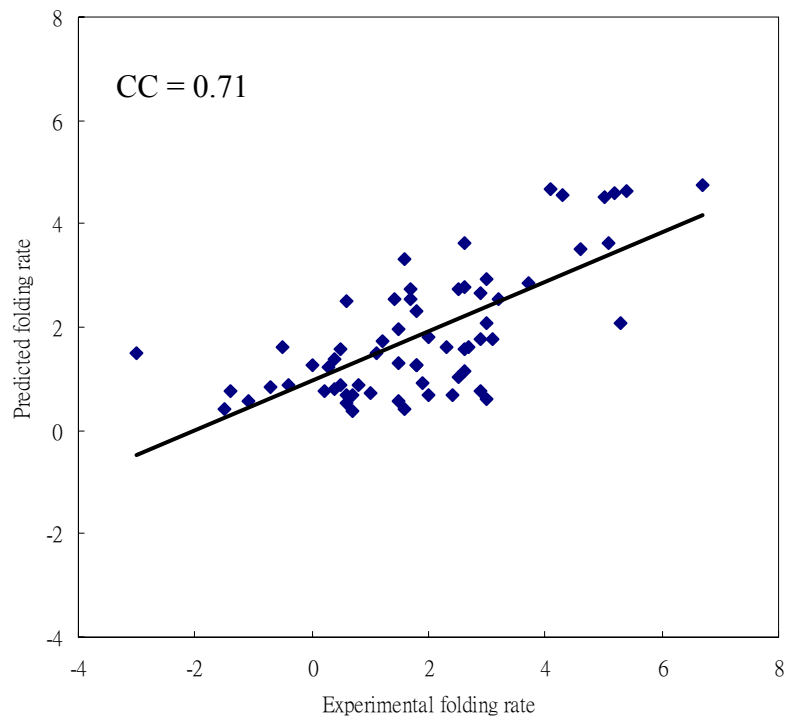


Figure 5 Correlation between experimental and predicted folding rates using  $L+L_S+N_S$  (PSIPRED) with all dataset



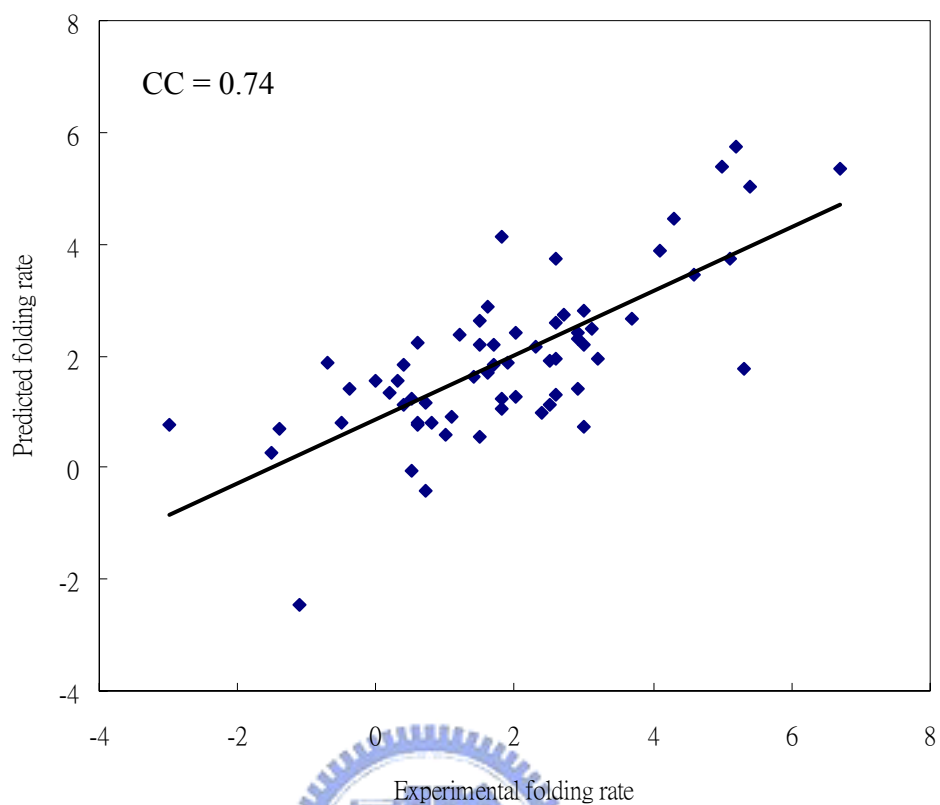


Figure 6 Correlation between experimental and predicted folding rates using L+L<sub>C</sub>+N<sub>C</sub> (PSIPRED) with all dataset



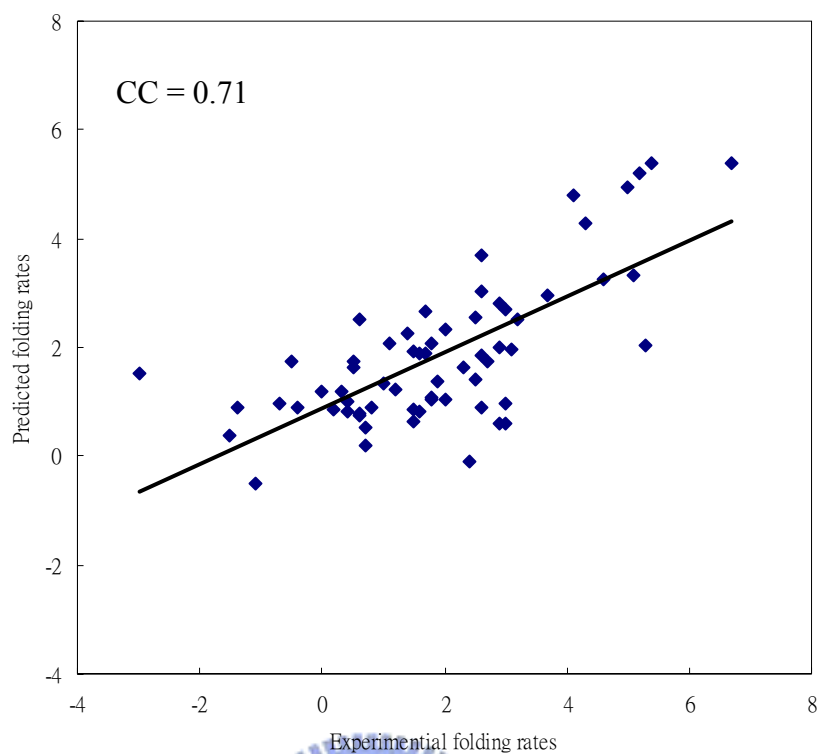


Figure 7 Correlation between experimental and predicted folding rates using  $L+L_S+N_S$  (ALB) with all dataset



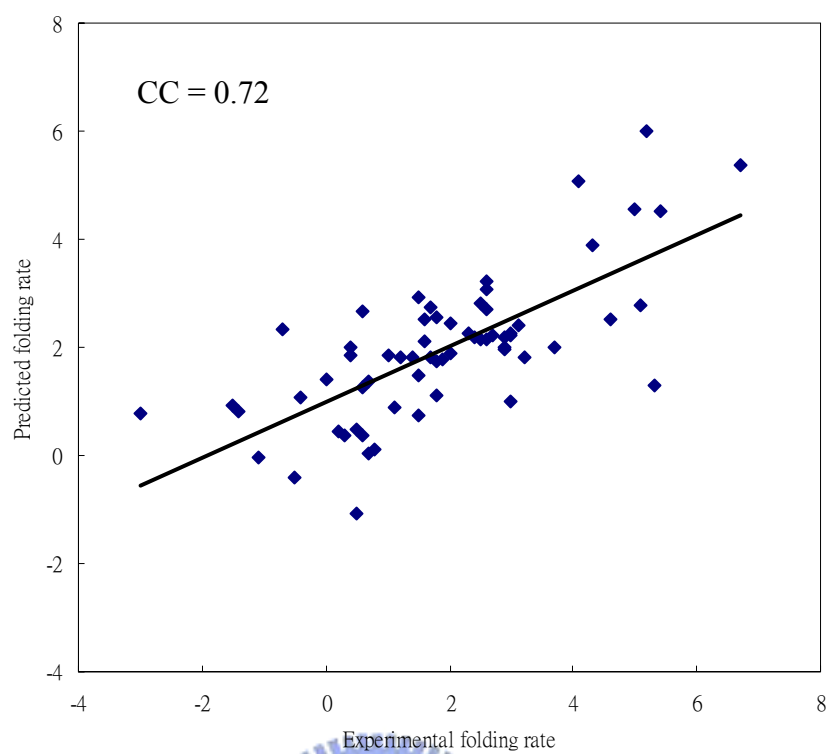


Figure 8 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (ALB) with all dataset



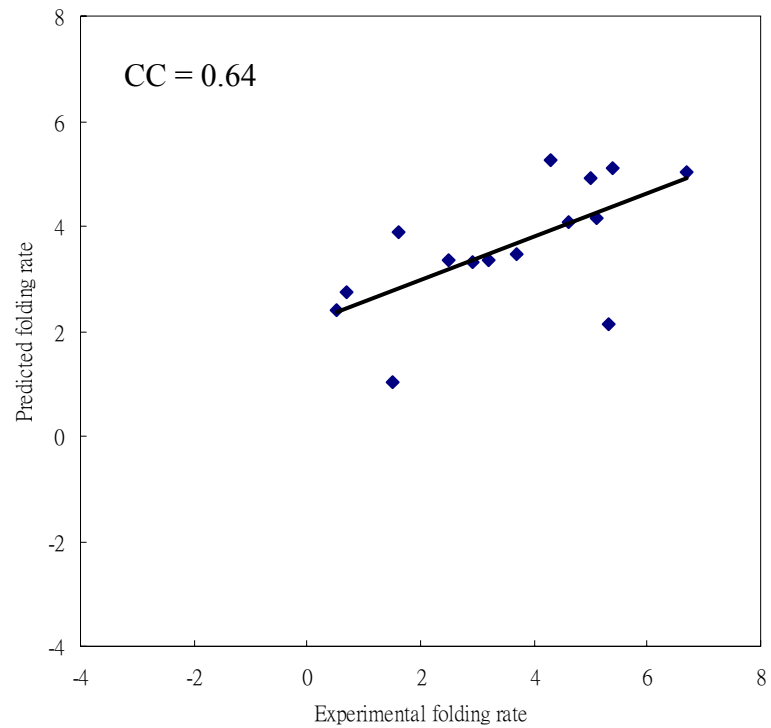


Figure 9 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (PSIPRED) with all- $\alpha$  protein



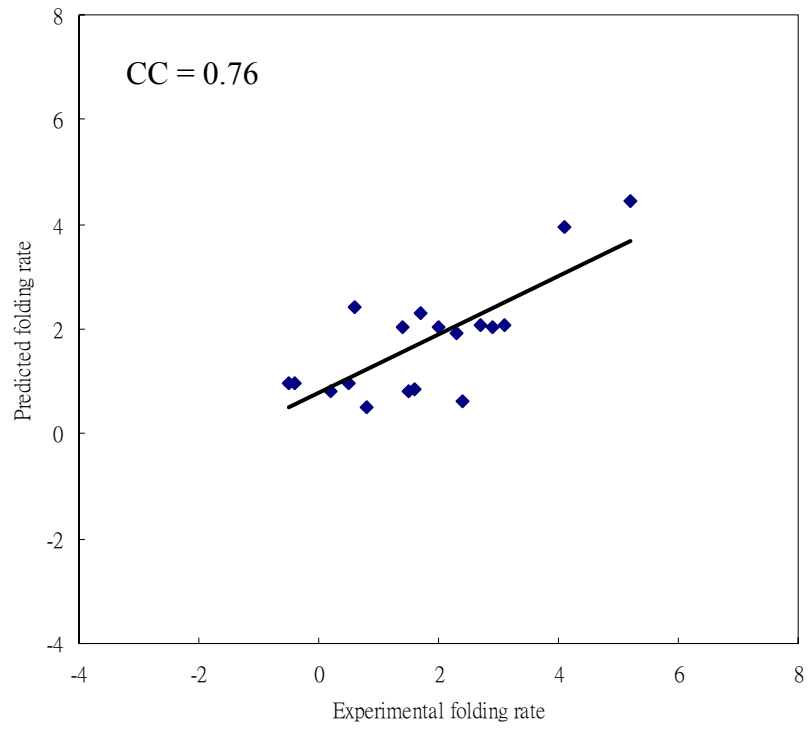


Figure 10 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (PSIPRED) with all- $\beta$  protein



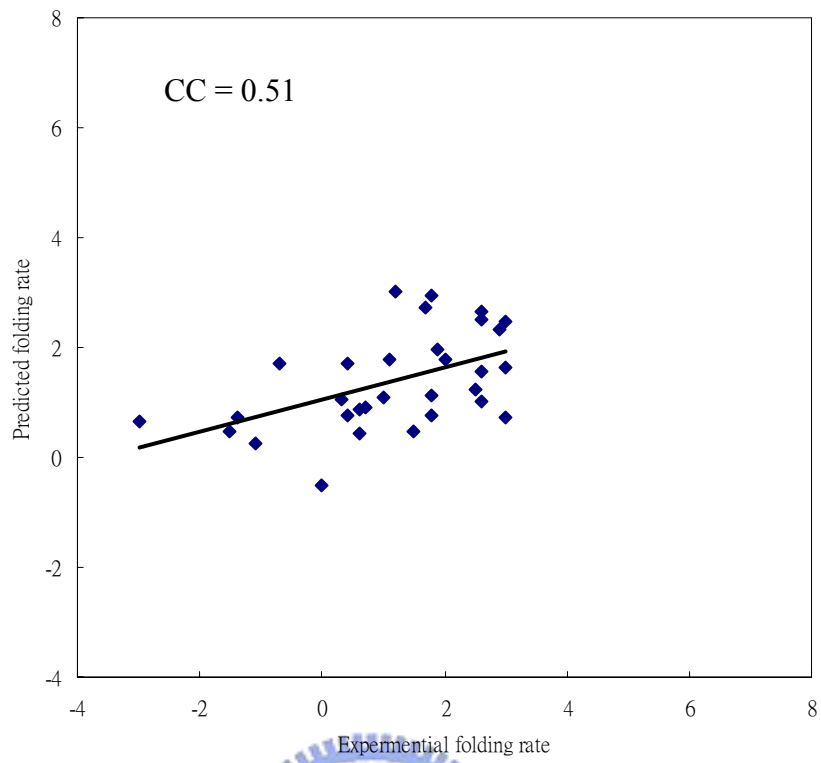


Figure 11 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (PSIPRED) with mixed-class protein

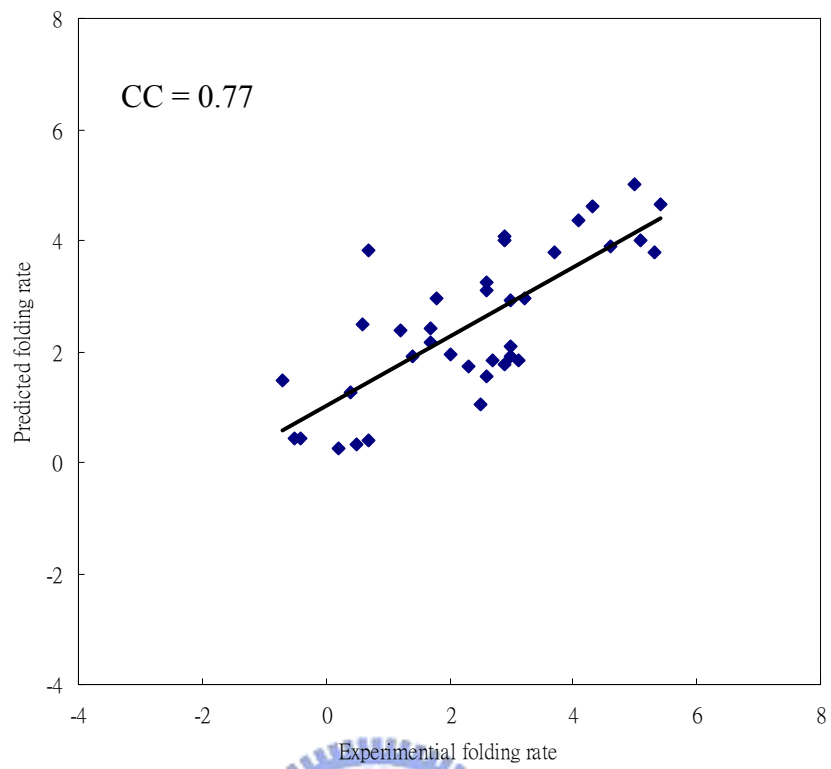
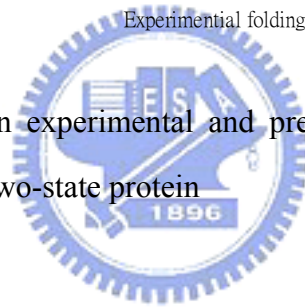


Figure 12 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (PSIPRED) with two-state protein



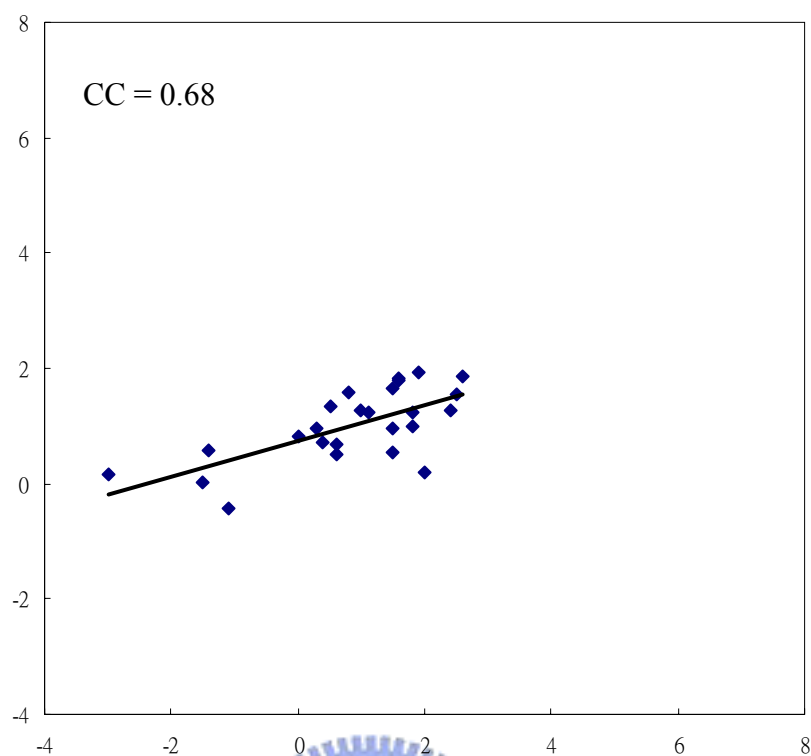


Figure 13 Correlation between experimental and predicted folding rates using  $L+L_H+N_H$  (PSIPRED) with multi-state protein

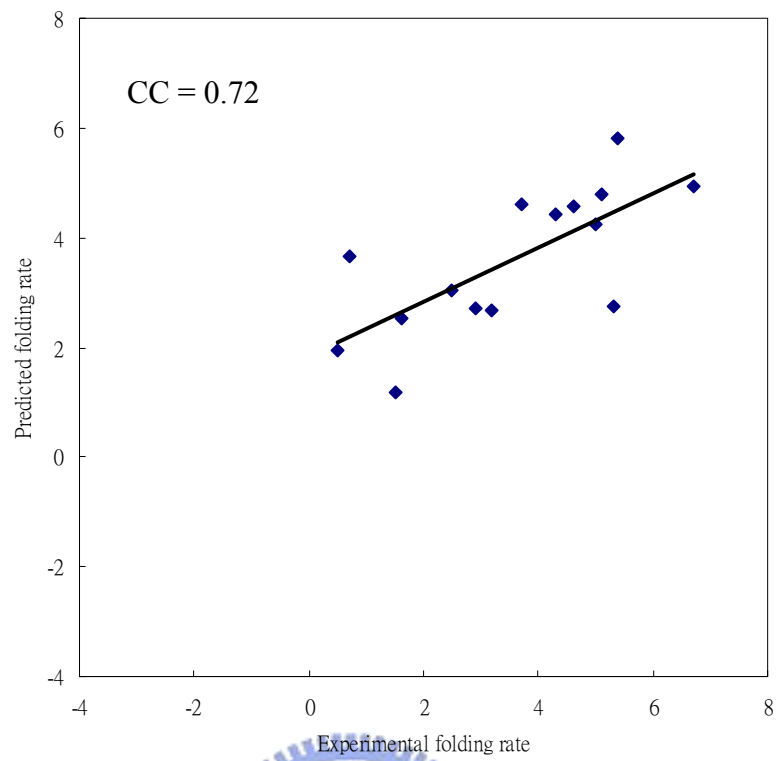


Figure 14 Correlation between experimental and predicted folding rates using L+L<sub>S</sub>+N<sub>S</sub> (PSIPRED) with all- $\alpha$  protein



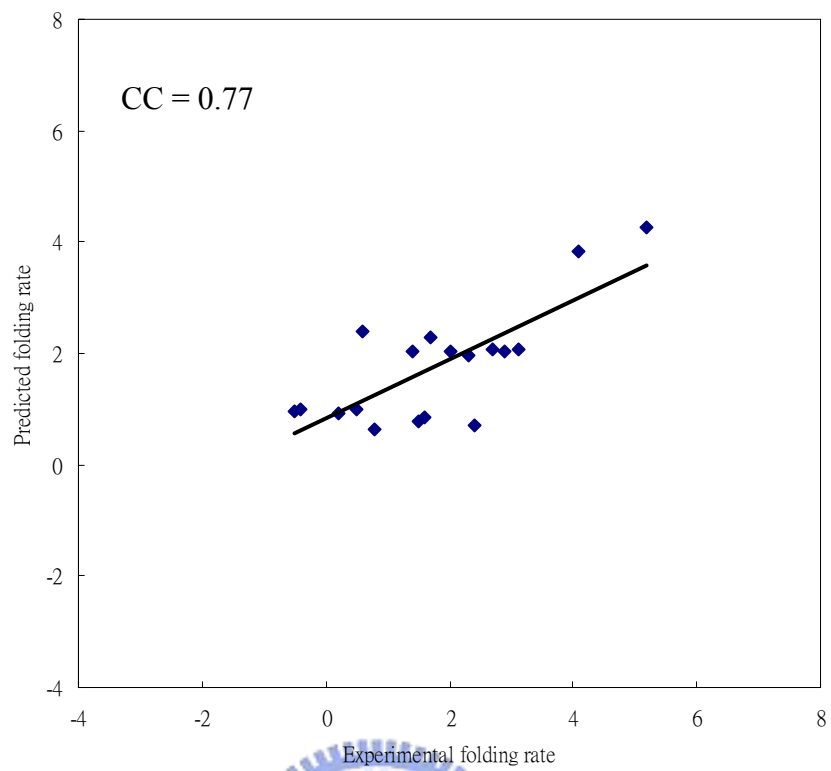


Figure 15 Correlation between experimental and predicted folding rates using L+L<sub>S</sub>+N<sub>S</sub> (PSIPRED) with all- $\beta$  protein

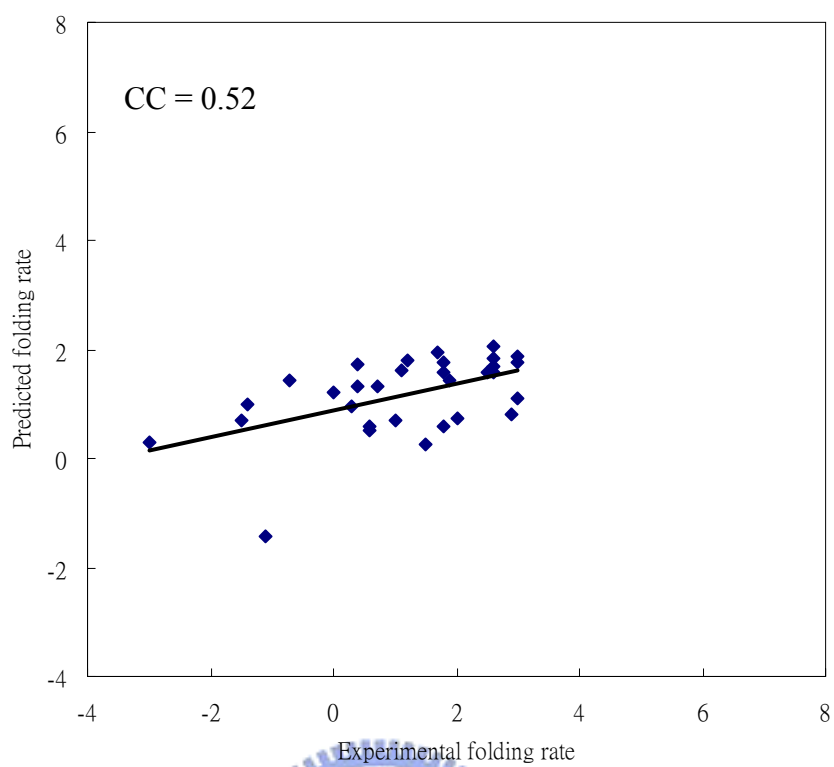


Figure 16 Correlation between experimental and predicted folding rates using L+L<sub>S</sub>+N<sub>S</sub> (PSIPRED) with mixed-class protein

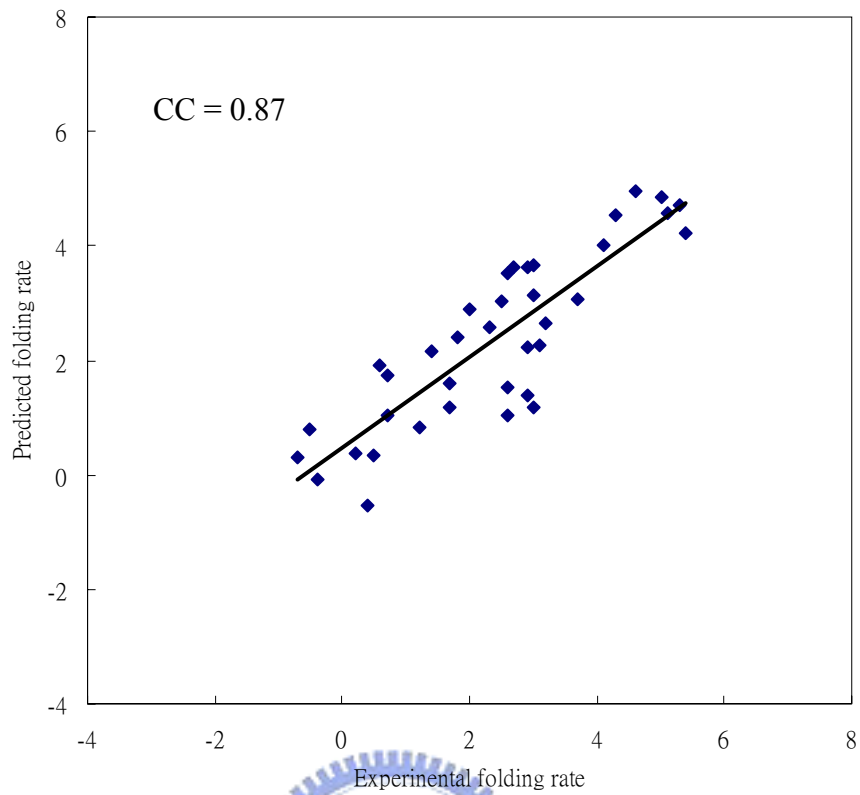


Figure 17 Correlation between experimental and predicted folding rates using  $L+L_S+N_S$  (PSIPRED) with two-state protein

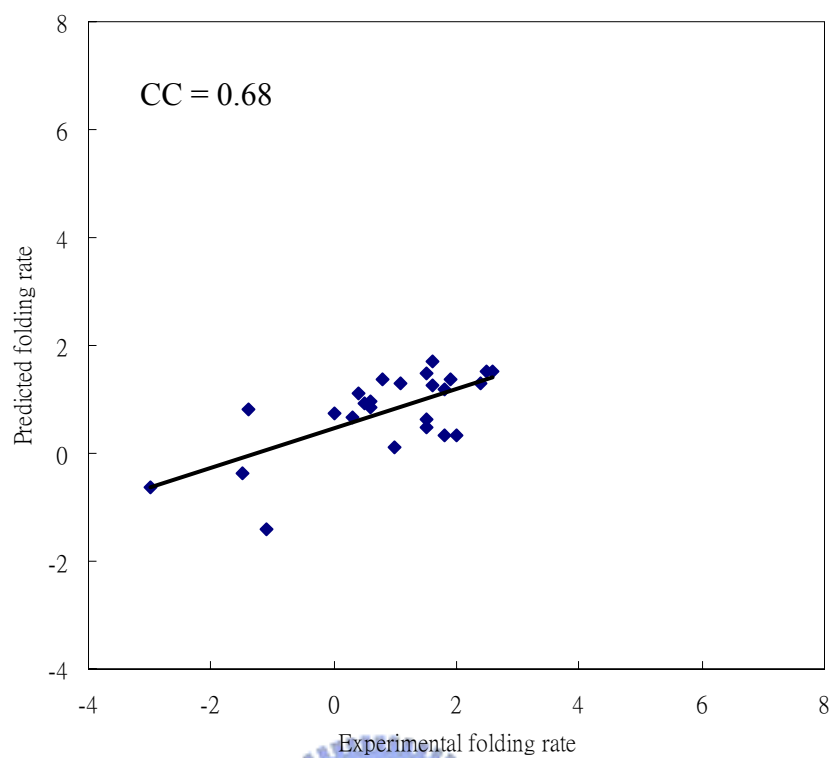


Figure 18 Correlation between experimental and predicted folding rates using L+L<sub>S</sub>+N<sub>S</sub> (PSIPRED) with multi-state protein

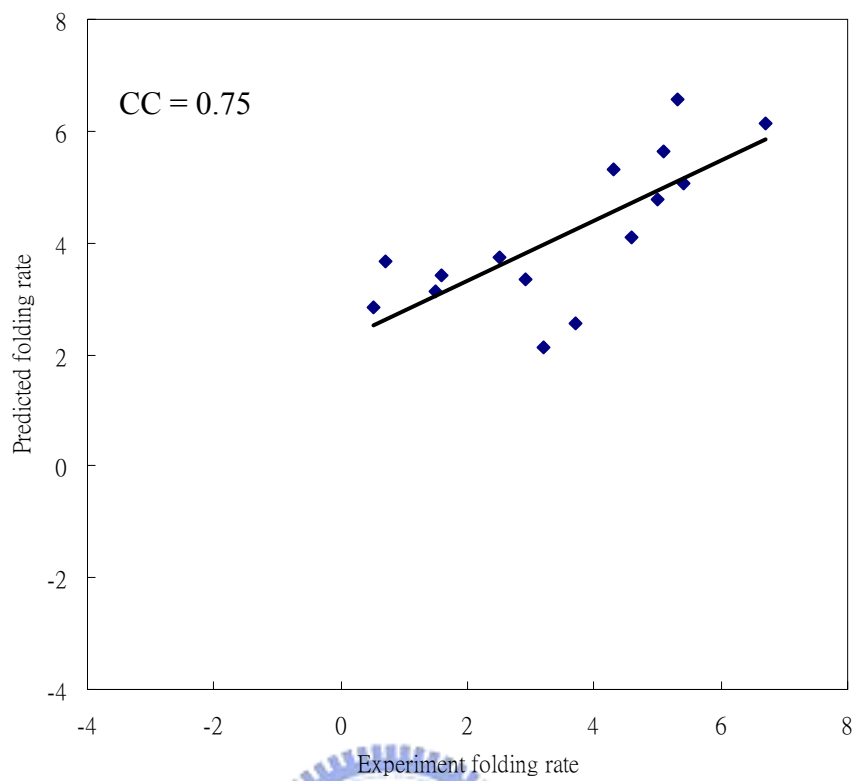


Figure 19 Correlation between experimental and predicted folding rates using L+L<sub>C</sub>+N<sub>C</sub> (PSIPRED) with all- $\alpha$  protein

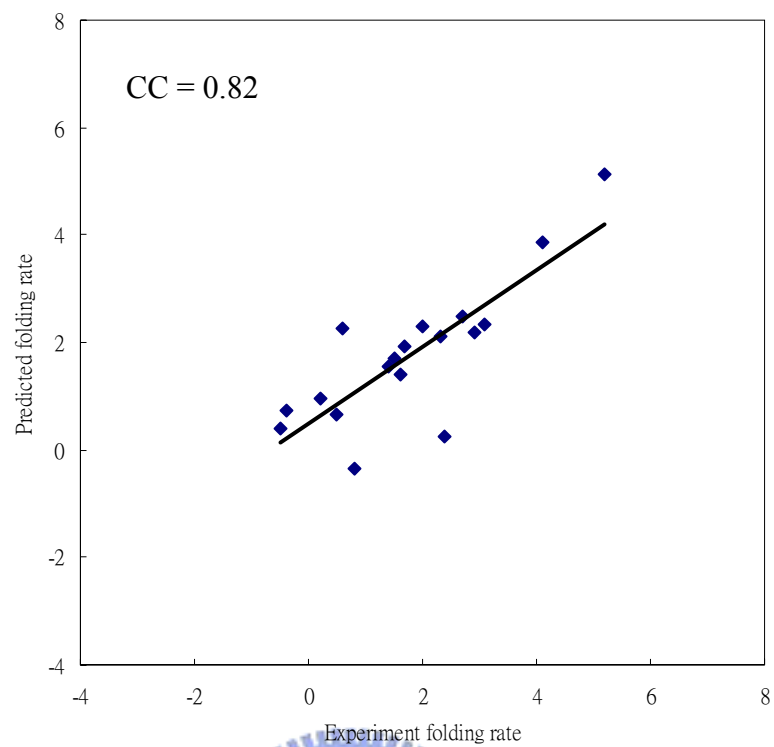
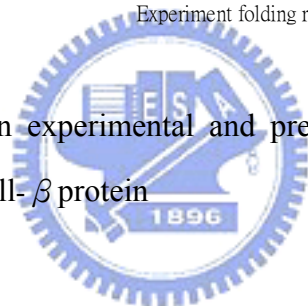


Figure 20 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (PSIPRED) with all- $\beta$  protein



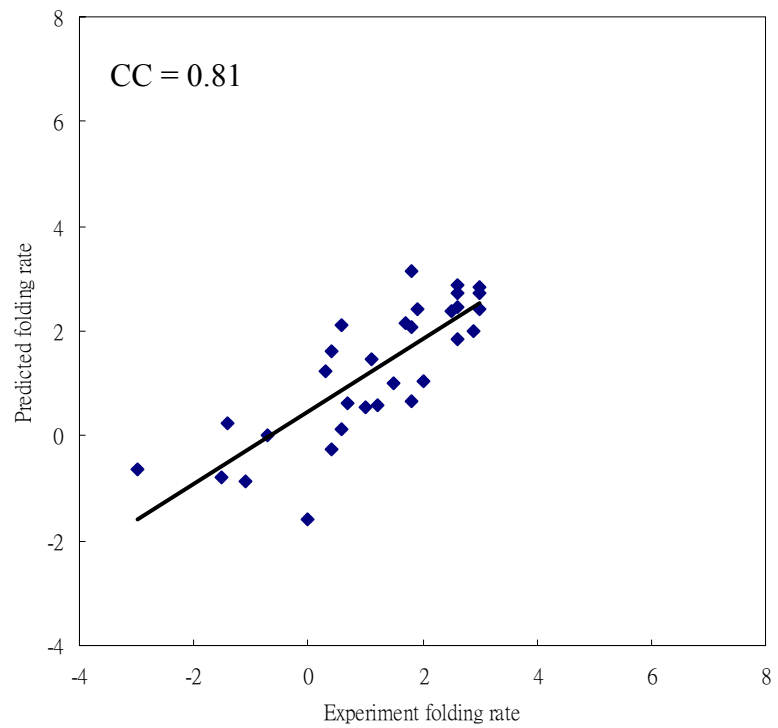
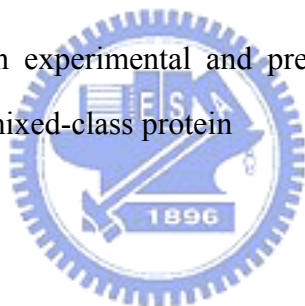


Figure 21 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (PSIPRED) with mixed-class protein



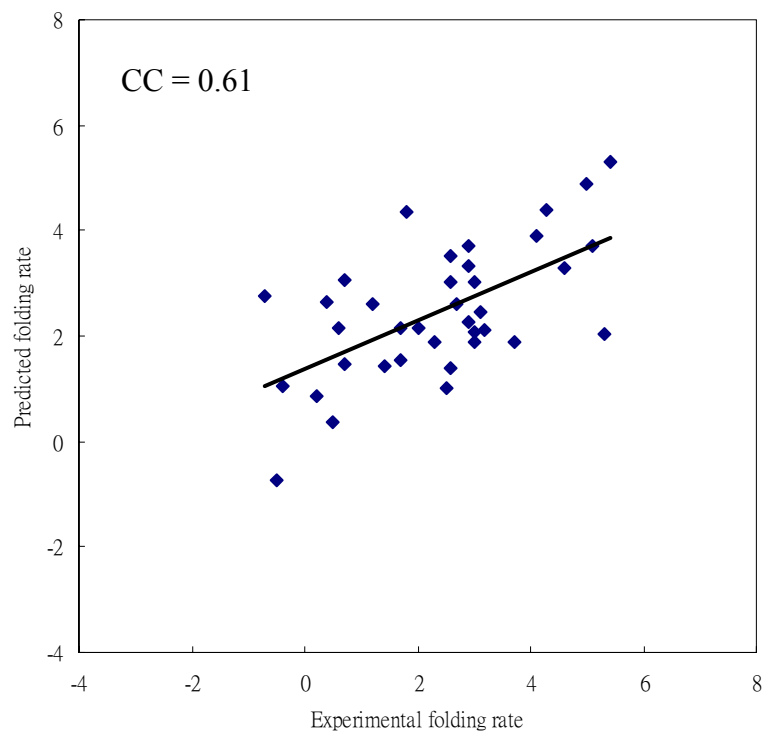
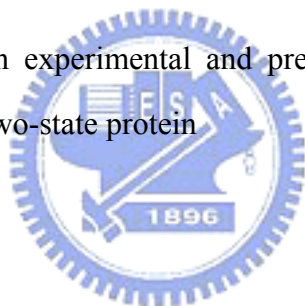


Figure 22 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (PSIPRED) with two-state protein





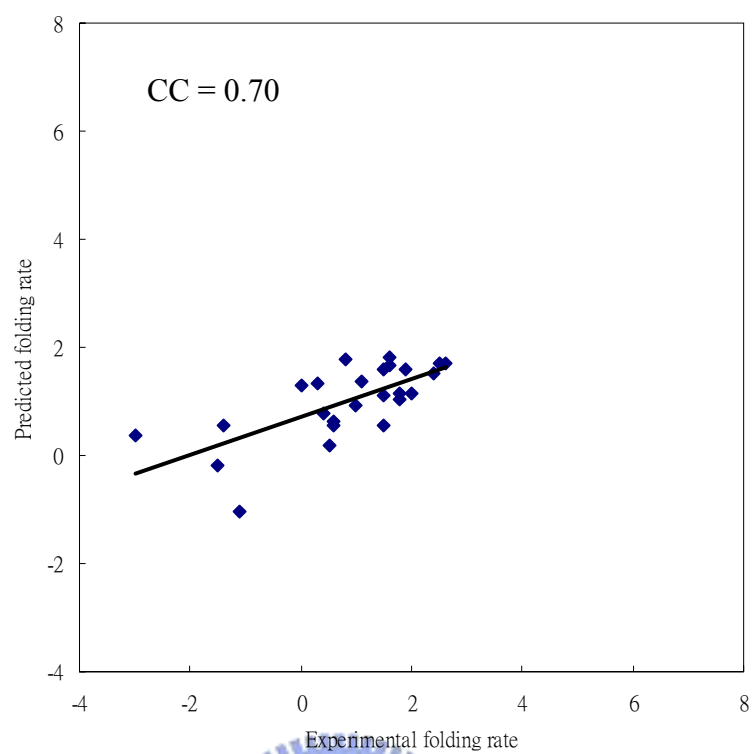


Figure 23 Correlation between experimental and predicted folding rates using  $L+L_C+N_C$  (PSIPRED) with multi-state protein