PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

SWFoldRate: Predicting protein folding rates from amino acid sequence with sliding window method

Xiang Cheng,[1] Xuan Xiao,[1,2*] Zhi-cheng Wu,[1] Pu Wang,[1] and Wei-zhong Lin[1]

[1] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China

[2] Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

ABSTRACT

Protein folding is the process by which a protein processes from its denatured state to its specific biologically active conformation. Understanding the relationship between sequences and the folding rates of proteins remains an important challenge. Most previous methods of predicting protein folding rate require the tertiary structure of a protein as an input. In this study, the long-range and short-range contact in protein were used to derive extended version of the pseudo amino acid composition based on sliding window method. This method is capable of predicting the protein folding rates just from the amino acid sequence without the aid of any structural class information. We systematically studied the contributions of individual features to folding rate prediction. The optimal feature selection procedures are adopted by means of combining the forward feature selection and sequential backward selection method. Using the jackknife cross validation test, the method was demonstrated on the large dataset. The predictor was achieved on the basis of multitudinous physicochemical features and statistical features from protein using nonlinear support vector machine (SVM) regression model, the method obtained an excellent agreement between predicted and experimentally observed folding rates of proteins. The correlation coefficient is 0.9313 and the standard error is 2.2692. The prediction server is freely available at http://www.jci-bioinfo.cn/swfrate/input.jsp.

Proteins 2012; 00:000–000.

Key words: protein folding rates; pseudo amino acid composition; sliding window; sequential backward selection; forward feature selection.

INTRODUCTION

Protein, which is an ordered array of amino acids, plays an important role in the creature body and the amino acids are organized into a large yet uniquely structured molecule. The correct three-dimensional structure is essential to function. Failure to fold into native structure produces inactive proteins that are usually toxic.[1] Several neurodegenerative and other diseases are believed to result from the accumulation of amyloid fibrils formed by misfolded proteins, and many allergies are also caused by the folding of the proteins, for the immune system does not produce antibodies for certain protein structures. The kinetic order and rate constant of the protein folding are the two main aspects for understanding the variations in protein folding kinetics. Folding rate is a measure of slow/fast folding of a protein from its unfolded state to its stable tertiary structure. Proteins have very different rates of folding. Some of them fold within microseconds; some need an hour to fold.

An explosion of protein sequences in the public databases has been witnessed in the post genomic era; but the genome-wide structure and function information has not been available, due to the technical difficulties and labor expenses incurred by existing experimental techniques. The rapid advancements in computer-based protein prediction methods have enabled automated and yet reliable methods for generating folding rates prediction models of proteins.

Grant sponsor: National Natural Science Foundation of China; Grant number: No. 31260273; 60961003; 61261027; Grant sponsor: Key Project of Chinese Ministry of Education; Grant number: 210116; Grant sponsor: Province National Natural Science Foundation of JiangXi; Grant number: 2010GZS0122; 2010GQS0127; 20114BAB211013; Grant sponsor: Department of Education of JiangXi Province; Grant number: GJJ11557; GJJ12490; Grant sponsor: Jiangxi Provincial Foundation for Leaders of Disciplines in Science; Grant number: 20113BCB22008

*Correspondence to: Xuan Xiao; Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China. E-mail: xiaoxuan0326@yahoo.com.cn.

Received 2 May 2012; Revised 20 July 2012; Accepted 25 August 2012
Published online 29 August 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24171

Recently, different methods have been proposed for predicting protein folding rates from amino acid sequence, secondary structure, and structural class information. Protein folding rates prediction methods can be divided into two categories:1. Protein folding rates prediction modeling based on structure information.

Plaxco *et al.*[2] proposed the concept of contact order (CO) using the information about the average sequence separation of all contacting residues in the native state of two-state proteins and found a significant correlation between CO and folding rated of two-state protein. Subsequently, many variations of this idea have been studied, which indicated that folding rates also correlated with long-range order,[3] the total contact distance,[4] a chain topology parameter,[5] and n-order contact distance.[6] These methods require the tertiary structure of a protein as input to predict its folding rate. As the vast majority of protein's tertiary structures are still not solved, several structural parameters have been developed to predict the protein folding rates from protein secondary structures, such as the effective length of a folding chain,[7] the local secondary structure contents,[8] and contact prediction.[9,10] It is important to design methods that can predict folding rate from protein sequence directly.

2. Protein folding rates prediction modeling based on amino acid sequence without knowledge of secondary structures, or information of structural class and without the aid of any other computational prediction of structural properties.

Ma *et al.*[11] explored the correlation between proteins folding rates and their amino acid compositions, and a new indicator called composition index was presented. Unfortunately, this method was only based on the conventional amino acid composition and did not take into account sequence order. Huang and Gromiha[12] have developed a method based on quadratic response surface models for predicting protein folding rates, this method consider the amino acid properties and quadratic response surface model is a nonlinear but low-order model. Guo *et al.*[13] present an algorithm that adopts the concept of the Chou's pseudo amino acid composition (PseAAC) feature extraction method, clearly, the PseAAC can be used to represent a protein sequence with a discrete model without completely discarding the sequence order information. Ever since the concept of PseAAC was introduced, various PseAAC approaches have been proposed to deal with different problems in proteins and protein-related systems. To successfully use the PseAAC for predicting various attributes of proteins, the key is how to optimally extract the features for the PseAA components.

In this study, the long-range and short-range contact in protein were used to derive extended version of the PseAAC based on sliding window method, and a nonlinear machine learning method (Support Vector Machine) predicting protein folding rates was developed using physical, chemical, energetic, and conformational properties of amino acid residues. Our method showed an excellent correlation of 0.9313 between predicted and experimental folding rates of proteins. The prediction server is available online at http://www.jci-bioinfo.cn/SWFoldRate/fold-rate.htm.

MATERIALS AND METHODS

To develop an effective statistical predictor, the following three things are indispensable: (1) a valid benchmark dataset; (2) a mathematical expression for the samples that can effectively reflect their intrinsic correlation with the object to be predicted; and (3) a powerful prediction algorithm or engine. The three necessities for establishing the current protein folding rate predictor were realized via the following procedures.

Data set

As a demonstration, let us use the benchmark dataset constructed in (Guo *et al.* 2011), 117 proteins with known experimentally determined folding rates have been collected from the literatures.[6,7,14–21] We chose 79 proteins and omitted the other 38 proteins for three reasons. (1) Sequences containing ambiguous residue like "X" were excluded; (2) the homology of the proteins will affect prediction accuracy; that is, the prediction accuracy will be overestimated when using highly homologous protein sequences. The homologous proteins by comparison with the Uniprot sequence (http://www.uniprot.org/) were removed from our dataset. (3) Their protein sequence lengths derived from the literatures are different from the data derived from the Protein Data Bank. Amino acid sequences of each protein are taken from the Protein Data Bank (http://www.rcsb.org/pdb/home/home.do).

Amino acid properties

We used a set of 49 diverse amino acid properties (physical–chemical, energetic, and conformational) from Gromiha and Selvaraj.[22,23]

Features representations of protein

Ever since the concept of PseAAC was introduced, it has been widely used to study various problems in proteins and protein-related systems.[24–38] Here, we are to propose a different PseAAC to represent proteins:

Representing target proteins with Sliding Window Method by incorporating long-range and short-range contact in protein

The recent study reveals that 85% of residues are involved in long-range contacts.[22,23]

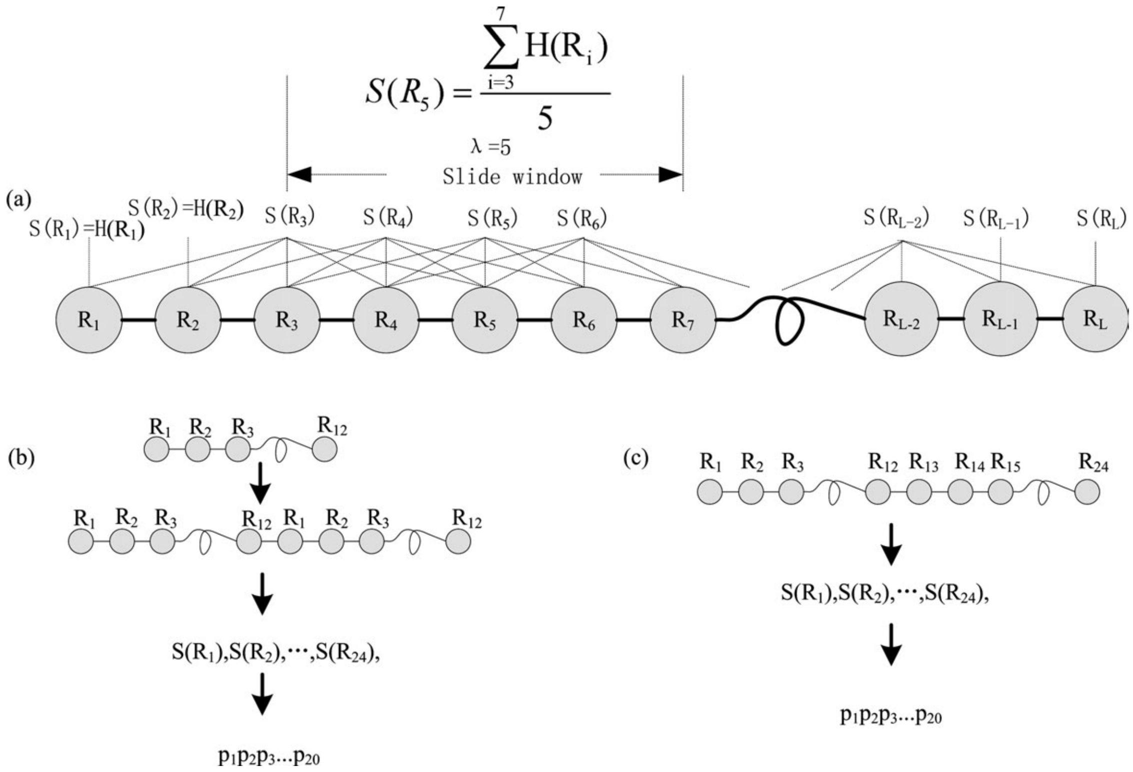


Figure 1

A schematic drawing to show the average sequence value within a sliding window of the certain length. (**a**) the sequence value within a sliding window of five residues while the sequence length *L* is equal to sliding window discrete model size ξ; (**b**) the sequence value with a sliding window of five residues while protein sequence length L is smaller than sliding window discrete model size ξ; (**c**) the sequence value with a sliding window of five residues, sequence length is bigger than sliding window discrete model size ξ.

Here, we present the sliding window method based on the long-range and short-range contact in protein.

Given a protein sequence P with L amino acid residues, that is,

$$\mathrm{P} = R_1 R_2 \cdots R_\mathrm{L} \tag{1}$$

where R_1 represents the 1st residue of a protein P, R_2 the 2nd residue, and they each belong to one of the 20 native amino acids. According to the PseAAC discrete model shown in Figure 1, the protein P of Eq. (1) can be formulated as

$$\mathrm{P} = p_1, p_2, \cdots, p_\xi \tag{2}$$

Where p represents the PseAAC of a protein P. ξ is the number of the PseAAC. If L is smaller than ξ, protein sequence of Eq. (1) need be redefined as:

$$\begin{aligned}\mathrm{P}' &= R_1 R_2 \cdots R_\mathrm{L} R_1 R_2 \cdots R_\mathrm{L} R_1 R_2 \cdots R_\mathrm{L} \\ &= \mathrm{R}_1 R_2 \cdots R_\mathrm{L} R_{\mathrm{L}+1} R_{\mathrm{L}+2} \cdots R_{2\mathrm{L}} \cdots R_{\mathrm{nL}} \quad (\mathrm{nL} = \mathrm{L}') \\ &= R_1 R_2 \cdots R_{\mathrm{L}'}\end{aligned} \tag{3}$$

The action of above will be repeated until the length of new protein sequence is bigger than ξ

$$n \times L \geq \xi \geq (n-1) \times L \tag{4}$$

While $\mathrm{p_u}$ PseAAC are given by

$$p_\mathrm{u} = \begin{cases} \mathrm{S}(R_\mathrm{u}) + \mathrm{S}(R_{\mathrm{u}+\xi}) & u = 1, 2, \cdots, L' - \xi \\ \mathrm{S}(R_\mathrm{u}) & u = L' - \xi + 1, \cdots, \xi \end{cases} \tag{5}$$

$$S(R_i) = \begin{cases} \sum_{k=0}^{\lambda-1} \mathrm{H}(R_{\mathrm{i+k}})/\lambda, & \lambda/2 \leq i \leq \mathrm{L}' - \lambda/2 \\ \mathrm{H}(R_\mathrm{i}), & i \geq \mathrm{L}' - \lambda/2 \ or \ \mathrm{i} \leq \lambda/2 \end{cases} \tag{6}$$

where the symbol $\mathrm{H(R_i)}$ is the original physicochemical property values of amino acids $\mathrm{R_i}$, which can be obtained from Table I, $S(R_i)$ is the mean of the *i*-th amino acid and its λ−1 neighbor amino acids physicochemical property values, and it reflects the sequence order correlation between all of the λ contiguous residues shown in Figure 1. So it can incorporate short-range contact in protein, λ is the size of

Table I
Physico-Chemical Property Values of Amino Acids

	Polarity	pK′	El	Br
A	0	0.9600	0.3600	0.64
C	0.0300	0.2800	0.7000	1
D	0.9600	0.6400	0.0900	0.19
E	0.9600	0.8100	0.1300	0.09
F	0.0100	0.5200	0.7900	0.89
G	0	0.9600	0.4300	0.64
H	0.9900	0.4500	0.4500	0.51
I	0	0	0.8700	0.98
K	0.9500	0.8000	0	0
L	0	0.9800	0.6600	0.87
M	0.0300	0.9000	0.6600	0.72
N	0.0700	0.6500	0.1500	0.21
P	0.0300	0.6200	0.3000	0.26
Q	0.0700	0.7900	0.1900	0.13
R	1.0000	0.4400	0.4700	0.06
S	0.0300	0.8300	0.2800	0.36
T	0.0300	0.7300	0.4200	0.36
V	0	0.9400	0.8100	0.83
W	0.0400	1.0000	1.0000	0.68
Y	0.0300	0.8200	0.6600	0.42

Polarity is one of the most important amino acid property; pK′ refers to equilibrium constant with reference to the ionization property of COOH group; El refers to long-range nonbonded energy; Br refers to Buriedness.

sliding window. $\mathrm{p_u}$ is sum of the u-th and $(u+\xi)^{th}$ amino acid physicochemical property values when $u = 1, 2, \cdots, L' - \xi$, so it incorporates long-range contact in protein.

If $L \geq \xi$, then

$$p_{\mathrm{u}} = \begin{cases} \sum_{i=0}^{\lfloor L/\xi \rfloor} \mathrm{S}(R_{\mathrm{i}\times\xi+\mathrm{u}}) & u = 1, 2, \cdots, L - \lfloor L/\xi \rfloor \times \xi \\ \sum_{i=0}^{\lfloor L/\xi \rfloor - 1} \mathrm{S}(R_{i\times\xi+\mathrm{u}}) & u = L - \lfloor L/\xi \rfloor \times \xi, \cdots, \xi \end{cases} \tag{7}$$

where $\mathrm{p_u}$ incorporates long-range contact in protein. The sliding window pseudo code is shown in Figure 2. For three different amino acid, physical–chemical properties: polarity, pK′, El (Table I), the feature set $\mathrm{SW_{polarity}}$, $\mathrm{SW_{pK'}}$, and $\mathrm{SW_{El}}$ can be got:

$$\mathrm{SW_{Polarity}} = [p_1, p_2, \cdots, p_\xi]^{\mathrm{T}} \tag{8}$$

$$\mathrm{SW_{pK'}} = [p_{\xi+1}, p_{\xi+2}, \cdots, p_{2\times\xi}]^{\mathrm{T}} \tag{9}$$

$$\mathrm{SW_{E1}} = [p_{2\times\xi+1}, p_{2\times\xi+2}, \cdots, p_{3\times\xi}]^{\mathrm{T}} \tag{10}$$

where T is the transpose operator.

Representing target proteins with PseAAC by correlation factor

The correlation factor φ_i reflects the sequence order correlation between all the i-th most contiguous residues as formulated by

$$\varphi_i = \frac{\sum_{j=1}^{L-i} (\mathrm{H}(R_j) \times \mathrm{H}(R_{j+i}))}{L - i} \tag{11}$$

$$P = [\varphi_i, \varphi_2, \cdots \varphi_\lambda]^{\mathrm{T}} \tag{12}$$

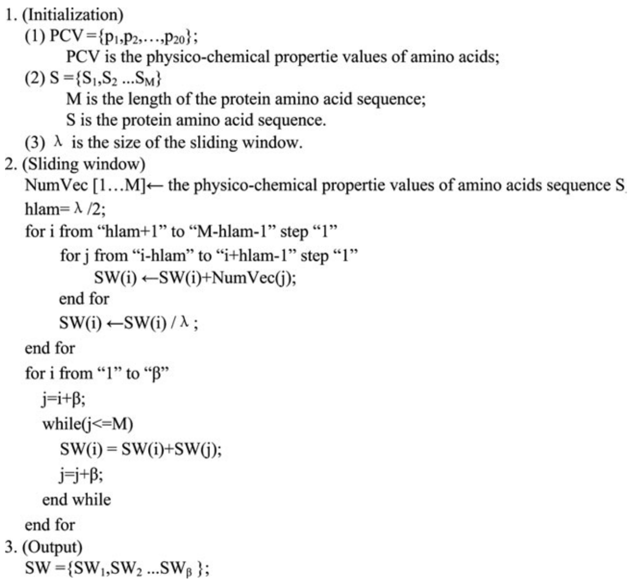
Where the symbol $\mathrm{H(R_i)}$ is the original physicochemical property values of amino acids for $\mathrm{R_i}$. Previous investigations indicated that the optimal value for λ should be the one that results in the best overall jackknife test. We have tried different values of λ in our method, and finally found $\lambda = 10$ can be used as the optimal value for the dataset. A set of 49 diverse amino acid properties (physical–chemical, energetic, and conformational) was used. The total number of components thus obtained for a given protein is $49 \times 10 = 490$, the protein can be formulated as a 490-D vector.

Complexity Factor

The complexity measure factor of a protein sequence can be used to reflect its pattern or sequence feature and has been successfully used in some protein attribute prediction.[34] Among the known measures of complexity, the Lempel–Ziv (LZ) complexity reflects the order that is retained in the sequence.

The complexity measure factor, CF(P), of a nonempty sequence synthesized according to the following procedure is defined by

$$\mathrm{Syn}(P) = P[1 : i_1] \bullet P[i_1 + 1 : i_2] \bullet \cdots \bullet P[i_{m-1} + 1 : L] \tag{13}$$



```
1. (Initialization)
  (1) PCV ={$p_1$,$p_2$,…,$p_{20}$};
        PCV is the physico-chemical propertie values of amino acids;
  (2) S ={$S_1$,$S_2$ ...$S_M$}
        M is the length of the protein amino acid sequence;
        S is the protein amino acid sequence.
  (3) λ is the size of the sliding window.
2. (Sliding window)
  NumVec [1…M]← the physico-chemical propertie values of amino acids sequence S
  hlam= λ /2;
  for i from "hlam+1" to "M-hlam-1" step "1"
      for j from "i-hlam" to "i+hlam-1" step "1"
          SW(i) ←SW(i)+NumVec(j);
      end for
      SW(i) ←SW(i) / λ ;
  end for
  for i from "1" to "β"
    j=i+β;
    while(j<=M)
      SW(i) = SW(i)+SW(j);
      j=j+β;
    end while
  end for
3. (Output)
  SW ={$SW_1$,$SW_2$ ...$SW_β$ };
```

Figure 2
Sliding Window Pseudocode.

where P[i:j] is defined by

$$P[i:j] = R_i R_{i+1} R_{i+2} \cdots R_j \quad (1 \le i \le j \le L) \tag{14}$$

For example, for the sequence P=TMPPPETPSEGRQPSPSPSPTT, the LZ schema of synthesis generates the following components Syn(P) and the corresponding complexity CF(P):

$$\begin{aligned}\mathrm{Syn(P)} = \mathrm{T} \bullet \mathrm{M} \bullet \mathrm{P} \bullet \mathrm{PPE} \bullet \mathrm{TP} \bullet \mathrm{S} \bullet \mathrm{EG} \\ \bullet \mathrm{R} \bullet \mathrm{Q} \bullet \mathrm{PSP} \bullet \mathrm{SPSPT} \bullet \mathrm{T}\end{aligned} \tag{15}$$

$$\mathrm{CF(P)} = 12 \tag{16}$$

Hybridization discrete model

The complexity factor (CF), length of the protein, correlation factor, and long-range and short-range contact factor all contain the information of constituent amino acids in a protein as well as some of its sequence pattern or order. In principle, the more the components the PseAAC is formed, the more sequence-order information it contains. A hybridization approach has been introduced by fusing the above three factors, we used a 2 + 490 + 3 × ξ dimension vector to represent a protein sequence.

LIBSVM

With all samples represented by a feature vector, now it is possible for us to construct our predictor using the nu-support vector regression (nu-SVR) which is integrated in the software LIBSVM. We can download it freely from http://www.csie.ntu.edu.tw/∼ cjlin/libsvm/. The kernel function was set as linear.

The data should be scaled before applying SVM. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation. Each attribute is scaled to the range [−1;+1] linearly.

Forward feature selection

However, it might cause over-fitting problem if the PseAAC contains too many components. Therefore, an optimal PseAAC should consist of as many key and as few trivial components as possible. Here, the forward feature selection (FFS) procedure is used to solve the problem. Previous investigations indicated that Complexity measure factor CF(P), protein sequence length, and $\mathrm{SW_{polarity}}$, $\mathrm{SW_{pK'}}$, $\mathrm{SW_{El}}$ feature sets can effectively describe the sequence-order information, they are enclosed to the already-selected feature set first, and the correlation factor is the to-be-selected feature set.

$$\begin{aligned}&\text{Already} - \text{selected feature set}\\ &\quad = \{\mathrm{CF(P)}, \text{length}, \ \mathrm{SW_{Polarity}}, \mathrm{SW_{pK'}}, \mathrm{SW_{E1}}\}\end{aligned} \tag{17}$$

$$\begin{aligned}&\text{to} - \text{be} - \text{selected feature set} =\\ &\{\{\varphi_1\varphi_2\cdots\varphi_{10}\}, \{\varphi_{1\times10+1}\varphi_{1\times10+2}\cdots\varphi_{1\times10+10}\},\\ &\cdots\{\varphi_{48\times10+1}\varphi_{48\times10+2}\cdots\varphi_{48\times10+10}\}\}\end{aligned} \tag{18}$$

We added each candidate feature subset to the already-selected feature set, and 49 new feature sets can be gotten. A nu-SVR algorithm predictor was constructed, and all the new feature sets are tested with the jackknife cross-validation test. We could draw an FFS curve with the index i to be the *x*-axis and the corresponding overall accurate rate to be the *y*-axis.

$$\begin{aligned}S_{\mathrm{opt}} = \{&\mathrm{CF(P)}, \text{length}, \mathrm{SW_{polarity}}, \mathrm{SW_{pK'}}, \mathrm{SW_{E1}},\\ &\{\varphi_{j\times10+1}\varphi_{j\times10+2}\cdots\varphi_{j\times10+10}\}\}\end{aligned} \tag{19}$$

is regarded as the optimal feature set if the curve reach its peak where the value of its *x*-axis is j $(0 \le j \le 48)$.

Sequential backward selection

To refine feature selection, the sequential backward selection (SBS) procedure based on the result of FFS was used, in which features are sequentially removed from a full candidate set until the removal of further features increase the criterion.

RESULTS AND DISCUSSION

Cross validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. One round of cross validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the training set), and validating the analysis on the other subset (called the validation set or testing set). The common types of cross validation include K-fold and Leave-one-out cross validation. For the K-fold cross validation, the partition is random, the result is variable. Leave-one-out cross validation is also called Jackknife test, which involves using a single observation from the original sample as the validation data, and the remaining observations as the training data. This is repeated such that each observation in the sample is used once as the validation data. It is a rigorous and objective statistical test that can always yield a unique result for a given test dataset. Therefore, it is used to examine the power of our predictor.

Results of sliding window

We used a set of 49 diverse amino acid properties (physical–chemical, energetic, and conformational), which fall into various clusters analyzed by Tomii and

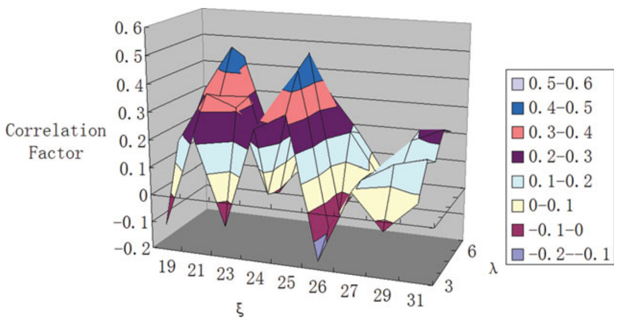


Figure 3

A schematic drawn to show the correlation coefficient between predicting and experimental data of protein folding rates according to different sliding window size. The peak of curved surface reaches 0.6161 while $\lambda = 5$ and $\xi = 25$.

Kanehisa[39] in this study. At last, three diverse amino acid properties, polarity, pK′, and El (Table I) are selected to calculate the PseAACs. El is the long-range nonbonded energy property of protein, pK′ equilibrium constant with reference to the ionization property of COOH group, polarity one of the most used property. The three kind of PseAAC should be gotten from the three diverse amino acid properties using sliding window method when $\lambda = 5$ and $\xi = 25$. It is found that correlation coefficient from E1 is 0.2430 which is the highest, and pK′ is −0.031 which is the lowest. Test proves that all the three diverse amino acid properties are indispensable to predict folding rates. Correlation coefficient is 0.46 which is higher than any single property while polarity, pK′, El are all taken into consideration and $\lambda = 5$, $\xi = 25$. To test the sliding window method, we predict the protein folding rates when λ changing from 3 to 6 and ξ changing from 19 to 31. The results are shown in Figure 3. Optimal correlation coefficients has been got when $\lambda = 5$ and $\xi = 25$. The result agrees with the long-range and short-range contact in protein depicted by Gromiha and Selvaraj,[22,23] who described that the long-range contacts computed for different intervals in four structural classes all play important roles, and the α/β class of proteins prefers the 21–30 range. In Table II, predicting results are given which are based on sliding window of 25 range and the correlation coefficient of α/β class of proteins is highest, which agrees with the result of Gromiha. The all-α class proteins have more long-range contacts in the 4–10 range and the all-β class proteins have more long-range contacts in the 11–20 range. The range 4–10 is favored by $\alpha+\beta$ class of proteins.[22,23] In Table II, predicting results of other three structure classes are lower comparing to the α/β class.

Results of FFS and SBS

Each feature subset in FFS-to-be-selected feature set would be taken out and added to the FFS-selected feature

Table II

Predicted Folding Rates in a Set of 79 Proteins

PDB ID	Length	Structural type	Experimental	Predicted	Correlation coefficient
1A6N	154	All-α	1.1	1.5469	0.8658
1AON	97	All-α	−1.48	−3.0136	
1AON	548	All-α	0.8	1.3444	
1ARR	53	All-α	6.8	7.5924	
1BA5	439	All-α	11.75	6.0726	
1BDD	508	All-α	2.6	2.6863	
1CEI	87	All-α	3.87	3.5123	
1EBD	470	All-α	10.53	10.9483	
1ENH	552	All-α	1.46	1.4138	
1FEX	399	All-α	4.05	4.7463	
1HRC	109	All-α	4.1	4.0015	
1IDY	636	All-α	7.3	7.11	
1IMQ	86	All-α	8.37	8.9935	
1L8W	356	All-α	8.5	10.0012	
1LMB	237	All-α	6.6	6.553	
1PRB	394	All-α	1.17	1.1371	
1VII	7,158	All-α	0.41	2.6003	
1YCC	105	All-α	12.2	9.0796	
256B	128	All-α	12.7	12.547	
2ABD	87	All-α	6.55	4.5289	
2CRO	71	All-α	3.7	3.5956	
2PDD	724	All-α	9.8	3.2139	
1C8C	64	All-β	−3.2	−2.7804	0.9513
1C90	313	All-β	5.8	5.2218	
1CBI	137	All-β	3.87	4.1398	
1CSP	67	All-β	10.37	9.6034	
1E0L	1,100	All-β	1.3	0.9376	
1EAL	128	All-β	9.68	7.3609	
1FMK	536	All-β	6.3	6.2111	
1FNF	2,386	All-β	4.38	3.998	
1G6P	66	All-β	2.7	2.8379	
1HNG	344	All-β	0.74	1.4967	
1HX5	100	All-β	8.73	5.121	
1IFC	132	All-β	−0.71	1.9879	
1K8M	482	All-β	12.4	8.5147	
1LOP	164	All-β	5.24	4.0887	
1MJC	70	All-β	4.54	4.6129	
1NYF	537	All-β	1.4	1.5745	
1OPA	134	All-β	6.8	6.5889	
1PIN	163	All-β	−1.05	−0.1692	
1PKS	387	All-β	−1.1	1.2938	
1PNJ	724	All-β	2.7	2.8053	
1PSE	70	All-β	3.2	3.1142	
1QTU	107	All-β	−2.5	−1.6801	
1SHG	2,477	All-β	4.04	4.1194	
1SRL	533	All-β	1.06	1.1075	
1TEN	2,201	All-β	3.47	3.6611	
1TIT	34,350	All-β	3.45	3.3408	
1WIT	685	All-β	9.62	10.399	
2ait	104	All-β	4.2	4.6327	
1APS	99	α+β	9.27	8.2274	0.9830
1BNI	157	α+β	3.4	4.9118	
1div	149	α+β	8.85	8.8981	
1FKB	108	α+β	−0.9	−0.3194	
1GXT	750	α+β	2.89	2.6585	
1HDN	85	α+β	8.76	8.1217	
1HZ6	719	α+β	3.4	3.4452	
1PBA	416	α+β	6.8	6.7969	
1PGB	448	α+β	−3.45	−3.0131	
1RFA	648	α+β	5.9	5.9549	
1RIS	101	α+β	4.2	4.1501	
1SCE	113	α+β	4.5	3.6133	
1UBQ	397	α+β	5.73	6.536	

Table II
(Continued)

PDB ID	Length	Structural type	Experimental	Predicted	Correlation coefficient
1URN	282	$\alpha+\beta$	11.52	11.1771	
2A5E	156	$\alpha+\beta$	3.5	3.2543	
2ACY	101	$\alpha+\beta$	0.92	1.4853	
2hqi	91	$\alpha+\beta$	0.18	0.7877	
2LZM	164	$\alpha+\beta$	4.1	2.5961	
2VIK	826	$\alpha+\beta$	6.8	6.2019	
1AYE	419	α/β	5.91	6.0137	0.9926
1BRS	90	α/β	7	7.2257	
1pca	419	α/β	6	5.3891	
1PHP	567	α/β	9.44	10.0838	
1qop	268	α/β	−0.36	−0.3194	
1RA9	159	α/β	7	7.0243	
2RN2	155	α/β	0.1	0.9541	
3CHY	129	α/β	1	0.89	
1cis	84	Designed protein	6.98	7.5008	1
1ubo	317	Multidomain protein (alpha and beta)	5.9	5.5652	

set. Each predictor based on each new FFS-selected feature set would be tested, and the feature set obtained the highest overall accurate rate would be used as the new FFS-selected feature set. The FFS curve is generated shown in Figure 4. $\{\varphi_{20\times10+1}\varphi_{20\times10+2}\cdots\varphi_{20\times10+10}\}$ is selected into the S_{FFS} and the correlation coefficients reaches 0.6161.

With the FFS-selected feature set, SBS was processed for each of the features. Twenty-three features were deleted according to the sequential backward selection algorithm. The result is shown in Figure 5. The optimal correlation coefficient reaches 0.9313. The protein folding rates correlation coefficient of all structural type are shown in Table II. The correlation coefficient of all-α is lowest, and the correlation coefficient of α/β is highest.

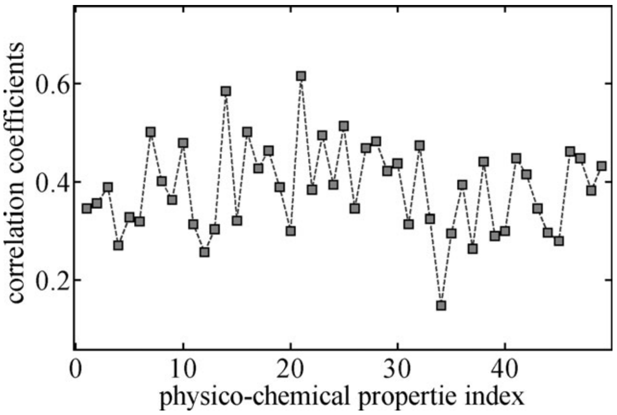


Figure 4
A schematic drawing to show the correlation coefficients between predicting and experimental data of protein folding rate using different physicochemical property values of amino acids and feature set from sliding window.

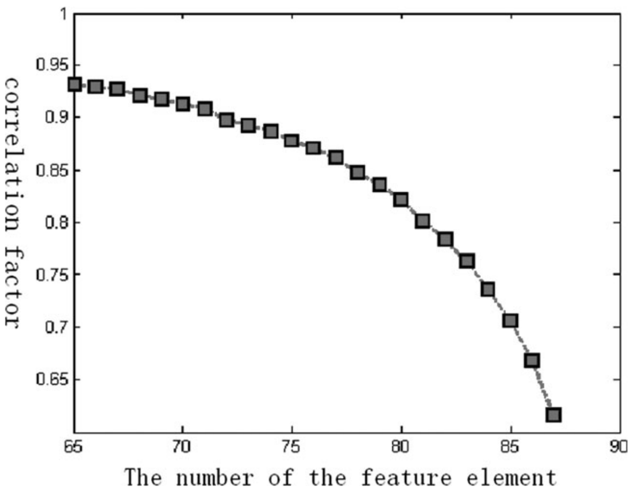


Figure 5
A schematic drawing to show the 23 correlation coefficients between predicting and experimental data of protein folding rate using Sequential backward selection method.

Comparison with different methods

To judge how well our predictor is, we have used the Fold-rate[15] without structural information to predict the folding rates of the 79 proteins, and we get the correlation coefficient 0.6414 between experimental folding rates and predicting folding rates. The set is also tested on the Pred-PFR,[40] CI[11], and Nα[21], and the testing results prove that our multifeature SVM-regression method is better than other sequence-based methods (CI, Fold-rate, Pred-PFR, and Nα) in Table III. Our method not only has better correlation between predicted rates and experimental rates than all the sequence-based method but also

Table III
Comparison among Different Folding Rate Prediction Methods Based on the Set of 79 Proteins

	R *value*	σ value
Pred-PFR^a	−0.0479	9.6308
Fold-rate^b	0.1764	7.8818
CI^c	−0.0729	8.6628
$N_\alpha{}^d$	−0.0964	22.2669
Present method	0.9313	2.2692

All the methods were tested on the set of 79 proteins. *R* value is correlation coefficient, and σ value is standard error.[a]Result from the Pred-PFR web server at http://www.csbio.sjtu.edu.cn/bioinf/FoldingRate/.
[b]Result from the Fold-Rate web server at http://psfs.cbrc.jp/fold-rate/.
[c]Result from the CI web server at http://ibi.hzau.edu.cn/FDserver/.
[d]Result from the N_α web server at http://gila.bioengr.uic.edu/lab/tools/foldingrate/fr0.html.

has smaller standard error values between predicted and real rates than all the sequence-based methods.

Limitations of the present method and possible improvements

In the present investigation, we develop a method to predict the folding rates of proteins based only on protein sequence, without any explicit structural information. However, the folding rate of two and three state protein may be governed by different factors. Consequently, revealing the influence of these different factors on the rates of protein folding represents future research efforts.

CONCLUSIONS

The extremely large numbers of sequence order patterns in proteins and their diverse lengths have made it very difficult to accommodate the protein sequence order effects. To tackle this issue, Chou's PseAAC combined with a new method, sliding window incorporating long-range and short-range contact in protein was presented to approximate the sequence order effects. We observed that the predicted folding rates using the method show an excellent agreement with experimental results. A web server has been developed for the prediction purpose, and the results are available online, which may be very helpful for the users to get the folding rate of any protein with its sequence.

REFERENCES

1. Wikipedia. Available at: http://www.en.wikipedia.org/wiki/Protein_-folding. Accessed April 11, 2012.
2. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol 1998;277:985–994.
3. Gromiha MM, Selvaraj S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: application of long-range order to folding rate prediction. J Mol Biol 2001;310:27–32.
4. Zhou H, Zhou Y. Folding rate prediction using total contact distance. Biophys J 2002;82:458–463.
5. Nolting B, Schalike W, Hampel P, Grundig F, Gantert S, Sips N, Bandlow W, Qi PX. Structural determinants of the rate of protein folding. J Theor Biol 2003;223;299–307.
6. Zhang L, Sun T. Folding rate prediction using n-order contact distance for proteins with two- and three-state folding kinetics. Biophys Chem 2005;113:9–16.
7. Ivankov ,DN, Finkelstein AV. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc Natl Acad Sci USA 2004;101:8942–8944.
8. Gong H, Isom DG, Srinivasan R, Rose GD. Local secondary structure content predicts folding rates for simple, two-state proteins. J Mol Biol 2003;327:1149–1154.
9. Punta M, Rost B. PROFcon: novel prediction of long-range contacts. Bioinformatics 2005;21:2960–2968.
10. Punta M, Rost B. Protein folding rates estimated from contact predictions. J Mol Biol 2005;348:507–512.
11. Ma BG, Guo JX, Zhang HY. Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. Proteins 2006;65:362–372.
12. Huang LT, Gromiha MM. Analysis and prediction of protein folding rates using quadratic response surface models. J Comput Chem 2008;29:1675–1683.
13. Guo JX, Rao NN, Liu GX, Yang Y, Wang G. Predicting protein folding rates using the concept of Chou's pseudo amino acid composition. J Comput Chem 2011;32:1612–1617.
14. Capriotti E, Casadio R. K-Fold: a tool for the prediction of the protein folding kinetic order and rate. Bioinformatics 2007;23:385–386.
15. Debe DA, Goddard WA. First principles prediction of protein folding rates. J Mol Biol 1999;294:619–625.
16. Fulton KF, Devlin GL, Jodun RA, Silvestri L, Bottomley SP, Fersht AR, Buckle AM. PFD: a database for the investigation of protein folding kinetics and stability. Nucleic Acids Res 2005;33:D279–D283.
17. Gromiha MM. A statistical model for predicting protein folding rates from amino acid sequence with structural class information. J Chem Inf Model 2005;45:494–501.
18. Gromiha MM, Thangakani AM, Selvaraj S. FOLD-RATE: prediction of protein folding rates from amino acid sequence. Nucleic Acids Res 2006;34:W70–W74.
19. Ivankov DN, Garbuzynskiy SO, Alm E, Plaxco KW, Baker D, Finkelstein AV. Contact order revisited: influence of protein size on the folding rate. Protein Sci 2003;12:2057–2062.
20. Jiang Y, Iglinski P, Kurgan L. Prediction of protein folding rates from primary sequences using hybrid sequence representation. J Comput Chem 2009;30:772–783.
21. Ouyang Z, Liang J. Predicting protein folding rates from geometric contact and amino acid sequence. Protein Sci 2008;17:1256–1263.
22. Gromiha MM, Selvaraj S. Importance of long-range interactions in protein folding. Biophys Chem 1999;77:49–68.
23. Gromiha MM, Selvaraj S. Influence of medium and long range interactions in protein folding. Prep Biochem Biotechnol 1999;29:339–351.
24. Cai YD, Chou KC. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 2005;4:967–971.
25. Cai YD, Chou KC. Predicting membrane protein type by functional domain composition and pseudo-amino acid composition. J Theor Biol 2006;238:395–400.
26. Cai YD, Zhou GP, Chou KC. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 2005;234:145–149.
27. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 2005;21:10–19.
28. Chou KC, Cai YD. Predicting protein quaternary structure by pseudo amino acid composition. Proteins 2003;53:282–289.
29. Chou KC, Cai YD. Prediction and classification of protein subcellular location-sequence-order effect and pseudo amino acid composition. J Cell Biochem 2003;90:1250–1260.
30. Liu L, Hu XZ, Liu XX, Wang Y, Li SB. Predicting protein fold types by the general form of Chou's pseudo amino acid composition: approached from optimal feature extractions. Protein Pept Lett 2012;19:439–449.
31. Shen HB, Yang J, Chou KC. Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. J Theor Biol 2006;240:9–13.
32. Wang J, Li Y, Wang Q, You X, Man J, Wang C. Gao X. ProClusEnsem: predicting membrane protein types by fusing different modes of pseudo amino acid composition. Comput Biol Med 2012;42:564–574.
33. Xiao X, Lin WZ, Chou KC. Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 2008;29:2018–2024.

34. Xiao X, Shao SH, Huang ZD, Chou KC. Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 2006;27:478–482.

35. Xiao X, Wang P, Chou KC. Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J Theor Biol 2008;254:691–696.

36. Xiao X, Wang P, Chou KC. GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol Biosyst 2011;7:911–919.

37. Xiao X, Wang P, Chou KC. Quat-2L: a web-server for predicting protein quaternary structural attributes. Mol Divers 2011;15:149–155.

38. Zhou GP, Cai YD. Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins 2006;63: 681–684.

39. Tomii ,K, Kanehisa ,M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 1996;9:27–36.

40. Shen HB, Song JN, Chou KC. Prediction of protein folding rates from primary sequence by fusing multiple sequential features, J Biomed Sci Eng 2009;2:135–207.