Analysis and Prediction of Protein Folding Rates Using Quadratic Response Surface Models

LIANG-TSUNG HUANG,[1] M. MICHAEL GROMIHA[2]
[1]*Department of Computer Science and Information Engineering, Ming-Dao University, Changhua 523, Taiwan*
[2]*Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), AIST Tokyo Waterfront Bio-IT Research Building, 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan*

Received 28 June 2007; Accepted 28 December 2007
DOI 10.1002/jcc.20925
Published online 19 March 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: Understanding the relationship between amino acid sequences and folding rates of proteins is an important task in computational and molecular biology. In this work, we have systematically analyzed the composition of amino acid residues for proteins with different ranges of folding rates. We observed that the polar residues, Asn, Gln, Ser, and Lys, are dominant in fast folding proteins whereas the hydrophobic residues, Ala, Cys, Gly, and Leu, prefer to be in slow folding proteins. Further, we have developed a method based on quadratic response surface models for predicting the folding rates of 77 two- and three-state proteins. Our method showed a correlation of 0.90 between experimental and predicted protein folding rates using leave-one-out cross-validation method. The classification of proteins based on structural class improved the correlation to 0.98 and it is 0.99, 0.98, and 0.96, respectively, for all-α, all-β, and mixed class proteins. In addition, we have utilized Baysean classification theory for discriminating two- and three-state proteins, which showed an accuracy of 90%. We have developed a web server for predicting protein folding rates and it is available at http://bioinformatics.myweb.hinet.net/foldrate.htm.

 J Comput Chem 29: 1675–1683, 2008

Key words: protein folding rates; amino acids; structural class; prediction

Introduction

Predicting the folding rate of a protein from its amino acid sequence is an important problem for understanding the variations in protein folding kinetics, which may lead to several pathologies such as prion and Alzheimer diseases. Folding rate is a measure of slow/fast folding of a protein from its unfolded state to the native three-dimensional structure. Several structural parameters have been developed to predict the protein folding rates from the information of inter-residue contacts.[1] These parameters emphasize the importance of topology of the folded state to understand the protein folding rates.

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 rates of two-state proteins. Gromiha and Selvaraj[3] defined a novel parameter, long-range order (LRO) from the knowledge of long-range contacts (contact between two residues that are close in space and far in the sequence) in protein structure and established a simple statistical model for predicting the protein folding rates. Miller et al.[4] experimentally demonstrated that LRO is one of the best parameters that correlate with protein-refolding rates including circular permutations of ribosomal proteins S6 from *Thermus thermophilus*. These two parameters, CO and LRO, are incorporated into a new parameter, total contact distance (TCD), which shows a good relationship with protein folding rates.[5]

In addition, several investigations have been carried out to understand/predict the folding rates of proteins from protein three-dimensional structures. These studies include the first principles of protein folding,[6] elementary statistical model,[7] combination of CO and stability,[8] number of native contacts,[9] topomer search model,[10] the topological properties of protein conformation,[11] neural networks based on CO, LRO and TCD,[12] amino acid properties,[13] chain length,[14] size,[15] helix parameter[16] and native state geometry,[17] chain topology,[18] and n-order contact distance.[19] Recently, different methods have been proposed for

This article contains supplementary material available via the Internet at http://www.interscience.wiley.com/jpages/0192–8651/suppmat

Correspondence to: M. M. Gromiha; e-mail: Michael-gromiha@aist.go.jp

predicting protein folding rates from amino acid sequence, secondary structure, and structural class information.[20–26]

In this work, we have developed a method based on quadratic response surface models (QRSM) for predicting protein folding rates from amino acid sequence. Our method could predict the folding rates of 77 two- and three-state proteins with a correlation of 0.90 and the classification of proteins into all-α, all-β, and mixed class improved the correlation up to 0.98. Further, we proposed a method based on discriminant rules for discriminating two- and three-state proteins, which showed an accuracy of 90%. A web server has been set up for predicting protein folding rates and it is available at http://bioinformatics.myweb.hinet.net/foldrate.htm. Further, the composition of amino acid residues for proteins with different ranges of folding rates will be discussed.

Materials and Methods

Experimental Folding Rates

The experimental folding rates of 77 two- and three-state proteins used in related works[14,21,22,26] form the basis for this study. The Protein Data Bank codes[25] and experimental $\ln(k_f)$ values are given in Table 1. The structural classification of these proteins yielded 16 all-α (dominated by α-helices; $\alpha > 40\%$ and $\beta < 5\%$), 26 all-β (dominated by β-strands; $\beta > 40\%$ and $\alpha < 5\%$), and 35 mixed class proteins (contain both α-helices and β-strands; $\alpha > 15\%$ and $\beta > 10\%$).

Amino Acid Properties

We used a set of 49 diverse amino acid properties (physical–chemical, energetic, and conformational), which fall into various clusters analyzed by Tomii and Kanehisa[27] in this study. The amino acid properties were normalized between 0 and 1 using the expression, $P_{\text{norm}}(i) = [P(i) - P_{\text{min}}]/[P_{\text{max}} - P_{\text{min}}]$, where $P(i)$, and $P_{\text{norm}}(i)$ are, respectively, the original and normalized values of amino acid i for a particular property, and P_{min} and P_{max} are, respectively, the minimum and maximum values. The numerical and normalized values for all the 49 properties used in this study along with their brief descriptions have been explained in our earlier articles[28,29] and are available at http://www.cbrc.jp/~gromiha/fold_rate/property.html as well as in supplementary information (Table S1).

Computational Procedure

The average amino acid property for each protein, $P_{\text{ave}}(i)$ was computed using the standard formula,

$$P_{\text{ave}}(i) = \sum_{j=1}^{n} P(j)/n \tag{1}$$

where, $P(j)$ is the property value of jth residue and the summation is over n, the total number of residues in a protein. The computed property value $P_{\text{ave}}(i)$ for each class of proteins was related with experimental folding rate $\ln k_f(i)$ using single correlation coefficient.

Quadratic Response Surface Models

A response surface is a model that tries to reproduce the system responses due to changes in input variables. In this study, the QRSM is used to establish a simplified relationship between multiple input variables and one output variable by using the equation,

$$y = b_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m} \sum_{j=i}^{m} b_{ij} x_i x_j + e \tag{2}$$

where, y is the output variable (folding rate); b_i and b_{ij} are coefficients of the input variables x_i and x_j (amino acid properties); m is the total number of selected amino acid properties; uncontrolled factors and errors are modeled by e. Given n independent observations $(\mathbf{x}^1,y^1),\ldots,(\mathbf{x}^n,y^n)$, the model becomes an n-by-p system of equations:

$$\begin{bmatrix} y^1 \\ \vdots \\ y^n \end{bmatrix} = \begin{bmatrix} f_1(x^1) & \cdots & f_p(x^1) \\ \vdots & \ddots & \vdots \\ f_1(x^n) & \cdots & f_p(x^n) \end{bmatrix} \begin{bmatrix} c_1 \\ \vdots \\ c_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}, \tag{3}$$

where y^k is the output variable value of folding rate; $\mathbf{x}^k = (x_1^k, \ldots, x_m^k)$ is the input variable vector of property values for the k-th observation; $f_l(\mathbf{x}^k)$, $l = 1\ldots p$, is the l-th transferred term of the model (e.g. terms are $f_1(\mathbf{x}^1) = 1$, $f_2(\mathbf{x}^1) = x_1^1$ and $f_3(\mathbf{x}^1) = x_2^1$), and the value of p is represented as $p = 1 + m + m(m + 1)/2$ (m, the total number of selected amino acid properties); c_l and e_l are the corresponding coefficient and error values, respectively. Therefore, the estimates of the model coefficients are determined by using the least square method which minimizes the statistics derived from errors.

The main advantages of applying the model to predict protein folding rates include: (i) Being nonparametric to suit data with unknown as well as skewed distribution, no assumptions are required or made regarding the underlying distribution of independent variables; (ii) The quadratic response surface model is a nonlinear but low-order model. Hence, relatively few observations are required to build a model relating inputs and outputs. (iii) Lower model complexity may directly help to reduce computational cost and time. Because of those abilities mentioned earlier, its related concepts and techniques have been broadly applied in many branches of engineering, especially in the chemical and manufacturing areas.[30–34]

Maximum Likelihood Discriminant Rules

On the basis of Bayes decision theory, the maximum likelihood (ML) discriminant rule[35] discriminates the class of a feature vector x by assigning the one which yields maximal likelihood. For multivariate Gaussian distributions, the likelihood function of ω_i with respect to x in the l-dimensional feature space is given by

$$p(x|\omega_i) = \frac{1}{(2\pi)^{l/2}|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_i)^T \sum_i^{-1} (x-\mu_i)\right], \tag{4}$$

where μ_i is the mean of x for the ω_i class, $\sum_i$ the l by l covariance matrix. When the covariance matrices are diagonal, $\sum_i = \text{diag}(\sigma_{i1}^2, \ldots, \sigma_{il}^2)$, the ML discriminate rule can be written

Table 1. Predicted Folding Rates in a Set of 77 Two- and Three-State Proteins with Different Test Types.

	Protein			$\ln(k_f)$ predicted (s^{-1})		
					Jack-knife	
PDB code	Structural class	State	$\ln(k_f)$ experimental (s^{-1})	Self-consistency	All	Classified
1A6N	All-α	Three	1.10	0.96	1.20	1.10
1BDD	All-α	Two	11.75	11.76	12.49	11.75
1CEI	All-α	Three	5.80	5.66	7.39	5.80
1EBD	All-α	Two	9.68	9.55	11.25	9.72
1ENH	All-α	Two	10.53	10.47	14.17	10.39
1HRC	All-α	Two	8.76	8.93	6.39	8.76
1IMQ	All-α	Two	7.31	7.44	7.20	7.39
1L8W	All-α	Two	1.61	1.60	−6.00	1.54
1LMB	All-α	Two	8.50	8.28	8.03	8.50
1VII	All-α	Two	11.52	11.60	15.24	11.53
1YCC	All-α	Two	9.62	9.47	12.53	9.62
256B	All-α	Two	12.20	12.24	10.92	12.20
2A5E	All-α	Three	3.50	3.51	3.56	3.50
2ABD	All-α	Two	6.55	6.47	4.75	6.56
2CRO	All-α	Three	3.70	3.70	1.11	3.71
2PDD	All-α	Two	9.80	9.89	8.27	9.79
1C8C	All-β	Two	6.91	6.98	8.78	6.87
1C9O	All-β	Two	7.20	7.32	6.39	6.90
1CBI	All-β	Three	−3.20	−2.93	−3.50	−3.12
1CSP	All-β	Two	6.98	6.98	8.56	8.17
1EAL	All-β	Three	1.30	1.60	1.02	0.97
1FNF-10	All-β	Three	5.48	5.52	7.07	5.11
1FNF-9	All-β	Two	0.91	0.80	1.90	2.09
1G6P	All-β	Two	6.30	6.28	4.69	7.11
1HNG	All-β	Three	2.89	3.09	3.47	3.24
1HX5	All-β	Three	0.74	0.70	4.46	0.35
1IFC	All-β	Three	3.40	3.19	2.59	3.39
1LOP	All-β	Two	6.60	6.31	6.72	7.29
1MJC	All-β	Two	5.24	5.18	5.77	5.20
1NYF	All-β	Two	4.54	4.70	2.97	5.28
1OPA	All-β	Three	1.40	1.42	1.66	1.16
1PIN	All-β	Two	9.44	9.70	5.47	9.02
1PKS	All-β	Two	−1.05	−1.05	−5.73	−0.83
1PNJ	All-β	Two	1.10	0.94	2.52	−0.38
1PSF	All-β	Three	3.22	3.28	7.77	3.10
1SHF	All-β	Two	4.50	4.37	6.70	3.85
1SHG	All-β	Two	1.41	1.42	2.05	1.54
1SRL	All-β	Two	4.04	4.00	1.99	4.27
1TEN	All-β	Two	1.06	1.04	0.06	1.02
1TIT	All-β	Three	3.47	3.51	5.86	2.61
1WIT	All-β	Two	0.41	0.53	−0.95	0.89
2AIT	All-β	Two	4.20	4.22	3.92	4.92
1AON	Mixed	Three	0.80	0.74	2.21	1.25
1APS	Mixed	Two	−1.48	−1.65	−2.93	−0.93
1AYE	Mixed	Two	6.80	6.92	5.02	6.74
1BNI	Mixed	Three	2.60	2.63	−0.51	1.78
1BRS	Mixed	Two	3.40	3.37	6.62	2.56
1CIS	Mixed	Two	3.87	3.93	2.95	4.07
1COA	Mixed	Two	3.87	3.71	5.08	3.80
1DIV	Mixed	Two	6.58	6.46	9.73	7.35
1FKB	Mixed	Two	1.46	1.20	−0.58	1.66
1GXT	Mixed	Three	4.38	4.36	4.82	4.88
1HDN	Mixed	Two	2.70	2.70	2.69	2.78
1HZ6	Mixed	Two	4.10	4.06	2.53	2.67

(*continued*)

Table 1. (*Continued*).

PDB code	Protein		$\ln(k_f)$ experimental (s^{-1})	$\ln(k_f)$ predicted (s^{-1})		
					Jack-knife	
	Structural class	State		Self-consistency	All	Classified
1PBA	Mixed	Two	6.80	6.80	8.57	9.25
1PCA	Mixed	Two	6.80	6.69	4.07	5.95
1PGB	Mixed	Two	6.00	6.03	4.23	7.11
1PHP C-terminal	Mixed	Three	−3.45	−3.33	−2.27	−2.94
1PHP N-terminal	Mixed	Three	2.30	2.28	3.97	1.30
1POH	Mixed	Two	2.70	2.70	2.69	2.78
1qop alpha-subunit	Mixed	Three	−2.53	−2.59	−2.34	−3.40
1qop beta-subunit	Mixed	Three	−6.91	−6.80	−8.93	−6.43
1RA9	Mixed	Three	−2.50	−2.63	−2.56	−2.17
1RIS	Mixed	Two	5.90	6.03	7.84	6.77
1SCE	Mixed	Three	4.20	4.25	3.81	1.72
1UBQ	Mixed	Two	7.33	6.63	5.97	5.98
1UBQ	Mixed	Two	5.90	6.63	7.33	7.18
1URN	Mixed	Two	5.73	5.72	6.59	6.41
2A5E	Mixed	Three	3.50	3.51	3.56	5.13
2ACY	Mixed	Two	0.92	0.94	1.87	0.89
2CI2	Mixed	Two	3.90	3.99	2.85	4.03
2HQI	Mixed	Two	0.18	0.12	3.64	−1.34
2LZM	Mixed	Three	4.10	4.08	2.71	4.21
2PTL	Mixed	Two	4.10	4.09	6.34	1.78
2RN2	Mixed	Three	0.10	0.08	1.22	−0.10
2VIK	Mixed	Two	6.80	6.82	8.23	7.58
3CHY	Mixed	Three	1.00	1.30	1.01	0.02

All, A single dataset with all the 77 considered proteins.
Classified, proteins are classified into three classes, all-α, all-β, and mixed.

as $C(x) = \arg\min_l \sum_{j=1}^{l} [(x_j - \mu_{ij})^2/\sigma_{ij}^2 + \log \sigma_{ij}^2]$. In practice, μ_i and $\sum_i$ are estimated by corresponding sample quantities. In this study, we used the combination of properties as the feature vector to discriminate two and three-state proteins.

Self-Consistency and Jack-Knife Tests

We have used both self consistency and jack-knife test (leave-one-out cross-validation method) for assessing the performance of the method.[36,37] In self consistency test, we have used all the 77 proteins to derive the coefficients in QRSM model and used the same to predict the folding rate. We have performed jack-knife test by determining the coefficients of QRSM using $(n - 1)$ data (i.e., omitting one protein at a time) and then computing the folding rate of the omitted protein.

Results and Discussions

Compositional Preference of Amino Acid Residues in Proteins with Various Ranges of Folding Rates

We have computed the preference of amino acid residues in proteins with different ranges of folding rates, −9 to −6 s^{-1}, −6 to −3 s^{-1}, −3 to 0 s^{-1}, 0– 3 s^{-1}, 3–6 s^{-1}, 6–9 s^{-1} and >9 s^{-1} and the results are presented in Table 2. We observed that the polar residues Asn, Gln, Lys, and Ser are predominant in fast folding proteins. The results have been compared with the tendency of residues that form short, medium, and long-range contacts in protein structures as explained in our earlier articles.[1,38] In this procedure, the residues in a protein molecule are represented by their α-carbon atoms. Using the C_α coordinates, a sphere of radius 8 Å is fixed around each residue and the residues occurring in this volume are identified. The contacting residue pairs are analyzed in terms of the location at the sequence level and the contributions from less than ±3 residues are treated as short-range, ±3 or ±4 residues as medium-range and more than ±4 residues are treated as long-range contacts.[1,38] We suggest that polar residues are dominated by short and medium-range contacts in the formation of hydrogen bonds and ion pairs. The structural analysis on fast folding proteins showed that the short and medium-range contacts between polar residues, such as NN, SQ, QE, QK, and QS are dominant in fast folding proteins. On the other hand, the occurrence of hydrophobic residues Ala, Cys, Gly, and Leu is high in slow folding proteins. This might be due to the fact that the formation of hydrophobic core involves long-range interactions, which slows down the folding process.[3] To verify this speculation we have computed the long-range contacts in slow folding proteins.[1,38] We observed that the long-range contacts in slow folding proteins are mainly influenced with the hydrophobic residue pairs, such as AA, AG, GG, WL, MG, and CY.

Table 2. Frequency and Percentage Distribution of Folding Rates for 20 Types of Residues.

Residue	ln(k_f) range (s^{-1})						
	−9 to −6 (1)	−6 to −3 (2)	−3 to 0 (4)	0 to 3 (21)	3 to 6 (26)	6 to 9 (15)	>9 (8)
A	42 [10.80]	34 [10.00]	57 [9.50]	236 [9.70]	192 [8.70]	91 [6.70]	56 [9.40]
R	17 [4.40]	15 [4.40]	36 [6.00]	97 [4.00]	130 [5.90]	37 [2.70]	40 [6.70]
N	11 [2.8]0	13 [3.80]	24 [4.00]	84 [3.40]	91 [4.10]	58 [4.30]	33 [5.50]
D	18 [4.60]	26 [7.60]	36 [6.00]	150 [6.20]	133 [6.00]	72 [5.30]	33 [5.50]
C	5 [1.30]	3 [0.90]	6 [1.00]	11 [0.50]	15 [0.70]	6 [0.40]	5 [0.80]
Q	17 [4.40]	3 [0.90]	19 [3.20]	79 [3.20]	74 [3.30]	66 [4.90]	31 [5.20]
E	27 [6.90]	28 [8.20]	45 [7.50]	195 [8.00]	189 [8.60]	133 [9.80]	44 [7.40]
G	42 [10.80]	30 [8.80]	49 [8.20]	216 [8.90]	158 [7.20]	122 [9.00]	45 [7.60]
H	14 [3.60]	5 [1.50]	11 [1.80]	47 [1.90]	37 [1.70]	32 [2.40]	11 [1.80]
I	23 [5.90]	20 [5.90]	38 [6.40]	138 [5.70]	123 [5.60]	73 [5.40]	26 [4.40]
L	38 [9.70]	27 [7.90]	47 [7.90]	198 [8.10]	193 [8.70]	85 [6.30]	51 [8.60]
K	18 [4.60]	27 [7.90]	26 [4.30]	219 [9.00]	138 [6.20]	137 [10.10]	61 [10.20]
M	14 [3.60]	8 [2.30]	12 [2.00]	37 [1.50]	39 [1.80]	34 [2.50]	12 [2.00]
F	13 [3.30]	15 [4.40]	24 [4.00]	83 [3.40]	71 [3.20]	70 [5.20]	24 [4.00]
P	18 [4.60]	11 [3.20]	32 [5.40]	80 [3.30]	96 [4.30]	37 [2.70]	22 [3.70]
S	20 [5.10]	14 [4.10]	40 [6.70]	121 [5.00]	110 [5.00]	61 [4.50]	37 [6.20]
T	21 [5.40]	21 [6.20]	23 [3.80]	169 [6.90]	136 [6.20]	79 [5.80]	24 [4.00]
W	1 [0.30]	5 [1.50]	8 [1.30]	29 [1.20]	34 [1.50]	18 [1.30]	6 [1.00]
Y	12 [3.10]	7 [2.10]	22 [3.70]	61 [2.50]	78 [3.50]	39 [2.90]	14 [2.30]
V	19 [4.90]	29 [8.50]	43 [7.20]	187 [7.70]	172 [7.80]	102 [7.50]	21 [3.50]

Number of proteins in each range is given in parenthesis.
Values in square bracket indicate percentage distribution of folding rates.

In addition, we have carried out statistical analyses to examine the confidence in the differences of amino acid compositions among slow and fast folding proteins. We have performed the χ^2-test of independence for two categorical variables, namely, the residue composition and the protein folding rate range. Along with Table 2, the residue is regarded as 20 levels and the folding rate ranges into 5 levels, where the range from −9 to –3 s^{-1} with fewer observations was removed.

As a test of independence, we followed the null and alternative hypotheses as (i) the two variables are independent and (ii) the two variables are dependent (i.e., there is a relationship between them). We obtained a p-value of 4.10 E−10, which is less than the significance level of $\alpha = 0.05$, and this result confirmed the existence of relationship between amino acid composition and protein folding rates.

Table 3. The Correlation Coefficients (r) Between the Compositions of Amino Acid Residues at Different Ranges of Folding Rates.

r	−9 to −6	−6 to −3	−3 to 0	0 to 3	3 to 6	6 to 9	>9
−9 to −6	1.00	0.78	0.86	0.83	0.82	0.67	**0.74**
−6 to −3	0.78	1.00	0.85	0.96	0.92	0.84	**0.76**
−3 to 0	0.86	0.85	1.00	0.84	0.92	0.68	**0.75**
0 to 3	0.83	0.96	0.84	1.00	0.94	0.91	0.84
3 to 6	0.82	0.92	0.92	0.94	1.00	0.81	0.80
6 to 9	**0.67**	0.84	0.68	0.91	0.81	1.00	0.77
>9	**0.74**	0.76	0.75	0.84	0.80	0.77	1.00

The r values between the amino acid compositions in slow and fast folding proteins are shown in bold.

Further, we have analyzed the relationship of amino acid compositions in slow and fast folding proteins and the results are presented in Table 3. Interestingly, the correlation is very high (>0.8) between the proteins with the marginal difference in their folding rates. On the other hand, there is a marked difference between slow and fast folding proteins and the correlation lies in the range of 0.67–0.76. This result suggest that there is a difference of amino acid compositions in slow and fast folding proteins and the composition is an important parameter for understanding protein folding rates.

Prediction of Protein Folding Rates

We have analyzed the relationship between amino acid properties and protein folding rates and we observed that the correlation lies in the range of −0.47 to 0.44 (Table 4). Hence, we combined the amino acid properties using QRSM method. The results obtained with self consistency and jack-knife tests are presented in Table 1. We observed that our model could fit all the data together and the mean absolute error between experimental and predicted folding rates using self-consistency test is 4%. In the jack-knife test our method showed a correlation of 0.90 between experimental and predicted folding rates. We noticed that few proteins have a large difference from the experimental ones and especially, 1ENH, 1L8W, 1HX5, 1PIN, 1PKS, 1PSF, 1BNI, 1BRS, 1DIV, and 2HQI. These proteins fall into different structural classes and the folding behavior may be different in these proteins.[1,38] It has been reported that topology is a major factor for determining the folding rates.[13,39] Hence, we classified the proteins based on different structural classes and

Table 4. Correlation Between Folding Rates of Proteins and 49 Various Amino Acid Properties.

Property no.	Property name	Correlation coefficient
1	K^0	−0.18
2	H_t	−0.10
3	H_p	−0.35
4	P	0.26
5	pH_i	0.24
6	pK'	0.02
7	Mw	0.23
8	B_l	−0.02
9	R_f	−0.25
10	μ	0.20
11	H_{nc}	−0.42
12	E_{sm}	−0.27
13	E_l	−0.38
14	E_t	−0.47
15	P_α	0.17
16	P_β	−0.24
17	P_t	0.02
18	P_c	−0.07
19	C_a	0.26
20	F	0.20
21	B_r	−0.42
22	R_a	−0.24
23	N_s	−0.41
24	α_n	0.08
25	α_c	−0.08
26	α_m	0.33
27	V^0	0.21
28	N_m	0.29
29	N_l	−0.43
30	H_{gm}	−0.35
31	ASA_D	0.27
32	ASA_N	0.44
33	ΔASA	0.02
34	ΔG_h	−0.21
35	G_{hD}	−0.29
36	G_{hN}	−0.21
37	ΔH_h	−0.16
38	$-T\Delta S_h$	0.01
39	ΔG_{ph}	−0.20
40	ΔG_c	0.19
41	ΔH_c	0.04
42	$-T\Delta S_c$	0.19
43	ΔG	0.00
44	ΔH	−0.16
45	$-T\Delta S$	0.20
46	v	0.23
47	s	0.12
48	f	0.37
49	P_{f-s}	0.26

derived three equations for all-α, all-β, and mixed class proteins. This classification remarkably improved the correlation between experimental and predicted folding rates as seen in Figure 1. The correlation coefficient is 0.99, 0.98, and 0.96, respectively for all-α, all-β, and mixed class proteins and the overall correlation is 0.98. From Table 1, we can see that the protein folding rates can be predicted within the mean relative error of 20%. Further, the mean error is 10% for fast folding proteins and the folding rates of 68% of the proteins are predicted within the relative error of 15%. Detailed analysis of the results presented in Table 1 showed that the relative error is high for few proteins and this might be due to the difference of experimental conditions used to measure the folding rates, which are not considered in this work.

Factors Governing the Folding Rates of Proteins

We have utilized the information about amino acid composition and amino acid properties for understanding/predicting protein folding rates. In the prediction model, we have used the average amino acid property of a protein, which are the combined effect of compositional preference as well as the specific nature of amino acids (property). This has been well explained with the correlation between total amino acid property and proteins folding rates (Table 4). We found a negative correlation for the properties reflecting long-range interactions and hydrophobicity (for example, H_p, H_{nc}, P_β, N_s, N_l etc.). On the other hand, properties reflecting electrostatic, physical, and short- and medium-range interactions have positive correlation with protein folding rates (e.g. P, pH_i, P_α, N_m etc.).

The analysis of protein three-dimensional structures showed that topology is a major determinant for protein folding rates,[39] which is evidenced by the development of the parameters, CO,[2] and LRO[3]. In addition, chain length,[14] size,[15] helix parameter[16] and native state geometry,[17] and secondary structure content[20,21] are reported to play important roles to understand/predict protein

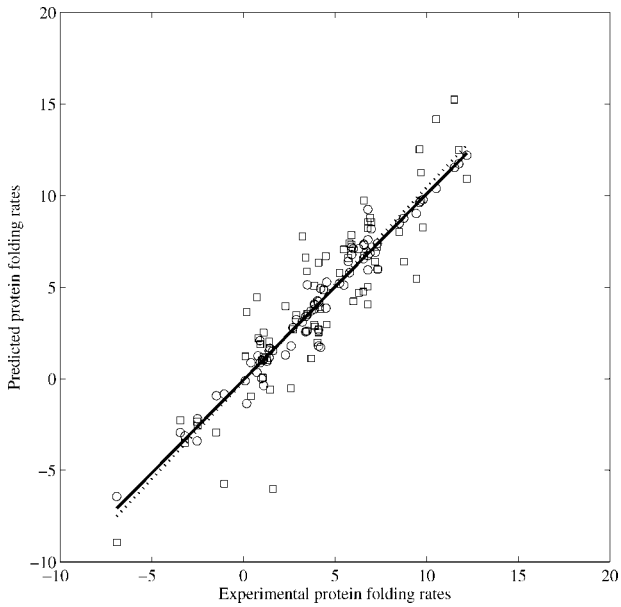
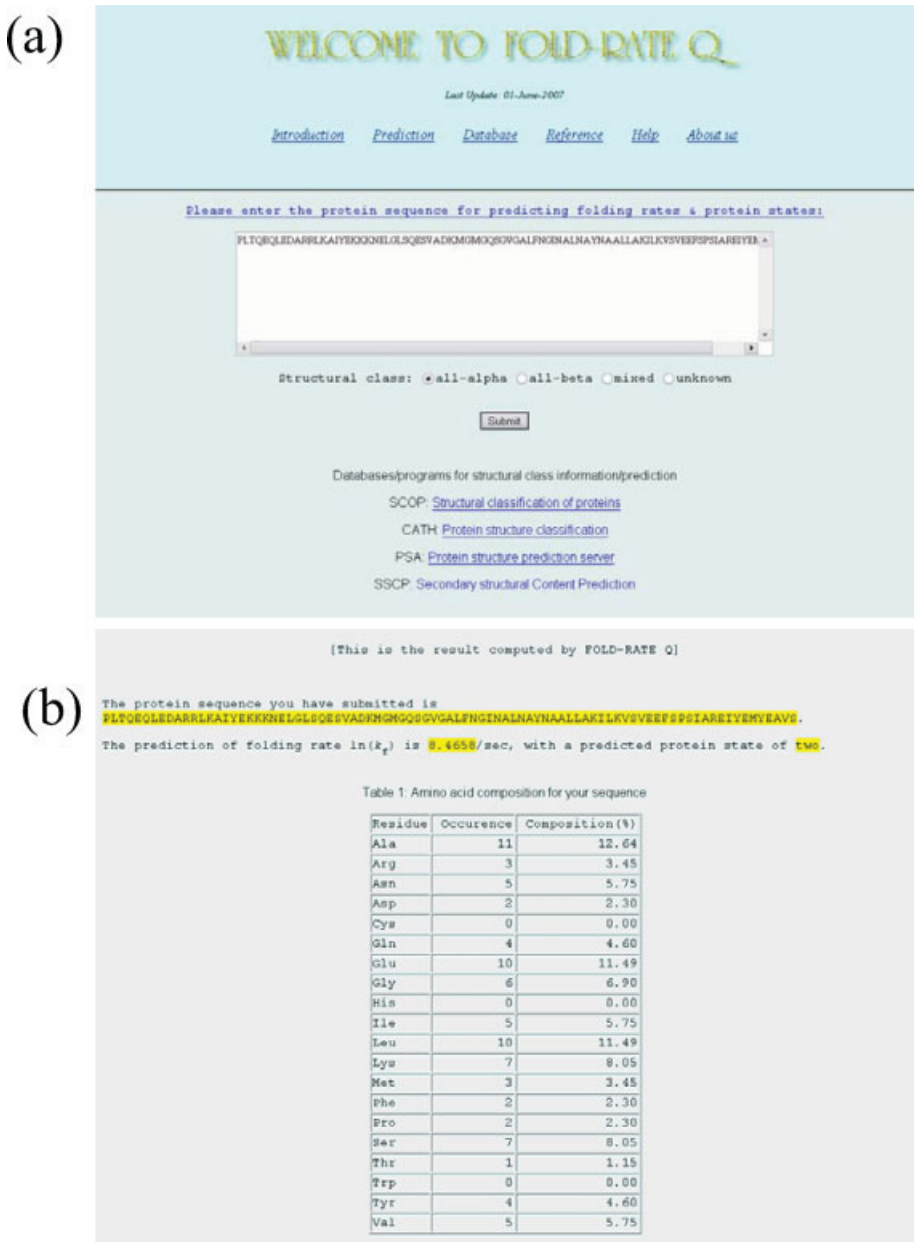


Figure 1. Relationship between experimental and predicted ln (k_f) values with jack-knife test in a set of 77 two- and three-state proteins; Squares and the dotted line denote the relationship before classification based on structural class, and circles and the solid line denote the same after classification.



Residue	Occurence	Composition(%)
Ala	11	12.64
Arg	3	3.45
Asn	5	5.75
Asp	2	2.30
Cys	0	0.00
Gln	4	4.60
Glu	10	11.49
Gly	6	6.90
His	0	0.00
Ile	5	5.75
Leu	10	11.49
Lys	7	8.05
Met	3	3.45
Phe	2	2.30
Pro	2	2.30
Ser	7	8.05
Thr	1	1.15
Trp	0	0.00
Tyr	4	4.60
Val	5	5.75

Figure 2. Snapshot showing the necessary items to be given as input for prediction of protein folding rates (a) and protein states (b). The results obtained for predicting the protein folding rates and the protein states along with the composition of amino acid residues for lambda-repressor (1LMB).

folding rates. The parameters, CO, and LRO are derived from the residue contacts in protein structures and the amino acid properties, number of medium-range contacts, long-range contacts, short and medium-range nonbonded energy, long-range nonbonded energy, number of surrounding residues, etc., reflect the tendency of CO and LRO. The chain length and size may be represented with volume, size, bulkiness, etc. The propensities of amino acids in α-helical, β-strand, turn, and coil regions resemble the helix geometry and secondary structure content. In essence, all the parameters used to predict protein folding rates have been obtained with protein sequence/structure information (amino acid sequence, secondary structure, and tertiary structure) and hence they have direct/indirect relationship with the amino acid properties used in this study.

Folding Rates of Naturally Disordered Proteins

We have predicted the folding rates of 69 disordered proteins[40] using the present method and the results are presented in Supplementary Table S2. We observed an average folding rate,

$\ln(k_f)$ of 0.82 s^{-1} for a set of 53 sample proteins that have the range of -21 to 21 s^{-1}, which is significantly less than that for the 77 two- and three-state proteins considered in this work (4.06 s^{-1}). However, most of the disordered proteins have the folding rates, which are similar to that of native proteins and hence our method may not be utilized directly for discriminating foldable and disordered proteins. Instead, our method can be successfully used for discriminating slow and fast folding proteins as well as predicting the folding rates of native proteins. On the other hand, other sequence/structure based methods can be used for discriminating disordered proteins.[40]

Discrimination of Two- and Three-State Proteins Using Combination of Properties

We have used the information about amino acid properties to discriminate two- and three-state proteins by maximum likelihood discriminant rules. We have tried several combinations of amino acid properties and the combination of 10 amino acid properties (K^0, H_t, pH_i, pK', B_1, P_β, ΔG_h, ΔH_h, ΔH, and ΔG) correctly discriminated 77 two- and three-state proteins with an accuracy of 89.6% with leave-one out cross-validation (jackknife) test. The discrimination accuracy is 88.2% for 51 two-state proteins and 92.3% for 26 three-state proteins.

Prediction on the Web

We have developed a web server for predicting the folding rates of two- and three-state proteins. Figure 2a shows the details of our web server including the input options. It takes the amino acid sequence in one letter format as the input and automatically omits gaps. It also gets the information about the structural class. The secondary structure and structural class information for a protein of known structure can be obtained either from SCOP[41] (http://scop.mrc-lmb.cam.ac.uk/scop/) or CATH[42] (http://cathwww.biochem.ucl.ac.uk/) databases and prediction results for structural class can be obtained with other servers, such as, protein structure prediction server (PSA; http://bmerc-www.bu.edu/psa/), secondary structural content prediction (SSCP; http://www.bork.embl-heidelberg.de/SSCP/), etc. The output formats are shown in Figure 2b. It shows the amino acid composition of the query protein, selected type of the protein, and the predicted folding rate. As an example, for λ repressor belonging to all-α protein, the predicted folding rate, $\ln(k_f)$, is 8.47 s^{-1}, which agrees remarkably well with experimental observations (8.50 s^{-1}). The prediction results are freely available at http://bioinformatics.myweb.hinet.net/foldrate.htm.

Comparison with Other Methods

The methods based on three-dimensional structures of proteins reveals the relationship between structural parameters, such as CO, LRO, etc., and protein folding rates. These methods showed a correlation in the range of 0.8 to 0.9. Gong et al.[20] reported the correlation of 0.91 using secondary structure content. Punta and Rost[23] predicted protein folding rates from amino acid sequence using the information about predicted long-range contacts and reported a correlation of 0.61 for a set of 37 proteins. Our earlier method using multiple regression technique showed a correlation of 0.93 between experimental and predicted folding rates. The present method with QRSM method raised the leave-one-out cross-validated correlation up to 0.90 and 0.98 between experimental and predicted folding rates, respectively, without and with structural class information. These accuracy levels are better than other methods in the literature. It is noteworthy that the direct comparison of correlation coefficients obtained in this work with the other methods is not appropriate due to the usage of different datasets and parameters in developing the methods. However, the empirical relationships derived for different structural classes predict the folding rates with high accuracy.

Conclusions

We have systematically analyzed the difference of amino acid compositions in slow and fast folding proteins. Utilizing this information, we have developed a method based on QRSM for predicting the folding rates of two- and three-state proteins. Our method showed a correlation of 0.90 between experimental and predicted folding rates. The correlation has been improved up to 0.98 when the proteins are classified into all-α, all-β, and mixed proteins. Further, the two- and three-state proteins have been correctly classified with an accuracy of 90% using discriminant rules. A web server has been developed for predicting the protein folding rates, which takes the amino acid sequence as input and displays the folding rate of the protein in the output.

References

1. Gromiha, M. M.; Selvaraj, S. Prog Biophys Mol Biol 2004, 86, 235.
2. Plaxco, K. W.; Simons, K. T.; Baker, D. J Mol Biol 1998, 277, 985.
3. Gromiha, M. M.; Selvaraj, S. J Mol Biol 2001, 310, 27.
4. Miller, E. J.; Fischer, K. F.; Marqusee, S. Proc Natl Acad Sci USA 2002, 99, 10359.
5. Zhou, H.; Zhou, Y. Biophys J 2002, 82, 458.
6. Debe, D. A.; Goddard, W. A., III. J Mol Biol 1999, 294, 619.
7. Munoz, V.; Eaton, W. A. Proc Natl Acad Sci USA 1999, 96, 11311.
8. Dinner, A. R.; Karplus, M. Nat Struct Biol 2001, 8, 21.
9. Makarov, D. E.; Keller, C. A.; Plaxco, K. W.; Metiu, H. Proc Natl Acad Sci USA 2002, 99, 3535.
10. Makarov, D. E.; Plaxco, K. W. Protein Sci 2003, 12, 17.
11. Dokholyan, N. V.; Li, L.; Ding, F.; Shakhnovich, E. I. Proc Natl Acad Sci USA 2002, 99, 8637.
12. Zhang, L.; Li, J.; Jiang, Z.; Xia, A. Polymer 2003, 44, 1751.
13. Gromiha, M. M. J Chem Inf Comput Sci 2003, 43, 1481.
14. Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. Proteins 2003, 51, 162–166.
15. Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. Protein Sci 2003, 12, 2057.
16. Shao, H.; Peng, Y.; Zeng, Z. H. Protein Pept Lett 2003, 10, 277.
17. Micheletti, C. Proteins 2003, 51, 74.
18. Nolting, B.; Schalike, W.; Hampel, P.; Grundig, F.; Gantert, S.; Sips, N.; Bandlow, W.; Qi, P. X. J Theor Biol 2003, 223, 299.
19. Zhang, L.; Sun, T. Biophys Chem 2005, 113, 9.
20. Gong, H.; Isom, D. G.; Srinivasan, R.; Rose, G. D. J Mol Biol 2003, 327, 1149.
21. Ivankov, D. N.; Finkelstein, A. V. Proc Natl Acad Sci USA 2004, 101, 8942.
22. Gromiha, M. M. J Chem Inf Model 2005, 45, 494.

23. Punta, M.; Rost, B. J Mol Biol 2005, 348, 507.
24. Ma, B. G.; Guo, J. X.; Zhang, H. Y. Proteins 2006, 65, 362.
25. Huang, J. T.; Tian, J. Proteins 2006, 63, 551.
26. Gromiha, M. M.; Thangakani, A. M.; Selvaraj, S. Nucleic Acids Res 2006, 34, W70.
27. Tomii, K.; Kanehisa, M. Protein Eng 1996, 9, 27.
28. Gromiha, M. M.; Oobatake, M.; Sarai, A. Biophys Chem 1999, 82, 51.
29. Gromiha, M. M.; Oobatake, M.; Kono, H.; Uedaira, H.; Sarai, A. J Biomol Struct Dyn 2000, 18, 281.
30. McDonald, D. B.; Grantham, W. J.; Tabor, W. L.; Murphy, M. J. Appl Math Model 2007, 31, 2095.
31. Dong, Q.; Tu, K.; Guo, L.; Li, H.; Zhao, Y. Food Microbiol 2007, 24, 624.
32. Kaul, N.; Agrawal, H.; Paradkar, A. R.; Mahadik, K. R. J Pharma Biomed Anal 2007, 43, 471.
33. Kaul, N.; Agrawal, H.; Paradkar, A. R.; Mahadik, K. R. J Biochem Biophys Methods 2005, 64, 121.
34. Iooss, B.; Van Dorpe, F.; Devictor, N. Reliab Eng Sys Saf 2006, 91, 1241.
35. Theodoridis, S.; Koutroumbas, K. Pattern Recognition; Elsevier/Academic Press: Amsterdam, Boston, 2006.
36. Sanders, D. H.; Eng, R. J.; Murph, A. F. Statistics: A Fresh Approach; McGraw-Hill: New York, 1985.
37. Gromiha, M. M. J Theor Biol 1993, 165, 87.
38. Gromiha, M. M.; Selvaraj, S. Biophys Chem 1999, 77, 49.
39. Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. Biochemistry 2000, 39, 11177.
40. Shimizu K, Muraoka Y, Hirose S, Tomii K, Noguchi T. BMC Bioinformatics 2007, 8, 78.
41. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Nucleic Acids Res 2004, 32, D226.
42. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA. Nucleic Acids Res 2007, 35, D291.