A Statistical Method for Predicting Protein Unfolding Rates from Amino Acid Sequence

M. Michael Gromiha,*,† S. Selvaraj,‡ and A. Mary Thangakani§

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, Department of Bioinformatics, Bharathidasan University, Tiruchirappalli 620 024,
Tamil Nadu, India, and Advanced Technology Institute Inc., Tokyo, Japan

Received September 21, 2005

The prediction of protein unfolding rates from amino acid sequences is one of the most important challenges in computational biology and chemistry. The analysis on the relationship between protein unfolding rates and physical−chemical, energetic, and conformational properties of amino acid residues provides valuable information to understand and predict the unfolding rates of two- and three-state proteins. We found that the classification of proteins into different structural classes shows an excellent correlation between amino acid properties and unfolding rates of two- and three-state proteins, indicating the importance of native-state topology in determining the protein unfolding rates. We have formulated three independent linear regression equations to different structural classes of proteins for predicting their unfolding rates from amino acid sequences and obtained an excellent agreement between predicted and experimentally observed unfolding rates of proteins; the correlation coefficients are 0.999, 0.990, and 0.992, respectively, for all-α, all-*β*, and mixed-class proteins. Further, we have derived a general equation applicable to all structural classes of proteins, which can be used for predicting the unfolding rates for proteins of an unknown structural class. We observed a correlation of 0.987 and 0.930, respectively, for back-check and jack-knife tests. These accuracy levels are better than those of other methods in the literature.

INTRODUCTION

Predicting the three-dimensional structure of a protein from its amino acid sequence is a challenging problem. A related interesting and important task is to understand the relationship between the sequences and folding rates of proteins.[1] Several investigations have been carried out to understand and predict the folding rates of proteins from protein three-dimensional structures. These studies include the concept of contact order;[2] first principles of protein folding;[3] long-range order;[4] elementary statistical models;[5] a combination of contact order and stability;[6] the number of native contacts;[7] the total contact distance;[8] topomer search models;[9] the topological properties of protein conformation;[10] neural networks based on contact order, long-range order, and total contact distance;[11] amino acid properties;[12] chain length;[13] size;[14] helix parameters;[15] and native-state geometry.[16] Recently, different methods have been proposed for predicting protein folding rates from amino acid sequence, secondary structure, and structural class information.[17−20]

The prediction of protein unfolding rates is an equally important problem as that of protein folding rates. The protein unfolding process plays an important role in controlling the function of proteins,[21] and there is a close relationship between protein unfolding rates and stability.[22] Recently, Jung et al.[23] investigated the topological parameters that determine protein unfolding rates and proposed the concept of edge removal for predicting the unfolding rates of 22 two-state proteins.

The folding and unfolding of a protein is mainly dictated by inter-residue interactions, which are influenced by physical, chemical, energetic, and conformational properties of amino acid residues. In our earlier work, we used amino acid properties to understand the transition-state structures of two-state proteins, predicting protein stability upon mutations and folding rates, and observed a good agreement with experimental results.[19,24−26] In this work, we have related various amino acid properties with the unfolding rates of two- and three-state proteins. We found that the classification of proteins into different structural classes shows a good correlation between amino acid properties and protein unfolding rates. We have set up an independent regression equation for each structural class and a general equation applicable to all structural classes using amino acid properties for predicting the unfolding rates of proteins. We found an excellent agreement between experimental and predicted protein unfolding rates, and the correlation coefficients are 0.987 and 0.930, with back-check prediction and jack-knife testing, respectively.

MATERIALS AND METHODS

Experimental Unfolding Rates. The experimental unfolding rates of 29 two- and three-state proteins used in related works[23,27,28] form the basis for the present study. The Protein Data Bank codes[29] and experimental $\ln(k_u)$ values are given in Table 1. The structural classification of these proteins yielded four all-α, 15 all-*β*, and 10 mixed-class proteins.

* Corresponding author. Tel: +81-3-3599-8046. Fax: +81-3-3599-8081. E-mail: michael-gromiha@aist.go.jp.

† National Institute of Advanced Industrial Science and Technology (AIST).

‡ Bharathidasan University.

§ Advanced Technology Institute Inc.

10.1021/ci050417u CCC: $33.50 © 2006 American Chemical Society
Published on Web 03/23/2006

Table 1. Predicted Unfolding Rates in a Set of 29 Two- and Three-State Proteins

PDB	ln(k_u) computed		ln(k_u) predicted		ln(k_u)
code	method 1[a]	method 2[b]	method 1[a]	method 2[b]	experimental
			All-α Proteins		
1lmb	3.43	3.48	3.60	3.72	3.40
2abd	−9.17	−7.63	−9.08	−6.45	−9.21
1hrc	−4.13	−3.94	−4.18	−3.44	−4.08
1imq	−4.44	−4.25	−4.70	−4.11	−4.42
			All-β Proteins		
2ait	−10.36	−10.10	−10.78	−10.34	−10.01
1csp	2.35	3.20	2.12	3.66	2.49
1c9o	−0.09	−0.19	0.39	8.73	−0.45
1g6p	−5.01	−4.44	−5.63	−4.99	−4.02
3mef	1.37	1.35	1.31	1.29	1.44
1mjc	1.37	1.35	1.51	1.46	1.19
1aey	−3.28	−3.90	−3.35	−4.75	−3.10
1shg	−4.96	−4.72	−4.93	−4.49	−5.01
1srl	−2.86	−2.91	−3.56	−3.58	−2.30
1pks	−6.94	−7.40	−6.33	−7.90	−7.31
1shf	−6.22	−6.21	−5.62	−5.07	−6.92
1fnf	−8.63	−7.25	−8.90	−4.21	−8.50
1tenf	−5.71	−6.47	−5.68	−7.11	−5.88
1tit	−6.77	−8.17	−6.34	−8.80	−7.62
1wit	−8.45	−8.35	−8.69	−8.48	−8.18
			Mixed-Class Proteins		
2ci2	−8.29	−7.58	−8.05	−6.17	−8.62
1pba	−0.88	−0.91	−1.44	−2.03	−0.43
1ubq	−8.44	−8.19	−8.63	−8.62	−7.74
2ptl	−3.54	−4.79	−2.95	−6.84	−3.91
1urn	−9.21	−9.37	−8.96	−8.29	−9.67
1hdn	−6.08	−6.47	−6.03	−6.97	−6.17
1fkb	−8.49	−9.83	−8.44	−10.65	−8.68
1aps	−9.47	−8.76	−10.40	−8.10	−9.12
2vik	−3.00	−2.71	−3.87	−2.52	−2.80
1arr	0.67	−0.08	1.32	−2.27	0.41

[a] Method 1: Computed using eqs 2−4, respectively, for all-α, all-β, and mixed-class proteins. [b] Method 2: Computed with the general equation applicable to all structural classes.

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,[30] in the present study. The amino acid properties were normalized between 0 and 1 using the expression $P_{\mathrm{norm}}(i) = [P(i) - P_{\min}]/\ [P_{\max} - P_{\min}]$, where $P(i)$ and $P_{\mathrm{norm}}(i)$ 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 of the 49 properties used in this study along with their brief descriptions have been explained in our earlier articles[31,32] and are available on the Web at http://www.cbrc.jp/~gromiha/fold_rate/property.html. These properties were obtained either directly from experiments or by computational methods using three-dimensional structures of proteins.

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

$$P_{\mathrm{ave}}(i) = \sum_{j=1}^{N} P(j)/N \qquad (1)$$

where $P(j)$ is the property value of the jth residue and the summation is over N, the total number of residues in a protein. The computed property value $P_{\mathrm{ave}}(i)$ for each class of proteins was related to the experimental unfolding rate ln $k_u(i)$ using a single correlation coefficient. Further, we have combined the amino acid properties using a multiple regression technique.[33] We have used the programs developed in our lab for most of the calculations including the derivation of multiple regression equations. The statistical significance of the results obtained in the present study has been verified with a t test and a p value by standard procedures using the tools available at http://fonsg3.let.uva.nl/Service/Statistics.html.

The average deviation (error) is computed by dividing the total absolute difference between the experimental and predicted folding rates by the total number of proteins. The standard deviation is computed using the standard expression $\sigma = [\Sigma(X - \bar{X})^2/(n - 1)]^{1/2}$.

Back-Check Prediction and the Jack-Knife (Leave-One-Out Cross-Validation) Test. We have used the data set of all of the 29 proteins for relating protein unfolding rates with amino acid properties using a multiple regression technique. The same regression equation has been used to predict the unfolding rates of these 29 proteins. This method is called back-check prediction (or self-consistency testing).

Further, we performed a jack-knife test for assessing the validity of the results. In this method, we have derived the multiple regression equation by omitting one protein and used this equation for predicting the unfolding rate of the left-out protein. From the comprehensive analysis on the validation methods for predicting protein structural classes, it has been reported that the jack-knife test is more rigorous than k-fold cross-validation methods.[34−36]

RESULTS AND DISCUSSIONS

Role of Protein Structural Classes. In our earlier work, we analyzed the influence of inter-residue interactions in different structural classes, and we found that the interacting pattern is distinct in each structural class.[37,38] Further, the analysis on the predictive accuracy of several secondary structure prediction algorithms indicates the necessity of structural classification for better performance.[39,40] On the other hand, several methods are available to predict the protein structural class with a high degree of accuracy.[41−43]

The relationship between amino acid properties and protein unfolding rates of all 29 proteins shows that the correlation is weak and the classification of proteins into all-α, all-β, and a mixed class remarkably enhanced the correlation. This result suggests that structural classification is necessary for the successful prediction of protein unfolding rates. The classification based on protein structures includes the information about the topology of the protein, which is found to be an important determinant for protein folding and unfolding rates.[23,44] Further, the regression equations developed for each structural class perform better than the general equation. The performance of the method is assessed with correlation, deviation, t-test and p values, and back-check and jack-knife tests. The additional classification, such as α+β, α/β, irregular, and so forth, may improve the accuracy as seen in protein structural class predictions.[42,45] However, because of the limited data set, we have used the classification method with three groups.

All-α Proteins. The relationship between amino acid properties and protein unfolding rates of all-α proteins shows that the property N_s (average number of surrounding residues)

Table 2. Cross-Correlation Coefficients between the Protein Unfolding Rates Obtained with Amino Acid Properties Used in eqs 2−4 (a) and eq 5 (b) and That with Protein Unfolding Rates

Part a

all-α class

	N_s	K^0	$\ln(k_u)$
N_s	1.00	0.80	0.97
K^0	0.80	1.00	0.93

all-β class

	P	P_c	ΔG_h	ΔH_h	ΔC_{ph}	f	$\ln(k_u)$
P	1.00	−0.70	−0.01	−0.20	0.57	0.91	−0.02
P_c	−0.70	1.00	−0.32	−0.10	−0.70	−0.73	−0.12
ΔG_h	−0.01	−0.32	1.00	0.95	0.16	−0.14	0.29
ΔH_h	−0.20	−0.10	0.95	1.00	−0.14	−0.35	0.34
ΔC_{ph}	0.57	−0.70	0.16	−0.14	1.00	0.62	−0.14
f	0.91	−0.73	−0.14	−0.35	0.62	1.00	0.11

mixed class

	P	R_f	ΔASA	$-T\Delta S_h$	$\ln(k_u)$
P	1.00	−0.56	0.08	0.14	0.34
R_f	−0.56	1.00	0.67	0.51	−0.26
ΔASA	0.08	0.67	1.00	0.96	0.05
$-T\Delta S_h$	0.14	0.51	0.96	1.00	−0.08

Part b

	pH_i	E_{sm}	E_l	E_t	P_α	P_β	P_c	C_α	α_m	ASA$_D$	ASA$_N$	ΔASA	ΔG_h	ΔG_c	ΔG	$P_{\phi-\Phi}$	$\ln(k_u)$
pH_i	1.00	−0.54	−0.04	−0.40	0.40	0.06	−0.46	0.75	0.67	0.71	0.62	0.54	0.11	−0.15	−0.14	0.01	−0.06
E_{sm}	−0.54	1.00	0.10	0.75	−0.29	0.02	0.30	−0.89	−0.34	−0.87	−0.76	−0.65	0.49	−0.45	0.09	−0.11	0.08
E_l	−0.04	0.10	1.00	0.74	−0.40	0.86	0.06	−0.02	−0.20	−0.09	−0.64	0.43	0.24	−0.22	0.04	−0.18	−0.07
E_t	−0.40	0.75	0.74	1.00	−0.46	0.58	0.24	−0.63	−0.37	−0.65	−0.95	−0.16	0.49	−0.44	0.10	−0.19	0.00
P_α	0.40	−0.29	−0.40	−0.46	1.00	−0.36	−0.92	0.54	0.80	0.58	0.58	0.37	0.11	−0.12	0.01	−0.21	0.16
P_β	0.06	0.02	0.86	0.58	−0.36	1.00	−0.04	0.09	−0.18	0.05	−0.47	0.48	0.29	−0.33	−0.17	−0.33	−0.17
P_c	−0.46	0.30	0.06	0.24	−0.92	−0.04	1.00	−0.61	−0.78	−0.63	−0.42	−0.60	−0.25	0.27	0.05	0.36	−0.10
C_α	0.75	−0.89	−0.02	−0.63	0.54	0.09	−0.61	1.00	0.61	0.98	0.76	0.84	−0.22	0.17	−0.11	−0.06	−0.05
α_m	0.67	−0.34	−0.20	−0.37	0.80	−0.18	−0.78	0.61	1.00	0.60	0.54	0.44	0.06	−0.04	0.13	0.03	0.19
ASA$_D$	0.71	−0.87	−0.09	−0.65	0.58	0.05	−0.63	0.98	0.60	1.00	0.79	0.84	−0.23	0.14	−0.25	−0.17	−0.12
ASA$_N$	0.62	−0.76	−0.64	−0.95	0.58	−0.47	−0.42	0.76	0.54	0.79	1.00	0.32	−0.35	0.27	−0.21	0.03	−0.07
ΔASA	0.54	−0.65	0.43	−0.16	0.37	0.48	−0.60	0.84	0.44	0.84	0.32	1.00	−0.03	−0.03	−0.18	−0.28	−0.11
ΔG_h	0.11	0.49	0.24	0.49	0.11	0.29	−0.25	−0.22	0.06	−0.23	−0.35	−0.03	1.00	−0.96	0.06	−0.15	0.08
ΔG_c	−0.15	−0.45	−0.22	−0.44	−0.12	−0.33	0.27	0.17	−0.04	0.14	0.27	−0.03	−0.96	1.00	0.22	0.32	0.03
ΔG	−0.14	0.09	0.04	0.10	0.01	−0.17	0.05	−0.11	0.13	−0.25	−0.21	−0.18	0.06	0.22	1.00	0.61	0.39
$P_{\phi-\psi}$	0.01	−0.11	−0.18	−0.19	−0.21	−0.33	0.36	−0.06	0.03	−0.17	0.03	−0.28	−0.15	0.32	0.61	1.00	0.57

has the highest correlation with protein unfolding rates ($r = 0.97$). Further, the properties related with long-range contacts show significant correlation with $\ln(k_u)$ values. This observation reveals that the presence of tightly packed residues slows down the unfolding process.

We have developed a simple regression model for predicting the unfolding rates of all-α proteins using amino acid properties, and the regression equation is

$$\ln(k_u) = 109.40\ (\pm 0.12)\cdot N_s + 114.84\ (\pm 0.11)\cdot K^0 - 87.44\ (\pm 0.04) \quad (2)$$

where N_s is the number of surrounding residues and K^0 is the compressibility. The standard deviation for the coefficients is less than 1% for each property, which might be due to a lesser number of data. We have computed the protein unfolding rates of all-α proteins using eq 2 and observed an excellent agreement between the predicted unfolding rates and experimental observations. The correlation coefficient is 0.999, and the average deviation (error) is 0.04. Further, we have verified that the results are statistically significant ($t = 173.8$ and $p \leq 3.31 \times 10^{-5}$). The cross correlation between the unfolding rates obtained with N_s and K^0 and the correlation of these properties with protein unfolding rates are presented in Table 2a. We observed a correlation of 0.80 between the unfolding rates obtained with these properties. Further, they have a strong correlation with protein unfolding rates.

We have also performed a jack-knife test by determining the coefficients of the regression equation using $(n - 1)$ data (i.e., omitting one protein at a time) and then computing the unfolding rate of the omitted protein. We found that all of the considered proteins agreed extremely well with experimental results. The r value is 0.999 ($t = 35.31$; $p \leq 8.03 \times 10^{-4}$), and the average error is 0.18.

In general, protein compressibility provides vital information about the forces that govern the stability of folded proteins. An interesting correlation has been observed between the compressibility of a protein and the heat capacity change upon unfolding, ΔC_p.[46] It is striking to note that the compressibility properties of amino acid residues are codeterminants of the protein unfolding rate along with the number of surrounding residues in the case of α-helical proteins.

All-β Proteins. In all-β proteins, the backbone dihedral probability property shows significant correlation with protein unfolding rates ($r = 0.72$). Similar calculations with random numbers yielded an average correlation of 0.20 ± 0.16. Further, we have combined different amino acid properties with a multiple regression fit for predicting protein unfolding rates. The corresponding regression equation is

$$\ln(k_u) = -142.67\ (\pm1.63)\cdot P + 58.27\ (\pm1.22)\cdot P_c - 2700.43\ (\pm0.74)\cdot\Delta G_h + 2991.55\ (\pm0.72)\cdot\Delta H_h + 1034.81\ (\pm1.18)\cdot\Delta C_{ph} + 370.33\ (\pm1.31)\cdot f - 764.44\ (\pm0.47) \quad (3)$$

where P, P_c, ΔG_h, ΔH_h, ΔC_{ph}, and f are, respectively, the polarity, coil tendency, Gibbs free-energy change of hydration for unfolding, enthalpy change of hydration for unfolding, heat capacity change of unfolding, and flexibility. This result shows the major role played by thermodynamic parameters along with physical−chemical and conformational parameters for determining protein unfolding rates.

In eq 3, we noticed that the standard deviation of each coefficient is less than 3%. The unfolding rates obtained with the amino acid properties used in eq 3 showed a correlation in the range of 0.01−0.95 among them. The combination of properties yielded a good correlation with the $\ln(k_u)$ values.

The predicted and experimental $\ln(k_u)$ values of 15 all-β proteins using eq 3 are given in Table 1. We found that the predicted protein unfolding rates have a very good agreement with experimental observations, and the correlation between them is 0.993. The calculated p value is $\leq 8.14 \times 10^{-57}$, and $t = 30.47$. We have also carried out the jack-knife test, and the correlation obtained with this method is 0.990 ($p \leq 1.80 \times 10^{-6}$ and $t = 16.39$). The standard error of the estimate is 0.68.

Mixed-Class Proteins. The relationship between amino acid properties and protein unfolding rates in a set of 10 mixed-class proteins shows that the backbone dihedral probability property has the highest positive correlation with the protein unfolding rates observed in the all-β class of proteins. On the other hand, a set of random numbers yielded an average correlation of 0.27 ± 0.20.

We have set up the following regression equation for predicting protein unfolding rates.

$$\ln(k_u) = -67.07\ (\pm1.31)\cdot P - 253.53\ (\pm0.95)\cdot R_f + 701.48\ (\pm0.95)\cdot\Delta\mathrm{ASA} - 506.66\ (\pm0.91)\cdot -T\Delta S_h + 49.41\ (\pm0.40) \quad (4)$$

where R_f, ΔASA, and $-T\Delta S_h$ are, respectively, the refractive index, solvent-accessible surface area upon unfolding, and unfolding entropy change of hydration.

As seen in the all-β proteins, there is a wide range of correlations among the unfolding rates obtained with the properties used in eq 4, and the correlation of each property with the unfolding rates of proteins is poor (Table 2a). The predicted and experimental $\ln(k_u)$ values for the mixed-class proteins used in this work derived by eq 4 are presented in Table 1. We found an excellent correlation of 0.994 between experimental and predicted protein unfolding rates. Further, we have corroborated the statistical significance of the results ($p \leq 1.27 \times 10^{-4}$ and $t = 26.45$). We have also performed a jack-knife test and obtained an r value of 0.992 ($p \leq 2.4 \times 10^{-4}$; $t = 11.8$). Further, we have estimated the deviation of $\ln(k_u)$ values, and the average deviation obtained for back-check prediction and jack-knife testing are, respectively, 0.34 and 0.78.

Prediction of Protein Unfolding Rates. We have used three different equations (eqs 2−4) for predicting the unfolding rates of proteins belonging to each structural class.

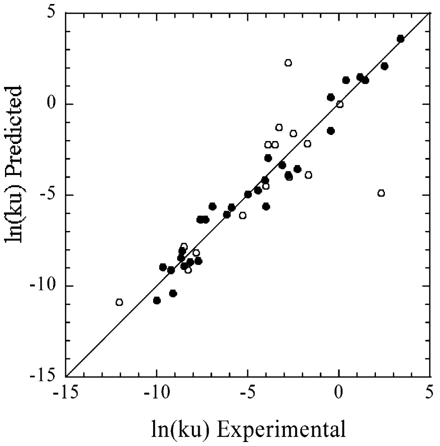

Figure 1. Relationship between experimental and predicted $\ln(k_u)$ values with a jack-knife test using a multiple regression model in 29 two- and three-state proteins (●). The results obtained for the set of 16 proteins are also included in this figure(○).

The results obtained with back-check prediction are presented in Table 1. We found an excellent correlation of 0.995 between predicted and experimental protein unfolding rates for the sample set of 29 proteins. The t-test and p values are, respectively, 51.46 and $\leq 5.74 \times 10^{-13}$, and the average error is 0.31. We have also performed a jack-knife test to examine the validity of the present method, and the results are shown in Figure 1. We found that about 76% of the considered proteins (22 out of 29 proteins) agreed very well with the experimental results, and the deviation is less than one unit. The correlation coefficient between experimental and predicted $\ln(k_u)$ values is 0.992 ($t = 25.85$; $p \leq 2.52 \times 10^{-12}$), and the average error is 0.65.

Prediction with the General Equation. The unfolding rates of proteins with a known structural class can be predicted with a high degree of accuracy using eqs 2−4. However, these equations are not applicable to proteins of an unknown structural class. Hence, we derived two regression equations by combining all 29 proteins together, which can be used for predicting the protein unfolding rates of unknown structural classes, using (i) the combination of amino acid properties that influence the unfolding rates of all-α, all-β, and mixed-class proteins and (ii) the addition of properties with the highest correlation coefficients. Both of the equations showed a good correlation with protein unfolding rates, and we used the following equation for predicting the unfolding rate of a protein:

$$\ln(k_u) = -278.0\ (\pm1.5)\cdot \mathrm{p}H_i + 898.4\ (\pm1.1)\cdot E_{sm} + 1247.0\ (\pm1.6)\cdot E_l - 272.1\ (\pm1.2)\cdot E_t + 6837.9\ (\pm1.3)\cdot P_\alpha + 4992.6\ (\pm1.4)\cdot P_\beta + 7870.6\ (\pm1.7)\cdot P_c + 143.6\ (\pm1.6)\cdot C_\alpha - 278.0\ (\pm1.4)\cdot \alpha_m - 3739.7\ (\pm1.1)\cdot \mathrm{ASA_D} + 3594.4\ (\pm1.5)\cdot \mathrm{ASA}_N + 2995.0\ (\pm1.6)\cdot\Delta\mathrm{ASA} - 3297.0\ (\pm1.0)\cdot\Delta G_h - 3548.9\ (\pm1.9)\cdot\Delta G_c + 1609.4\ (\pm2.5)\cdot\Delta G + 308.9\ (\pm2.8)\cdot P_{\phi-\psi} - 7274.1\ (\pm0.6) \quad (5)$$

Interestingly, the equation is a combination of physical−chemical, energetic, and conformational parameters of amino acid residues. The standard deviation for the coefficients of each property is less than 2%. The cross-correlation coefficients between the unfolding rates obtained with amino acid properties that are used in eq 5 are presented in Table 2b along with the correlation between each amino acid property and the protein unfolding rates. We observed that several properties have significantly high positive and negative correlations in this table. The combination of properties raised the correlation, and the predicted unfolding rates of proteins without structural class information are included in Table 1. We found that the correlation between experimental and predicted ln(k_u) values is 0.987, and the deviation is 0.49. The combination of 11 properties observed in eqs 2−4 along with pH_i, P_β, B_r, α_n, G_{hD} and G_{hN} showed a correlation of 0.97. The *t*-test and *p* values are, respectively, 32.25 and 1.27×10^{-12}. We have also examined the method with a jack-knife test and obtained a good agreement with experimental results (Table 1). The correlation, average error, *t*-test and *p* values are, respectively, 0.930, 1.55, 8.633, and 2.96×10^{-8}. Although the performance of the general equation is not better than that of the equations derived for each structural class, it can be used for a protein of an unknown structural class.

A key observation in relating the unfolding rates of proteins in terms of the physical−chemical properties of their constituent amino acid residues is that the rates are dependent on different thermodynamic quantities such as the Gibbs free energy change of hydration for unfolding, the enthalpy change of hydration for unfolding, the heat capacity change of unfolding, and the unfolding entropy change of hydration. All of these properties are related to each other as well as to other properties. For example, denaturation heat capacity change ΔC_p has been shown to be linearly dependent on the change in solvent accessible surface area, ΔASA.[47−49] We have also mentioned the correlation between compressibility and heat capacity change. In essence, the unfolding rates of proteins seem to be determined not by a single physical−chemical property but rather by the concerted effects of different physical−chemical properties of amino acid residues, the dominant being the thermodynamic properties. This observation is similar to that in which the component-coupled algorithm predicts the structural classes of proteins significantly better than component-independent algorithms.[50−53]

It has been reported that the multivariate regression models should have at least five data points for one adjustable parameter for predicting the target value of the sample data.[54] In the present work, we have used approximately 2−3 data points for each descriptor (property), and we noticed that the unfolding rates of proteins have been predicted with a high degree of accuracy using a greater number of adjustable parameters.

Validating the Present Method. Recently, Maxwell et al.[55] reported the unfolding rates of 30 two-state proteins, and 16 of them are not used in our predicted model. We have calculated the unfolding rates of these 16 proteins belonging to different structural classes and compared the predicted ln(k_u) values with experimental observations. For apo-azurin (PDB code: 1E65), we obtained a ln(k_u) value of −4.63/s, which shows an excellent agreement with the experimental unfolding rate, −4.02/s.[55] The amino terminal domain of ribosomal protein L9 gave a ln(k_u) value of 0.03/s, and the experimental value is 0.08/s. CheW (PDB code: 1K0S) has been classified as an all-β protein in the Structural Classification of Proteins database, and we obtained a ln(k_u) value of −10.9/s using structural class information. This is very close to the experimental unfolding rate, −12.05/s. Further, the relationships between the predicted and experimental unfolding rates for all 16 proteins are presented in Table 3. We observed a good agreement for most of the proteins. The predicted unfolding rates of 10 proteins are close to the experimental values using structural class information, and seven proteins showed good agreement without structural class information. Further, we noticed that the introduction of additional residues (tags) at the C-terminal end of the proteins led to wrong predictions, and it might be due to the fact that the equations are derived with the sequences of known structures. In addition, protein unfolding rates depend on experimental conditions, which are not considered in the present work.

Table 3. Prediction of Unfolding Rates in a Test Set of 16 Proteins

			ln(k_u) predicted		
protein name	PDB code	ln(k_u) experimental	method 1[a]	eq	method 2[b]
ABP1 SH3	1JO8	−2.72	−3.99	3	−1.65
apo-azurin	1E65	−4.02	−4.47	4	−4.63
CheW	1K0S	−12.05	−10.90	3	−5.20
GW1	1M9S	−1.66	−9.87	4	−3.86
L23	1N88	−3.88	−11.30	3	−2.22
NTL9	1DIV	0.08	3.05	3	0.03
protein G	3GB1	−1.72	−2.17	4	1.57
S6	1RIS	−8.28	−9.13	3	2.73
Sho1 SH3	SSU81_YEAST	−2.49	−1.62	4	−9.93
Src SH2	1SPR	−3.48	−2.21	3	−7.76
Tm1083	Q9X0H1_THEMA	−5.26	−6.11	2	0.70
Urm1	Q59JW3_CANAL	−3.30	−0.69	4	−1.28
VlsE	1L8W	−8.47	−4.50	4	−7.85
rafRBD	1RFA	−2.77	−6.61	3	2.29
CTL9	1DIV	−7.85	−8.16	3	−1.91
1Im7	1AYI	2.34	−4.89	2	−7.46

[a] Method 1: Computed using eqs 2−4, respectively, for all-α, all-β, and mixed-class proteins. [b] Method 2: Computed with the general equation applicable to all structural classes.

CONCLUSIONS

We have systematically analyzed the relationship between amino acid properties and protein unfolding rates in different structural classes of proteins. We have set up linear regression models for predicting the unfolding rates of two- and three-state proteins of known and unknown structural classes using a combination of amino acid properties. The present method is the first one that can predict the protein unfolding rates from amino acid sequences. The predicted unfolding rates show an excellent agreement with experimental observations; the correlation coefficients are 0.999, 0.990, and 0.992, respectively, for all-α, all-β, and mixed-class proteins. These accuracy levels are superior to those of other methods in the literature.

ACKNOWLEDGMENT

We thank the reviewers for constructive comments and Dr. Yutaka Akiyama for encouragement.

Supporting Information Available: The amino acid sequences of all the proteins used in the present work, experimental unfolding rates, and their 49 average property values. This material is available free of charge via the

Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Eaton, W. A.; Munoz, V.; Hagen, S. J.; Jas, G. S.; Lapidus, L. J.; Henry, E. R.; Hofrichter, J. Fast Kinetics and Mechanisms in Protein Folding. *Annu. Rev. Biophys. Biomol. Struct.* **2000**, *29*, 327−359.

(2) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact Order, Transition State Placement and the Refolding Rates of Single Domain Proteins. *J. Mol. Biol*. **1998**, *277*, 985−994.

(3) Debe, D. A.; Goddard, W. A., III. First Principles Prediction of Protein Folding Rates. *J. Mol. Biol.* **1999**, *294*, 619−625.

(4) Gromiha, M. M.; 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.

(5) Munoz, V.; Eaton, W. A. A Simple Model for Calculating the Kinetics of Protein Folding from Three-Dimensional Structures. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 11311−11316.

(6) Dinner, A. R.; Karplus, M. The Roles of Stability and Contact Order in Determining Protein Folding Rates. *Nat. Struct. Biol*. **2001**, *8*, 21−22.

(7) Makarov, D. E.; Keller, C. A.; Plaxco, K. W.; Metiu. H. How the Folding Rate Constant of Simple, Single-Domain Proteins Depends on the Number of Native Contacts. *Proc. Natl. Acad. Sci. U.S.A*. **2002**, *99*, 3535−3539.

(8) Zhou, H.; Zhou, Y. Folding Rate Prediction Using Total Contact Distance. *Biophys. J.* **2002**, *82*, 458−463.

(9) Makarov, D. E.; Plaxco, K. W. The Topomer Search Model: A Simple, Quantitative Theory of Two-State Protein Folding Kinetics. *Protein Sci*. **2003**, *12*, 17−26.

(10) Dokholyan, N. V.; Li, L.; Ding, F.; Shakhnovich, E. I. Topological Determinants of Protein Folding. *Proc. Natl. Acad. Sci. U.S.A*. **2002,** *99*, 8637−8641.

(11) Zhang, L.; Li, J.; Jiang, Z.; Zia, A. Folding Rate Prediction on Neural Network Model. *Polymer* **2003**, *44*, 1751−1756.

(12) Gromiha, M. M. Importance of Native State Topology for Determining the Folding Rate of Two-State Proteins. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1481−1485.

(13) Galzitskaya, O. V.; Garbuzynskiy, S. O.; Ivankov, D. N.; Finkelstein, A. V. Chain Length is the Main Determinant of the Folding Rate for Proteins with Three-State Folding Kinetics. *Proteins* **2003**, *51*, 162−166.

(14) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. Contact Order Revisited: Influence of Protein Size on the Folding Rate. *Protein Sci.* **2003**, *12*, 2057−2062.

(15) Shao, H.; Peng, Y.; Zeng, Z. H. A Simple Parameter Relating Sequences with Folding Rates of Small Alpha Helical Proteins. *Protein Pept. Lett*. **2003**, *10*, 277−280.

(16) Micheletti, C. Prediction of Folding Rates and Transition-State Placement from Native-State Geometry. *Proteins* **2003**, *51*, 74−84.

(17) Gong, H.; Isom, D. G.; Srinivasan, R.; Rose, G. D. Local Secondary Structure Content Predicts Folding Rates for Simple, Two-State Proteins. *J. Mol. Biol*. **2003**, *327*, 1149−1154.

(18) Ivankov, D. N.; Finkelstein, A. V. Prediction of Protein Folding Rates from the Amino Acid Sequence-Predicted Secondary Structure. *Proc. Natl. Acad. Sci. U.S.A*. **2004**, *101*, 8942−8944.

(19) Gromiha, M. M. A Statistical Model for Predicting Protein Folding Rates from Amino Acid Sequence with Structural Class Information. *J. Chem. Inf. Model.* **2005**, *45*, 494−501.

(20) Punta, M.; Rost, B. Protein Folding Rates Estimated from Contact Predictions. *J. Mol. Biol.* **2005**, *348*, 507−512.

(21) Pain, R. H. *Mechanisms of Protein Folding*; Oxford University Press: New York, 2000.

(22) Creighton, T. E. *Protein Folding*; W. H. Freeman: New York, 1992.

(23) Jung, J.; Lee, J.; Moon, H. T. Topological Determinants of Protein Unfolding Rates. *Proteins* **2005**, *58*, 389−395.

(24) Gromiha, M. M.; Oobatake, M.; Kono, H.; Uedaira, H.; Sarai, A. Role of Structural and Sequence Information in the Prediction of Protein Stability Changes: Comparison between Buried and Partially Buried Mutations. *Protein Eng.* **1999**, *12*, 549−555.

(25) Gromiha M. M.; Oobatake, M.; Kono, H.; Uedaira, H.; Sarai, A. Importance of Mutant Position in Ramachandran Plot for Predicting Protein Stability of Surface Mutations. *Biopolymers* **2002**, *64*, 210−220.

(26) Gromiha, M. M.; Selvaraj, S. Important Amino Acid Properties for Determining the Transition State Structures of Two-State Protein Mutants. *FEBS Lett.* **2002**, *526*, 129−134.

(27) Jackson, S. E. How Do Small Single-Domain Proteins Fold? *Fold Des*. **1998**, *3*, R81−91.

(28) Fulton, K. F.; Devlin, G. L.; Jodun, R, A; Silvestri, L.; Bottomley, S. P.; Fersht, A. R.; Buckle, A. M. PFD: A Database for the Investigation of Protein Folding Kinetics and Stability. *Nucleic Acids Res.* **2005**, *33*, D279-D283.

(29) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res*. **2000**, *28*, 235−242.

(30) 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.

(31) Gromiha, M. M.; Oobatake, M.; Sarai, A. Important Amino Acid Properties for Enhanced Thermostability from Mesophilic to Thermophilic Proteins. *Biophys. Chem.* **1999**, *82*, 51−67.

(32) Gromiha, M. M.; Oobatake, M.; Kono, H.; Uedaira H.; Sarai, A. Importance of Surrounding Residues for Protein Stability of Partially Buried Mutations. *J. Biomol. Struct. Dyn*. **2000**, *18*, 281−295.

(33) Grewal, P. S. *Numerical Methods of Statistical Analysis*; Sterling Publishers: New Delhi, India, 1987.

(34) Chou, K. C.; Zhang, C. T. Prediction of Protein Structural Classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275−349.

(35) Zhou, G. P.; Assa-Munt, N. Some Insights into Protein Structural Class Prediction. *Proteins* **2001**, *44*, 57−59.

(36) Zhou, G. P.; Doctor, K. Subcellular Location Prediction of Apoptosis Proteins. *Proteins* **2003**, *50*, 44−48.

(37) Gromiha, M. M.; Selvaraj, S. Importance of Long-Range Interactions in Protein Folding. *Biophys. Chem*. **1999**, *77*, 49−68.

(38) Gromiha, M. M.; Selvaraj, S. Inter-Residue Interactions in Protein Folding and Stability. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235−277.

(39) Rost, B.; Sander, C. Secondary Structure Prediction of All-Helical Proteins in Two States. *Protein Eng*. **1993**, *6*, 831−836.

(40) Gromiha, M. M.; Selvaraj, S. Protein Secondary Structure Prediction in Different Structural Classes. *Protein Eng*. **1998**, *11*, 249−251.

(41) Chou, K. C. Prediction of Protein Structural Classes and Subcellular Locations. *Curr. Protein Pept. Sci.* **2000**, *1*, 171−208.

(42) Chou, K. C. Progress in Protein Structural Class Prediction and Its Impact to Bioinformatics and Proteomics. *Curr. Protein Pept. Sci*. **2005**, *6*, 423−436.

(43) Chou, K. C.; Cai, Y. D. Predicting Protein Structural Class by Functional Domain Composition. *Biochem. Biophys. Res. Commun.* **2004**, *321*, 1007−1009.

(44) Plaxco, K. W.; Simons, K. T.; Ruczinski, I.; Baker, D. Topology, Stability, Sequence, and Length: Defining the Determinants of Two-State Protein Folding Kinetics. *Biochemistry* **2000**, *39*, 11177−11183.

(45) Chou, K. C.; Maggiora, G. M. Domain Structural Class Prediction. *Protein Eng.* **1998**, *11*, 523−538.

(46) Dadarlat, V. M.; Post, C. B. Adhesive−Cohesive Model for Protein Compressibility: An Alternative Perspective on Stability. *Proc. Natl. Acad. Sci. U.S.A.* **2003**, *100*, 14778−14783.

(47) Livingstone, J. R.; Spolar, R. S.; Record, M. T., Jr. Contribution to the Thermodynamics of Protein Folding from the Reduction in Water-Accessible Nonpolar Surface Area. *Biochemistry* **1991**, *30*, 4237−4244.

(48) Spolar, R. S.; Livingstone, J. R.; Record, M. T., Jr. Use of Liquid Hydrocarbon and Amide Transfer Data To Estimate Contributions to Thermodynamic Functions of Protein Folding from the Removal of Nonpolar and Polar Surface from Water. *Biochemistry* **1992**, *31*, 3947−3955.

(49) Myers, J. K.; Pace, C. N.; Scholtz, J. M. Denaturant m Values and Heat Capacity Changes: Relation to Changes in Accessible Surface Areas of Protein Unfolding. *Protein Sci.* **1995**, *4*, 2138−2148.

(50) Chou, K. C.; Zhang, C. T. Predicting Protein Folding Types by Distance Functions that Make Allowances for Amino Acid Interactions. *J. Biol. Chem.* **1994**, *269*, 22014−22020.

(51) Chou, K. C. A Novel Approach to Predicting Protein Structural Classes in a (20−1)-D Amino Acid Composition Space. *Proteins* **1995**, *21*, 319−344.

(52) Chou, K. C.; Liu, W.; Maggiora, G. M.; Zhang, C. T. Prediction and Classification of Domain Structural Classes. *Proteins* **1998**, 31, 97−103.

(53) Zhou, G. P. An Intriguing Controversy over Protein Structural Class Prediction. *J. Protein Chem.* **1998**, *17*, 729−738.

(54) Andrea, T. A.; Kalayeh, H. Applications of Neural Networks in Quantitative Structure−Activity Relationships of Dihydrofolate Reductase Inhibitors. *J. Med. Chem.* **1991**, *34*, 2824−2836.

(55) Maxwell, K. L.; Wildes, D.; Zarrine-Afsar, A.; De Los Rios, M. A.; Brown, A. G.; Friel, C. T.; Hedberg, L.; Horng, J. C.; Bona, D.; Miller, E. J.; Vallee-Belisle, A.; Main, E. R.; Bemporad, F.; Qiu, L.; Teilum, K.; Vu, N. D.; Edwards, A. M.; Ruczinski, I.; Poulsen, F. M.; Kragelund, B. B.; Michnick, S. W.; Chiti, F.; Bai, Y.; Hagen, S. J.; Serrano, L.; Oliveberg, M.; Raleigh, D. P.; Wittung-Stafshede, P.; Radford, S. E.; Jackson, S. E.; Sosnick, T. R.; Marqusee, S.; Davidson, A. R.; Plaxco, K. W. Protein Folding: Defining a “Standard” Set of Experimental Conditions and a Preliminary Kinetic Data Set of Two-State Proteins. *Protein Sci.* **2005**, *14*, 602−616.

CI050417U