A Statistical Model for Predicting Protein Folding Rates from Amino Acid Sequence with Structural Class Information

M. Michael Gromiha*

Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and
Technology (AIST), Aomi Frontier Building 17F, 2-43 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received August 6, 2004

Prediction of protein folding rates from amino acid sequences is one of the most important challenges in molecular biology. In this work, I have related the protein folding rates with physical-chemical, energetic and conformational properties of amino acid residues. I found that the classification of proteins into different structural classes shows an excellent correlation between amino acid properties and folding rates of two- and three-state proteins, indicating the importance of native state topology in determining the protein folding rates. I have formulated a simple linear regression model for predicting the protein folding rates from amino acid sequences along with structural class information and obtained an excellent agreement between predicted and experimentally observed folding rates of proteins; the correlation coefficients are 0.99, 0.96 and 0.95, respectively, for all-α, all-*β* and mixed class proteins. This is the first available method, which is capable of predicting the protein folding rates just from the amino acid sequence with the aid of generic amino acid properties and structural class information.

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 sequences and folding rates of proteins.[1] As an advance to this problem, 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 protein folding rates. Gromiha and Selvaraj[3] emphasized the importance of long-range contacts for determining the folding rates of two-state proteins and 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. Accordingly, theoretical models have been proposed for predicting the folding rates of two-state proteins based on long-range contacts.[3−5] Recently, thermodynamic and kinetic experiments demonstrated that long-range order is one of the best parameters that correlates with protein-refolding rates including circular permutations of ribosomal proteins S6 from *Thermus thermophilus*.[6]

Debe and Goddard[7] have predicted the folding rates for 21 small, single-domain, topologically distinct proteins based on the first principles of protein folding and observed a good correlation with experimentally observed folding rates. Munoz and Eaton[8] have proposed an elementary statistical mechanical model to calculate the protein folding rates from their three-dimensional structures. Dinner and Karplus[9] performed a statistical analysis to predict the protein folding rates and reported that both contact order and stability play important roles in determining the folding rate. Further, neural networks based models have been developed to relate folding rates of proteins from the combination of topological parameters, contact order, long-range order and total contact distance.[10] Recently, the role of chain length, native state geometry and the topological properties of protein conformation for determining the protein folding rates have been reported.[11−13]

Currently, all the available methods for predicting the folding rates of two- and three-state proteins are based on their three-dimensional structures, and no method has been proposed for predicting the folding rates from amino acid sequence, so far. Hence, it is necessary to develop a model for predicting the protein folding rates from their amino acid sequences, which could serve as a useful tool for predicting the folding rates of proteins with unknown structure. The folding of a protein is mainly dictated by interresidue interactions, which are influenced by physical, chemical, energetic and conformational properties of amino acid residues. In this work, I have related various amino acid properties with folding rates of two- and three-state proteins. I found that the classification of proteins into different structural classes shows a good correlation between amino acid properties and protein folding rates. I have set up linear regression models using amino acid properties for predicting the folding rates of proteins from the amino acid sequence along with structural class information. I found an excellent agreement between experimental and predicted protein folding rates, and the correlation coefficients are 0.97 and 0.93, with back-check prediction and jack-knife test, respectively.

MATERIALS AND METHODS

Experimental Folding Rates. The experimental folding rates of 32 two- and three-state proteins used in related

* Corresponding author phone: +81-3-3599-8046; fax: +81-3-3599-8081; e-mail: michael-gromiha@aist.go.jp.

10.1021/ci049757q CCC: $30.25 © 2005 American Chemical Society
Published on Web 01/26/2005

Table 1. Predicted Folding Rates in a Set of 32 Two- and Three-State Proteins[a]

PDB code	property						$\ln(k_f)$		
	α_c	K^0	P_β	R_a	ΔASA	ΔG_{hD}	pred	expt.	ref
				All-α *Proteins*					
1lmb	0.354						8.45	8.50	(45)
2abd	0.402						6.85	6.55	(46)
1imq	0.398						6.98	7.31	(47)
2pdd	0.313						9.80	9.80	(48)
1hrc	0.341						8.88	8.76	(49)
1ycc	0.320						9.57	9.62	(50)
				All-*β Proteins*					
1nyf		0.392	0.473	0.352	0.422		3.18	4.54	(14)
1pks		0.386	0.442	0.325	0.406		−1.39	-1.05	(51)
1shg		0.383	0.481	0.342	0.442		1.72	1.41	(38)
1srl		0.403	0.489	0.325	0.415		3.68	4.04	(52)
1fnf_9		0.466	0.479	0.345	0.410		−0.81	-0.91	(53)
1fnf_10		0.485	0.503	0.335	0.381		4.67	5.48	(54)
1hng		0.418	0.479	0.358	0.420		2.79	2.89	(54)
1ten		0.398	0.455	0.331	0.395		1.72	1.06	(55)
1tit		0.392	0.457	0.381	0.416		3.70	3.47	(54)
1wit		0.441	0.458	0.358	0.392		1.27	0.41	(54)
1csp		0.382	0.450	0.392	0.394		7.22	6.98	(56)
1mjc		0.431	0.457	0.346	0.366		4.47	5.24	(57)
2ait		0.438	0.502	0.322	0.394		5.54	4.20	(58)
				Mixed-Class Proteins					
1aps		0.421		0.328	0.403	0.632	−0.84	-1.48	(59)
1hdn		0.408		0.353	0.386	0.711	3.17	2.70	(60)
1urn		0.415		0.364	0.456	0.671	7.05	5.73	(61)
2hqi		0.452		0.333	0.368	0.729	−0.28	0.18	(62)
1pba		0.343		0.381	0.432	0.642	7.11	6.80	(63)
1ubo		0.394		0.354	0.420	0.678	6.00	7.33	(64)
2ptl		0.368		0.285	0.355	0.659	4.61	4.10	(65)
1fkb		0.442		0.347	0.407	0.696	1.63	1.46	(14)
1coa		0.414		0.423	0.454	0.708	4.92	3.87	(66)
1div		0.403		0.334	0.396	0.713	6.90	6.58	(67)
2vik		0.399		0.331	0.417	0.660	5.53	6.80	(14)
1cis		0.417		0.383	0.436	0.676	2.95	3.87	(14)
1pca		0.395		0.411	0.457	0.682	5.99	6.80	(14)

[a] α_c: Power to be at the C-terminal of α-helix;[17,26] K^0: Compressibility;[30] P_β: β-strand tendency;[17,26] R_a: Reduction in solvent accessibility;[31] ΔASA: Solvent accessible surface area for protein unfolding;[29] ΔG_{hD}: Gibbs free energy change of hydration for denatured protein;[29] $\ln(k_f)$: logarithms of folding rate. The numerical values of 20 amino acid residues for these six properties are given in Table 2. The regression equations are as follows: (i) ***all-α proteins***: $\ln(k_f) = -33.191\ (\pm\ 0.482)\ \alpha_c + 20.195\ (\pm\ 0.190)$; (ii) ***all-β proteins***: $\ln(k_f) = -81.48\ (\pm\ 1.687)\ K^0 + 163.08\ (\pm\ 1.484)\ P_\beta + 79.92\ (\pm\ 2.091)\ R_a - 134.99\ (\pm\ 1.781)\ \Delta ASA - 13.18\ (\pm\ 0.709)$; (iii) ***mixed class proteins***: $\ln(k_f) = -102.60\ (\pm\ 2.06)\ K^0 - 90.33\ (\pm\ 2.32)\ R_a + 131.80\ (\pm\ 1.95)\ \Delta ASA + 90.82\ (\pm\ 1.23)\ \Delta G_{hD} - 38.53\ (\pm\ 0.83)$.

works[3,4,14] form the basis for the present study. The Protein Data Bank codes[15] and experimental $\ln(k_f)$ values are given in Table 1. The structural classification of these proteins yielded six all-α, 13 all-β and 13 mixed class proteins. Further, I have validated the present method using a set of other 17 two- and three-state proteins.

Amino Acid Properties. I used a set of 49 diverse amino acid properties (physical-chemical, energetic and conformational), which fall into various clusters analyzed by Tommi and Kanehisa[16] in the present study. The amino acid properties were normalized between 0 and 1 using the expression, $P_{norm}(i) = [P(i) - P_{min}]/[P_{max} - P_{min}]$, where $P(i)$, $P_{norm}(i)$ 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 values of 20 amino acid residues for six selected properties are presented in Table 2. Further, 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[17,18] 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.

Table 2. Numerical Values of 20 Amino Acid Residues for Six Selected Properties

residue	property					
	α_c	K^0	P_β	R_a	ΔASA	ΔG_{hD}
Ala	1.44	−25.50	0.83	3.70	70.90	−0.58
Asp	2.13	−33.12	0.54	2.60	69.60	−6.10
Cys	0.76	−32.82	1.19	3.03	114.30	−1.91
Glu	2.01	−36.17	0.37	3.30	80.50	−7.37
Phe	1.01	−34.54	1.38	6.60	148.40	−1.35
Gly	0.62	−27.00	0.75	3.13	44.00	−0.82
His	0.56	−31.84	0.87	3.57	107.90	−5.57
Ile	0.68	−31.78	1.60	7.69	142.70	0.40
Lys	0.59	−32.40	0.74	1.79	87.50	−5.97
Leu	0.58	−31.78	1.30	5.88	129.80	0.35
Met	0.73	−31.18	1.05	5.21	147.90	−0.71
Asn	0.93	−30.90	0.89	2.12	74.00	−6.63
Pro	2.19	−23.25	0.55	2.12	73.50	0.56
Gln	1.20	−32.60	1.10	2.70	93.30	−7.12
Arg	0.39	−26.62	0.93	2.53	116.00	−12.78
Ser	0.81	−29.88	0.75	2.43	62.80	−6.18
Thr	1.25	−31.23	1.19	2.60	78.00	−3.66
Val	0.63	−30.62	1.70	7.14	115.60	0.18
Trp	1.40	−30.24	1.37	6.25	167.80	−4.71
Tyr	0.72	−35.01	1.47	3.03	145.90	−8.45

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

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

where $P(j)$ is the property value of jth residue and the summation is over N, the total number of residues in a protein. These property values have been obtained from the table of 20 amino acid residues (e.g. Table 2), and no structural information of a specific protein is required for computation. The computed property value $P_{ave}(i)$ for each class of proteins was related with experimental folding rate $\ln k_f(i)$ using single correlation coefficient. Further, I have combined the amino acid properties using multiple regression technique.[19] The statistical significance of the results obtained in the present study has been verified with *t*-test and *p*-value by standard procedures.

RESULTS AND DISCUSSION

Role of Protein Structural Classes. In our earlier work, we have analyzed the influence of interresidue interactions in different structural classes, and we found that the interacting pattern is distinct in each structural class.[20] Further, the analysis on the predictive accuracy of several secondary structure prediction algorithms indicates the necessity of structural classification for better performance.[21,22] On the other hand, several methods are available to predict the protein structural class with high accuracy.[23]

The relationship between LRO and protein folding rates shows that the classification of proteins into three structural classes significantly improved the correlation.[3] I have also reported that the single data set with the inclusion of proteins from all structural classes yielded a poor correlation ($|r| \leq 0.39$) between amino acid properties and protein folding rates.[24] In the present study, I found that the classification of proteins into all-α, all-β and mixed class remarkably enhanced the correlation from 0.39 to 0.97, and hence the structural classification is necessary for successful prediction of protein folding rates. This result reveals that the classification includes the information about the topology of protein, which is found to be an important determinant for protein folding rates.[25]

All-α Proteins. The relationship between amino acid properties and protein folding rates of all-α proteins shows that the property power to be at the C-terminal of α-helix (abbreviated as α_c)[17,26] has the highest negative correlation with protein folding rates ($r = -0.99$). Further, the thermodynamic and conformational properties show significant correlation (0.6 to 0.9) with $\ln(k_f)$ values. This observation reveals that the formation of local secondary structures enhances the folding rates of proteins and hence the folding process is the interplay between the local conformational preferences and long-range contacts.[27] The all-α proteins have higher folding rates than other classes of proteins, in general, and hence the folding rates of this class of proteins may be well explained with single amino acid property. In contrast, when I generated 49 sets of random numbers and computed the correlation coefficients, I have obtained the average r-value of 0.34 ± 0.23. This verifies that one could clearly discriminate between amino acid properties and random numbers and emphasizes the validity of selecting amino acid properties for understanding/predicting protein folding rates.

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

$$\ln(k_f) = -33.191\ (\pm\ 0.482)\ \alpha_c + 20.195\ (\pm\ 0.190) \qquad (2)$$

where α_c is the power to be at the C-terminal of α-helix. The numerical values of α_c and folding rates for the six all-α proteins are presented in Table 1. I observed an excellent agreement between the predicted folding rates and experimental observations. The correlation coefficient is 0.988 and average deviation is 0.142. Further, I have verified that the results are statistically significant ($t = 12.09$ and $p \leq$ 2.68e-4).

I have also performed the 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 folding rate of the omitted protein. I found that all the considered proteins agreed extremely well with experiment. The r-value is 0.958 ($t = 6.685$; $p \leq$ 2.69e-3), and the average deviation is 0.21. The cross-validation has also been performed with the complete data set of nine all-α class proteins (listed in Tables 1 and 3), and I obtained the correlation of 0.947 ($t = 7.735$; $p \leq$ 1.16e-3) between experimental and predicted folding rates. The R^2 value is 0.895, and the average deviation is 0.39.

Zhou and Zhou[4] used four all-α proteins and obtained the correlation of 0.92 between experimental and predicted folding rates. The r-values obtained with CO and LRO are, respectively, 0.56 and 0.72. Recently, Shao et al.[28] proposed a new term, "helix parameter", for predicting the protein folding rates and reported that it has the correlation of 0.927 with protein folding rate. The present model predicts the protein folding rates of all-α proteins at the highest accuracy of $r = 0.99$ and 0.96, respectively, with back-check prediction and jack-knife test. These correlation coefficients demonstrate the better performance of the present method compared to other methods in the literature.

Table 3. Validity Test for Predicting the Folding Rates of 17 Proteins

PDB code	$\ln(k_f)$		ref.
	pred	expt.	
	All-α *Proteins*		
1CEI	6.19	5.8	(47)
1ENH	9.44	10.53	(68)
1EBD	9.87	9.68	(48)
	All-β *Proteins*		
1PNJ	−3.16	−1.10	(51)
1SHF	2.91	4.50	(69)
1C9O	10.01	7.20	(70)
1G6P	5.32	6.30	(70)
1LOP	5.65	6.60	(71)
	Mixed Proteins		
1HZ6	5.09	4.10	(72)
1PGB	7.14	6.00	(73)
2CI2	4.87	3.90	(74)
1AYE	7.46	6.80	(63)
1POH	3.17	2.70	(60)
1AON	1.47	0.80	(75)
1BNI	1.79	2.60	(76)
2LZM	3.60	4.10	(77)
1UBQ	6.00	5.90	(78)

All-β Proteins. In all-β proteins, the thermodynamic properties show significant correlation with protein folding rates, and the highest single property correlation is −0.696 with unfolding entropy change.[29] Similar calculations with random numbers yielded the average correlation of 0.20 ± 0.16. As the single property with the highest r-value is not sufficient for accurate prediction I have combined different amino acid properties with a multiple regression fit. The computation has been carried out with the combinations of two to five amino acid properties, and I found that the combination of four properties substantially improved the correlation. The corresponding regression equation is

$$\ln(k_f) = -81.48\ (\pm\ 1.687)\ K^0 + 163.08\ (\pm\ 1.484)\ P_\beta + 79.92\ (\pm\ 2.091)\ R_a - 134.99\ (\pm\ 1.781)\ \Delta ASA - 13.18\ (\pm\ 0.709) \qquad (3)$$

where K^0, P_β, R_a, and ΔASA are, respectively, compressibility,[30] β-strand tendency,[26] reduction in solvent accessibility[31] and solvent accessible surface area for protein unfolding.[29] This result shows that the combination of physical, thermodynamic and conformational properties of amino acid residues enhances the folding rates of all-β proteins. Experimental observations support my findings that the local secondary structure and conformational parameters influence the folding rates of proteins[27,32] and thermodynamic parameters play an important role for determining the folding transition state structures of proteins.[33] The inclusion of other properties may improve the correlation, but it will cause over fitting of the parameters. On the other hand, the combination of four properties could predict the folding rates of two- and three-state proteins at high accuracy.

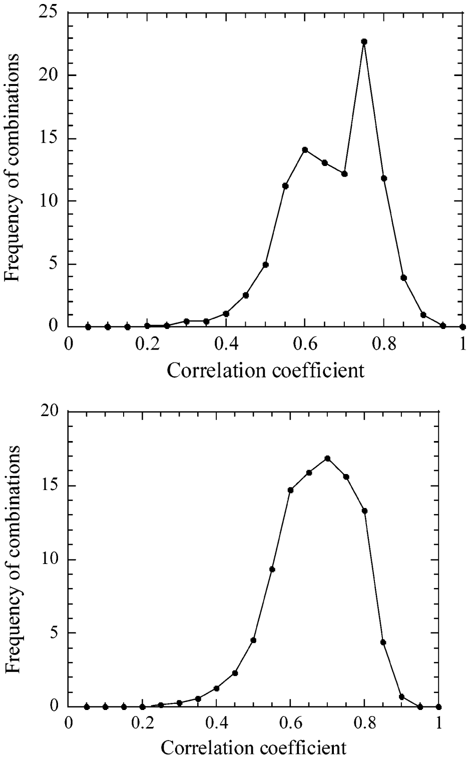

Figure 1. Frequency of the combinations of four amino acid properties at different ranges of correlation coefficients. (a) all-β proteins. (b) mixed class proteins.

I have analyzed the correlation coefficients obtained for a different combination of four amino acid properties, and the frequency of combinations at different ranges of correlation coefficients are shown in Figure 1a. From this figure, I noticed that most of the combinations have the correlation in the range of 0.5 to 0.8, and only two combinations have the correlation higher than 0.95. I have selected the properties that yielded the highest correlation coefficient for predicting protein folding rates.

The numerical values of the parameters, K^0, P_β, R_a and ΔASA for the considered all-β proteins along with their predicted and experimental $\ln(k_f)$ values are given in Table 1. I found that the predicted protein folding rates have a very good agreement with experimental observations and the correlation between them is 0.956. The calculated p value is $\leq$ 2.58e-5 and $t = 10.86$. I have also carried out the jack-knife test, and the correlation obtained with this method is 0.890 ($R^2 = 0.799$; $p \leq$ 1.85e-4 and $t = 6.616$).

Zhou and Zhou[4] used a set of 13 all-β proteins and reported that the correlation between predicted protein folding rates using total contact distance and experimental data is 0.69. With the same data set, the present method predicted the folding rate of two-state proteins with the remarkable correlation of 0.956. The deviation of $\ln(k_f)$ values obtained with total contact distance is 1.372, whereas the present method predicted the folding rate within the deviation of 0.575. Further, the standard error of each coefficient in Eqn. 3 is less than 5%. These results emphasize that the present method could be used for predicting the folding rates of unknown protein.

Mixed Class Proteins. The relationship between amino acid properties and protein folding rates in a set of 13 mixed class proteins shows that the average medium-range contacts[34,35] has the highest positive correlation ($r = 0.68$) with protein folding rates. This might be due to the fact that the short- and medium-range interactions may predominate and initiate protein folding with the formation of α-helices.[36] On the other hand, a set of random numbers yielded the average correlation of 0.27 ± 0.20.

I have analyzed the variation of correlation coefficients due to the combination of several amino acid properties by using multiple regression technique. The correlation coefficient obtained with single property is 0.68 and it raised to 0.79, 0.88, 0.95 and 0.95, respectively, with the combination of 2 to 5 properties. I observed that the combination of four properties significantly improved the correlation, and hence I have used the regression equation derived with four amino acid properties for predicting the folding rates of mixed class proteins. The regression equation is

$$\ln(k_f) = -102.60\ (\pm\ 2.06)\ K^0 - 90.33\ (\pm\ 2.32)\ R_a + 131.80\ (\pm\ 1.95)\ \Delta ASA + 90.82\ (\pm\ 1.23)\ \Delta G_{hD} - 38.53\ (\pm\ 0.83) \quad (4)$$

where ΔG_{hD} is the Gibbs free energy change of hydration for denatured protein.[29] Interestingly, the three properties, K^0, R_a and ΔASA, are also important for determining the folding rates of all-β proteins. The combination of free energy with other thermodynamic and physical properties influences the fast folding rates of two- and three-state proteins. Previously, it has been shown that the combination of free energy and contact order determines the folding rates.[9]

The analysis on the variation of correlation coefficients at different combinations of amino acid properties revealed that most of the combinations yielded the correlation in the range of 0.5 to 0.8, as observed in all-β proteins (Figure 1b). The combination of the properties, K^0, R_a, ΔASA and ΔG_{hD}, has the highest correlation. The numerical values for all the properties and predicted $\ln(k_f)$ values are presented in Table 1. I found an excellent correlation of 0.95 between experimental and predicted protein folding rates. Further, I have corroborated the statistical significance of the results ($p \leq$ 3.02e-5 and $t = 10.3$).

Recent studies on protein folding rates showed that the $\ln(k_f)$ values of mixed class proteins have been predicted with high accuracy. The correlation coefficient obtained with CO, LRO and TCD are, respectively, 0.82, 0.86 and 0.92. The present method predicted the folding rates of two-state proteins with the highest accuracy, and the correlation coefficient is 0.95. I have also performed a jack-knife test and obtained the r-value of 0.87 ($R^2 = 0.749$; $p \leq$ 3.9e-4; $t = 5.729$). The standard errors obtained for fitting the data are also included in Eqn. 4, which is less than 3%. Further, I have estimated the deviation of $\ln(k_f)$ values, and the average deviation obtained for back-check prediction and jack-knife test are, respectively, 0.737 and 1.253.

Prediction of Protein Folding Rates. I have used three different equations (Eqns. 2−4) for predicting the folding rates of proteins belonging to each structural class. The results obtained with back-check prediction are presented in Table 1. I found an excellent correlation of 0.97 between predicted and experimental protein folding rates for the sample set of 32 proteins as seen in Figure 2a. The t-test and p-values are, respectively, 23.54 and $\leq$ 2.98e-13, and

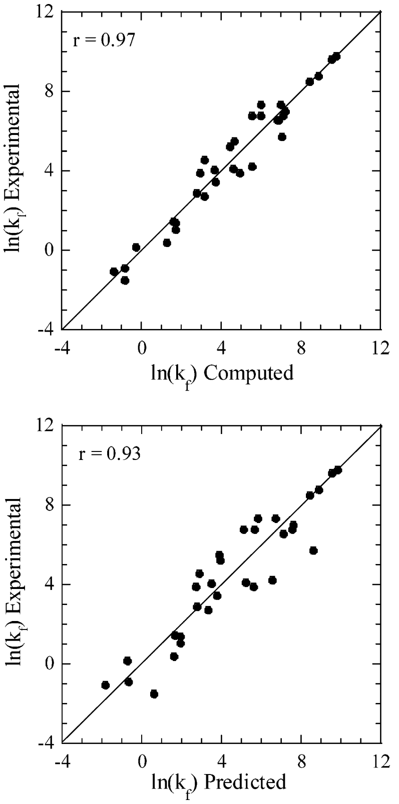

Figure 2. Relationship between experimental and predicted $\ln(k_f)$ values using multiple regression model in 32 two-state proteins. (a) back-check prediction. (b) jack-knife test.

the average deviation is 0.56. I have also performed a jack-knife test to examine the validity of the present method, and the results are shown in Figure 2b. I found that about 60% of the considered proteins (19 out of 32 proteins) agreed very well with the experiment and the deviation is less than one unit. The correlation coefficient between experimental and predicted $\ln(k_f)$ values is 0.93 ($t = 13.8$; $p \leq 2.11$ e-11), and the average deviation is 0.931.

Validating the Present Method. I have calculated the folding rates of other 17 two- and three-state proteins belonging to different structural classes and compared the predicted $\ln(k_f)$ values with experimental observations. I have presented the list of proteins along with predicted and experimental folding rates in Table 3. In all-α proteins, all the three proteins are predicted within the deviation of 1.1. The correlation between predicted and experimental $\ln(k_f)$ values in all-β proteins is 0.95. In mixed proteins, 8 out of 9 proteins are predicted within the deviation of 1.0. Considering all the 17 proteins together, the correlation coefficient is 0.94, and the average deviation is 0.96. Recently, Scott et al.[37] reported the folding rate of α-spectrin 15th domain for which the three-dimensional structure is not available. I have calculated the folding rate using Eqn. 2 and observed a good agreement between predicted and experimental folding rates (within the deviation of 2.0). This result emphasizes the validity of the present method for predicting the protein folding rates.

Prediction of Protein Folding Rates upon Mutations. I have examined the performance of the present method for predicting the protein folding rates upon mutations. Viguera et al.[38] measured the experimental folding rates of SH3 domain of src for the wild-type protein and seven mutants.

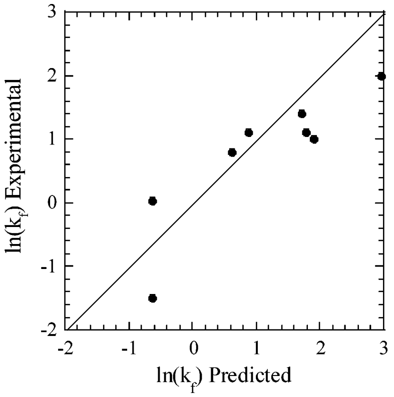

Figure 3. Comparison between experimental and predicted folding rates in the mutants of SH3 src domain. Experimental data are taken from ref 38.

Table 4. Prediction of Protein Folding Rates upon Mutations in Barnase

	$\Delta\ln(k_f)$	
mutant	experiment[a]	predicted
D 12 G	−0.01	−0.50
D 12 A	0.20	−0.30
Y 13 G	−0.48	−1.17
Q 15 G	−0.18	−0.44
Q 15 A	0.22	−0.37
Y 17 G	−0.94	−1.17
I 55 G	0.01	−0.79
R 72 G	−0.12	−0.02
E 73 G	−0.08	−0.53
I 88 G	−1.72	−0.79
L 89 G	0.32	−0.93
L 95 G	−0.75	−0.93
I 96 G	−3.86	−0.79
Y 97 G	−2.94	−1.17
T 100G	−0.38	−0.48

[a] Data taken from ref 39.

This protein belongs to all-β class, and I have calculated the folding rates using Eqn. 3. The relationship between experimental and predicted folding rates is shown in Figure 3, and I observed a good agreement. The r value is 0.87, and the average deviation is 0.60.

Vu et al.[39] constructed a pseudo wild type of barnase, in which residue His102 is mutated to Ala. They have measured the folding rates of 15 mutants under the background of pseudo wild-type protein. Barnase belongs to a mixed class protein, and I have calculated the difference of folding rates for each of the 15 mutants with respect to pseudo wild-type protein using Eqn. 4 and the results are presented in Table 4. I found that 12 out of 15 mutants are predicted within the deviation of 1.0. This data set contains several mutants with the same amino acid replacements at different positions (e.g., L89G and L95G; Y13G, Y17G and Y97G), and the present method is not able to distinguish them.

From these results I observed that the positional parameters play an important role in understanding the protein folding rates upon mutations, and the analysis based on the separation of mutants based on secondary structure and solvent accessibility may improve the accuracy of prediction as in the case of protein mutant stability and Φ value analysis.[40−42]

Comparison with Other Methods. The protein folding rate predictive ability of eight different methods along with the results obtained with the present multiple regression model are presented in Table 5.

Table 5. Comparison of Protein Folding Rate Predictive Ability of Eight Other Methods with the Present Method

method	parameter	information	*r*	reference
linear single regression	contact order	3D structure	0.74	Plaxco et al.[2]
first principles approach	residue−residue contacts	3D structure	0.78	Debe and Goddard[7]
statistical mechanical model	residue−residue contacts	3D structure	0.83	Munoz and Eaton[8]
linear single regression	long-range order	3D structure	0.81	Gromiha and Selvaraj[3]
neural network	CO and free energy change	3D structure	0.79	Dinner and Karplus[9]
linear single regression	total contact distance	3D structure	0.88	Zhou and Zhou[4]
topomer search model	long-range contacts	3D structure	0.89	Makarov et al.[43]
linear multiple regression	secondary structure content	3D structure	0.91	Gong et al.[44]
linear multiple regression	amino acid properties	**1D sequence**	**0.97**	present work

Plaxco et al.[2] proposed the concept of contact order and related it with protein folding rates. Gromiha and Selvaraj[3] introduced the parameter, long-range order from the knowledge of long-range contacts in protein structures for predicting protein folding rates. Zhou and Zhou[4] combined the parameters CO and LRO and formed the term, total contact distance to relate with protein folding rates. The correlation coefficients obtained with CO, LRO and TCD are, respectively, 0.74, 0.81 and 0.88. Debe and Goddard[7] proposed a method based on first principles approach and reported the correlation of 0.78 between experimental and predicted folding rates. Munoz and Eaton[8] developed a simple statistical model and obtained the correlation of 0.83 between theory and experiment. Dinner and Karplus[9] combined CO and free energy change to predict the folding rates of two-state proteins and reported the correlation of 0.79. Recently, Makarov et al.[43] developed a topomer search model, which increased the correlation to 0.89. Further, Gong et al.[44] proposed a multiple regression model for predicting the protein folding rates with their local secondary structure content and reported the correlation coefficient of 0.91. The present method shows the correlation of 0.97 and 0.93 between experimental and predicted folding rates with the back-check and jack-knife tests, respectively. These accuracy levels are better than other methods in the literature. The high accuracy attained by the present method may be due to the following reasons: (i) it revealed the important amino acid properties for accelerating protein folding rates, (ii) the combination of properties has been systematically selected for understanding/predicting the folding rates of proteins, and (iii) the selected properties are reliable in understanding protein folding rates as demonstrated from experiments.

Although the direct comparison of correlation coefficients obtained in the present work with the other methods is not appropriate, the empirical relationships derived for different structural classes predict the folding rates with greatest accuracy. Further, all the other methods use the three-dimensional structure information for predicting the protein folding rates. The present method is the first available one, which uses the amino acid sequence (along with structural class information) for predicting the folding rates of proteins. These comparisons reveal the superior performance of the present method for predicting the folding rates of proteins.

CONCLUSIONS

The interaction of amino acid residues among themselves and with surrounding medium dictates the structure of a protein and hence the rate of folding. Interresidue interactions are mainly influenced by physical-chemical, energetic and conformational properties of amino acid residues in protein structures. I have systematically analyzed the relationship between amino acid properties and protein folding rates in different structural classes of proteins. I have set up linear regression models for predicting the folding rates of two- and three-state proteins using the combination of amino acid properties. The present method is the first one, which can predict the protein folding rates from amino acid sequence along with structural class information. The predicted folding rates show an excellent correlation with experimental observations; the correlation coefficients are 0.99, 0.96 and 0.95, respectively, for all-α, all-β and mixed class proteins. These accuracy levels are superior to other methods in the literature.

ACKNOWLEDGMENT

The author thanks Dr. S. Selvaraj and Dr. T. Kin for helpful discussions.

Supporting Information Available: Numerical values (Table 1) and normalized values (Table 2) for and brief descriptions (Table 3) of 49 selected physicochemical, energetic and conformational properties of the 20 amino acids/residues. 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) 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.

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

(5) 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.

(6) Miller, E. J.; Fischer, K. F.; Marqusee, S. Experimental Evaluation of Topological Parameters Determining Protein-Folding Rates. *Proc. Natl. Acad. Sci. U.S.A.* **2002**, *99*, 10359−10363.

(7) Debe, D. A.; Goddard, W. A. 3rd. First principles prediction of protein folding rates. *J. Mol. Biol.* **1999**, *294*: 619−625.

(8) 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.

(9) Dinner, A. R.; Karplus, M. The roles of stability and contact order in determining protein folding rates. *Nat. Struct. Biol*. **2001**, *8*, 21−22.

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

(11) 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.

(12) Micheletti, C. Prediction of folding rates and transition-state placement from native-state geometry. *Proteins* **2003**, *51*, 74−84.

(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) Jackson, S. E. How do small single-domain proteins fold? *Fold Des*. **1998** *3*, R81−91.

(15) 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.

(16) 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.

(17) 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.

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

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

(20) Gromiha, M. M.; Selvaraj, S. Importance of long-range interactions in protein folding. *Biophys Chem*. **1999**, *77*, 49−68.

(21) Rost, B.; Sander, C. Secondary structure prediction of all-helical proteins in two states. *Protein Eng*. **1993**, *6*, 831−836.

(22) Gromiha, M. M.; Selvaraj, S. Protein secondary structure prediction in different structural classes. *Protein Eng*. **1998**, *11*, 249−251.

(23) Chou, K. C. Prediction of protein structural classes and subcellular locations. *Curr. Protein Pept. Sci.* **2000**, *1*, 171−208.

(24) 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.

(25) 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.

(26) Chou, P. Y.; Fasman, G. D. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzym*. **1978**, *47*, 45−148.

(27) Kaya, H.; Chan, H. S. Contact order dependent protein folding rates: kinetic consequences of a cooperative interplay between favorable nonlocal interactions and local conformational preferences. *Proteins* **2003**, *52*, 524−533.

(28) 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.

(29) Oobatake, M.; Ooi, T. Hydration and heat stability effects on protein unfolding. *Prog. Biophys. Mol. Biol*. **1993**, *59*, 237−284.

(30) Iqbal, M.; Verrall, R. E. Implications of protein folding. Additivity schemes for volumes and compressibilities. *J. Biol. Chem*. **1988**, *263*, 4159−4165.

(31) Ponnuswamy, P. K.; Prabhakaran, M.; Manavalan, P. Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins. *Biochim. Biophys. Acta* **1980**, *623*, 301−316.

(32) Mirny, L.; Shakhnovich, E. Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct*. **2001**, *30*, 361−396.

(33) Main, E. R.; Fulton, K. F.; Jackson, S. E. Folding pathway of FKBP12 and characterisation of the transition state. *J. Mol. Biol*. **1999**, *291*, 429−444.

(34) Gromiha, M. M.; Selvaraj, S. Influence of Medium and Long-Range Interactions in Different Structural Classes of Globular Proteins. *J. Biol. Phys.* **1997**, *23*, 151−162.

(35) Gromiha, M. M.; Selvaraj, S. Interresidue interactions in protein folding and stability. *Prog. Biophys. Mol. Biol.* **2004**, *86*, 235−277.

(36) Unger, R.; Moult, J. Local Interactions Dominate Folding in a Simple Protein Model. *J. Mol. Biol*. **1996**, *259*, 988−994.

(37) Scott, K. A.; Batey, S.; Hooton, K. A.; Clarke, J. The folding of spectrin domains I: wild-type domains have the same stability but very different kinetic properties. *J. Mol. Biol*. **2004**, *344*, 195−205.

(38) Viguera, A. R.; Serrano, L.; Wilmanns, M. Different folding transition states may result in the same native structure. *Nat. Struct. Biol*. **1996**, *3*, 874−880.

(39) Vu, N. D.; Feng, H.; Bai, Y. The folding pathway of barnase: the rate-limiting transition state and a hidden intermediate under native conditions. *Biochemistry* **2004**, *43*, 3346−3356.

(40) 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.

(41) 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.

(42) 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.

(43) 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.

(44) 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.

(45) Burton, R. E.; Huang, G. S.; Daugherty, M. A.; Fullbright, P. W.; Oas, T. G. Microsecond protein folding through a compact transition state. *J. Mol. Biol.* **1996**, *263*, 311−322.

(46) Kragelund, B. B.; Hojrup, P.; Jensen, M. S.; Schjerling, C. K.; Juul, E.; Knudsen, J.; Poulsen, F. M. Fast and one-step folding of closely and distantly related homologous proteins of a four helix bundle family. *J. Mol. Biol.* **1996**, *256*, 187−200.

(47) Ferguson, N.; Capaldi, A. P.; James, R.; Kleanthous, C.; Radford, S. E. Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J. Mol. Biol.* **1999**, *286*, 1597−1608.

(48) Spector, S.; Raleigh, D. P. Submillisecond folding of the peripheral subunit-binding domain. *J. Mol. Biol.* **1999**, *293*, 763−768.

(49) Chan, C. K.; Hu, Y.; Takahashi, S.; Rousseau, D. L.; Eaton, W. A.; Hofrichter, J. Submillisecond protein folding kinetics studied by ultrarapid mixing. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 1779−84.

(50) Mines, G. A.; Pascher, T.; Lee, S. C.; Winkler, J. R.; Gray, H. B. Cytochrome-c folding triggered by electron-transfer. *Chem. Biol.* **1996**, *3*. 491−497.

(51) Guijarro, J. I.; Morton, C. J.; Plaxco, K. W.; Pitkeathly, M.; Campbell, I. D.; Dobson, C. M. Folding kinetics of the SH3 domain of PI3 kinase by real-time NMR combined with optical spectroscopy. *J. Mol. Biol.* **1998**, *276*, 657−667.

(52) Grantcharova, V. P.; Baker, D. Folding dynamics of the src SH3 domain. *Biochemistry* **1997**, *36*, 15685−92.

(53) Plaxco, K. W.; Spitzfaden, C.; Campbell, I. D.; Dobson, C. M. A comparison of the folding kinetics and thermodynamics of two homologous fibronectin type III modules. *J. Mol. Biol.* **1997**, *270*, 763−770.

(54) Clarke, J.; Cota, E.; Fowler, S. B.; Hamill, S. J. Folding studies of immunoglobulin-like β-sandwich proteins suggest that they share a common folding pathway. *Struct. Fold. Des.* **1999**, *7*, 1145−1153.

(55) Clarke, J.; Hamill, S. J.; Johnson, C. M. Folding and stability of a fibronectin type III domain of human tenascin. *J. Mol. Biol*. **1997**, *270*, 771−8.

(56) Schindler, T.; Schmid, F. X. Thermodynamic properties of an extremely rapid protein folding reaction. *Biochemistry* **1996**, *35*, 16833−16842.

(57) Reid, K. L.; Rodriguez, H. M.; Hillier, B. J.; Gregoret, L. M. Stability and folding properties of a model beta-sheet protein, *Escherichia coli* CspA. *Protein Sci*. **1998**, *7*, 470−9.

(58) Schonbrunner, N.; Koller, K. P.; Kiefhaber, T. Folding of the disulfide-bonded β-sheet protein tendamistat: rapid two-state folding without hydrophobic collapse. *J. Mol. Biol.* **1997**, *268*, 526−538.

(59) Nuland, N. A. J. V.; Chiti, F.; Taddei, N.; Raugei, G.; Ramponi, G.; Dobson, C. M. Slow folding of muscle acylphosphatase in the absence of intermediates. *J. Mol. Biol.* **1998**, *283*, 883−891.

(60) Nuland, N. A. J. V.; Meijberg, W.; Warner, L.; Forge, V.; Scheek, R. M.; Robillard, G. T.; Dobson, C. M. Slow cooperative folding of a small globular protein HPr. *Biochemistry* **1998**, *37*, 622−637.

(61) Otzen, D. E.; Kristensen, O.; Proctor, M.; Oliveberg, M. Structural changes in the transition state of protein folding: alternative interpretations of curved chevron plots. *Biochemistry* **1999**, *38*, 6499−6511.

(62) Aronsson, G.; Brorsson, A.-C.; Sahlman, L.; Jonsson, B.-H. Remarkably slow folding of a small protein. *FEBS Lett.* **1997**, *411*, 359−364.

(63) Villegas, V.; Azuaga, A.; Catasus, L.; Reverter, D.; Mateo, P. L.; Aviles, F. X.; Serrano, L. Evidence for a 2-state transition in the folding process of the activation domain of human procarboxypeptidase-A2. *Biochemistry* **1995**, *34*, 15105−15110.

(64) Khorasanizadeh, S.; Peters, I. D.; Butt, T. R.; Roder, H. Folding and stability of a tryptophan-containing mutant of ubiquitin. *Biochemistry* **1993**, *32*, 7054−7063.

(65) Scalley, M. L.; Baker, D. Protein folding kinetics exhibit an Arrhenius temperature dependence when corrected for the temperature dependence of protein stability. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 10636−10640.

(66) Tan, Y.−J.; Oliveberg, M.; Fersht, A. R. Titration properties and thermodynamics of the transition state for folding: comparison of two-state and multistate folding pathways. *J. Mol. Biol.* **1996**, *264*, 377−389.

(67) Kuhlman, B.; Luisi, D. L.; Evans, P. A.; Raleigh, D. P. Global analysis of the effects of temperature and denaturant on the folding and unfolding kinetics of the N-terminal domain of the protein L9. *J. Mol. Biol.* **1998**, *284*, 1661−1670.

(68) Mayor, U.; Johnson, C. M.; Daggett, V.; Fersht, A. R. Protein folding and unfolding in microseconds to nanoseconds by experiment and simulation. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 13518−13522.

(69) Plaxco, K. W.; Guijarro, J. I.; Morton, C. J.; Pitkeathly, M.; Campbell, I. D.; Dobson, C. M. The folding kinetics and thermodynamics of the Fyn-SH3 domain. *Biochemistry* **1998**, *37*, 2529−2537.

(70) Perl, D.; Welker, C.; Schindler, T.; Schroder, K.; Marahiel, M. A.; Jaenicke, R.; Schmid, F. X. Conservation of rapid two-state folding in mesophilic, thermophilic and hyperthermophilic cold shock proteins. *Nat. Struct. Biol.* **1998**, *5*, 229−235.

(71) Ikura, T.; Hayano, T.; Takahashi, N.; Kuwajima, K. Fast folding of *Escherichia coli* cyclophilin A: a hypothesis of a unique hydrophobic core with a phenylalanine cluster. *J. Mol. Biol.* **2000**, *297*, 791−802.

(72) Kim, D. E.; Fisher, C.; Baker, D. A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **2000**, *298*, 971−984.

(73) McCallister, E. L.; Alm, E.; Baker, D. Critical role of beta-hairpin formation in protein G folding. *Nat. Struct. Biol.* **2000**, *7*, 669−673.

(74) Jackson, S. E.; Fersht, A. R. Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry* **1991**, *30*, 10428−10435.

(75) Golbik, R.; Zahn, R.; Harding, S. E.; Fersht, A. R. Thermodynamic stability and folding of GroEL minichaperones. *J. Mol. Biol*. **1998**, *276*, 505−515.

(76) Matouschek, A.; Kellis, J. T., Jr.; Serrano, L.; Bycroft, M.; Fersht, A. R. Transient folding intermediates characterized by protein engineering. *Nature* **1990**, *346*, 440−445.

(77) Parker, M. J.; Sessions, R. B.; Badcoe, I. G.; Clarke, A. R. The development of tertiary interactions during the folding of a large protein. *Fold. Des*. **1996**, *1*, 145−156.

(78) Khorasanizadeh, S.; Peters, I. D.; Roder, H. Evidence for a three-state model of protein folding from kinetic analysis of ubiquitin variants with altered core residues. *Nat. Struct. Biol.* **1996**, *3*, 193−205.

CI049757Q