doi:10.1006/jmbi.2001.4775 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) **310**, 27–32

JMB

COMMUNICATION

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

M. Michael Gromiha[1]* and S. Selvaraj[1,2]

[1]*RIKEN Tsukuba Institute*
3-1-1 Koyadai, Tsukuba
Ibaraki, 305-0074
Japan

[2]*Bharathidasan University*
Tiruchirapalli 620 024, Tamil
Nadu, India

**Corresponding author*

The contact order is believed to be an important factor for understanding protein folding mechanisms. In our earlier work, we have shown that the long-range interactions play a vital role in protein folding. In this work, we analyzed the contribution of long-range contacts to determine the folding rate of two-state proteins. We found that the residues that are close in space and are separated by at least ten to 15 residues in sequence are important determinants of folding rates, suggesting the presence of a folding nucleus at an interval of approximately 25 residues. A novel parameter "long-range order" has been proposed to predict protein folding rates. This parameter shows as good a relationship with the folding rate of two-state proteins as contact order. Further, we examined the minimum limit of residue separation to determine the long-range contacts for different structural classes. We observed an excellent correlation between long-range order and folding rate for all classes of globular proteins. We suggest that in mixed-class proteins, a larger number of residues can serve as folding nuclei compared to all-α and all-β proteins. A simple statistical method has been developed to predict the folding rates of two-state proteins using the long-range order that produces an agreement with experimental results that is better or comparable to other methods in the literature.

Keywords: contact order; folding rate; long-range order; protein folding; structural classes

Predicting the native structures of proteins from their amino acid sequences has remained an elusive goal for many years. A related and perhaps equally challenging task is to understand the relationship between sequences and folding rates of proteins.[1] As an advance on this problem, Plaxco *et al.*[2] found a significant correlation between the average sequence separation of all contacting residues in the native state, defined by the parameter contact order (CO) and the rate constants of folding of two-state proteins. Contact order is defined as $CO = \Sigma\Delta S_{ij}/LN$, where N is the total number of contacts in a protein, ΔS_{ij} is the number of residues separating contacts i and j, and L is the number of residues in the protein. This parameter reflects the relative importance of local and non-local contacts in protein structures. Recently, Fersht[3] has proposed an extended nucleus mechanism that relates contact order, chain topology and stability, and protein folding rates.

Debe & Goddard[4] 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 & Eaton[5] have proposed a simple statistical model to calculate the folding rates for 22 proteins from their three-dimensional structures and observed a correlation of 0.83 between predicted and experimental folding rates.

It has been well established that the inter-residue interactions (short, medium and long-range) play

Abbreviations used: *CO*, contact order; *LRO*, long-range order.

E-mail address of the corresponding author: gromiha@rtc.riken.go.jp an important role in the folding and stability of globular proteins.[6,7] Specifically, the long-range contacts within a limit of certain residues of separation in a sequence influence protein structures strongly.[7,8] In view of this fact, in this work we have analyzed the role of long-range contacts in determining the folding rate of two-state proteins. We have developed a simple, statistical model to predict the folding rate of 23 small, two-state proteins and observed an excellent agreement between experimental and predicted folding rates.

Concept of long-range order

We define a parameter, long-range order (*LRO*) for a protein from the knowledge of long-range contacts (contacts between two residues that are close in space and far in the sequence) in protein structure. It is defined as:

$$LRO = \Sigma n_{ij}/N \quad n_{ij} = 1 \text{ if } |i-j| > 12 \\ = 0 \text{ otherwise} \tag{1}$$

where *i* and *j* are two residues for which the C^{α}-C^{α} distance is $\leqslant 8$ Å and *N* is the total number of residues in a protein. The details of the computation of inter-residue contacts have been described in our recent review.[9]

Relationship between *LRO* and folding rate of proteins

The computed *LRO* using equation (1) for a set of 23 small, two-state proteins are presented in Table 1 along with the *CO* and folding rate of each protein. We found a strong inverse relationship between *LRO* and ln(k) for all the considered proteins ($r = -0.78$). Further, we have computed the *LRO*, varying the minimum distance of separation between two interacting residues from one to 50 residues and examined the correlation between *LRO* and folding rates. The results presented in Figure 1 indicate that the minimum distance of 12 residues in defining *LRO* has the best correlation between *LRO* and folding rates, and a significant correlation is obtained for the minimum residue separation of ten to 15 residues. This observation suggests the presence of key residues, responsible for the formation of folding nucleus at an interval of approximately 25 residues, in agreement with the experimentally observed distance of separation between residues forming folding nuclei in chymotrypsin inhibitor and src SH3 domain.[10,11] Further analysis on the role of short and medium-range contacts in determining the folding rate demonstrates that these effects are minimal ($r = 0.22$ and 0.46, respectively) compared with long-range contacts.

Interestingly, we observed a strong correlation between *LRO* and *CO* ($r = 0.77$), and both these parameters show a good relationship with folding rate. We have examined the influence of distance from the central residue, ranging from 4-20 Å in steps of 0.5 Å and analyzed the correlation between *LRO* and folding rate. We found that the limit of 8 Å is the best for predicting the folding rate of proteins. This limit has been used for studies of e.g. the hydrophobic character of proteins[12], prediction of protein stability upon mutations,[13] prediction of folding rates.[4]

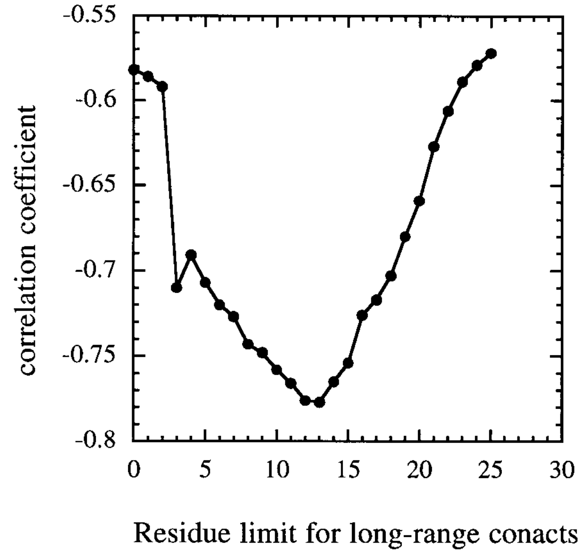

Figure 1. Plot connecting the correlation coefficient obtained between long-range contacts and folding rate of proteins and the minimum limit to define long-range contacts.

Folding rate of proteins in different structural classes

In our earlier work,[7] we observed different ranges of residue separation in each of the structural classes for contributing towards long-range contacts. On this basis, we have grouped the proteins into three classes, all-α, all-β and mixed, to analyze the relationship between *LRO*, folding rate and the minimum distance of separation between two interacting residues. We found an excellent relationship between *LRO* and folding rate of all classes of proteins (Figure 2(a)-(c)). Interestingly, we observed that the minimum limit of residue separation to represent long-range contacts is different for these classes of proteins; it is 27 for all-α proteins, 44 for all-β proteins and ten for mixed-class proteins. The structural analysis of all-β proteins considered in this work showed the presence of several inter-residue contacts between two β-strands at an interval of 40-50 residues. Further, we found an improved correlation only within the range of ±4 residues from the respective cut-off lengths. These cut-off optima might be refined when folding rates and three-dimensional structures are known for several two-state proteins. This result indicates the presence of more residues serving as folding nuclei in mixed proteins than in

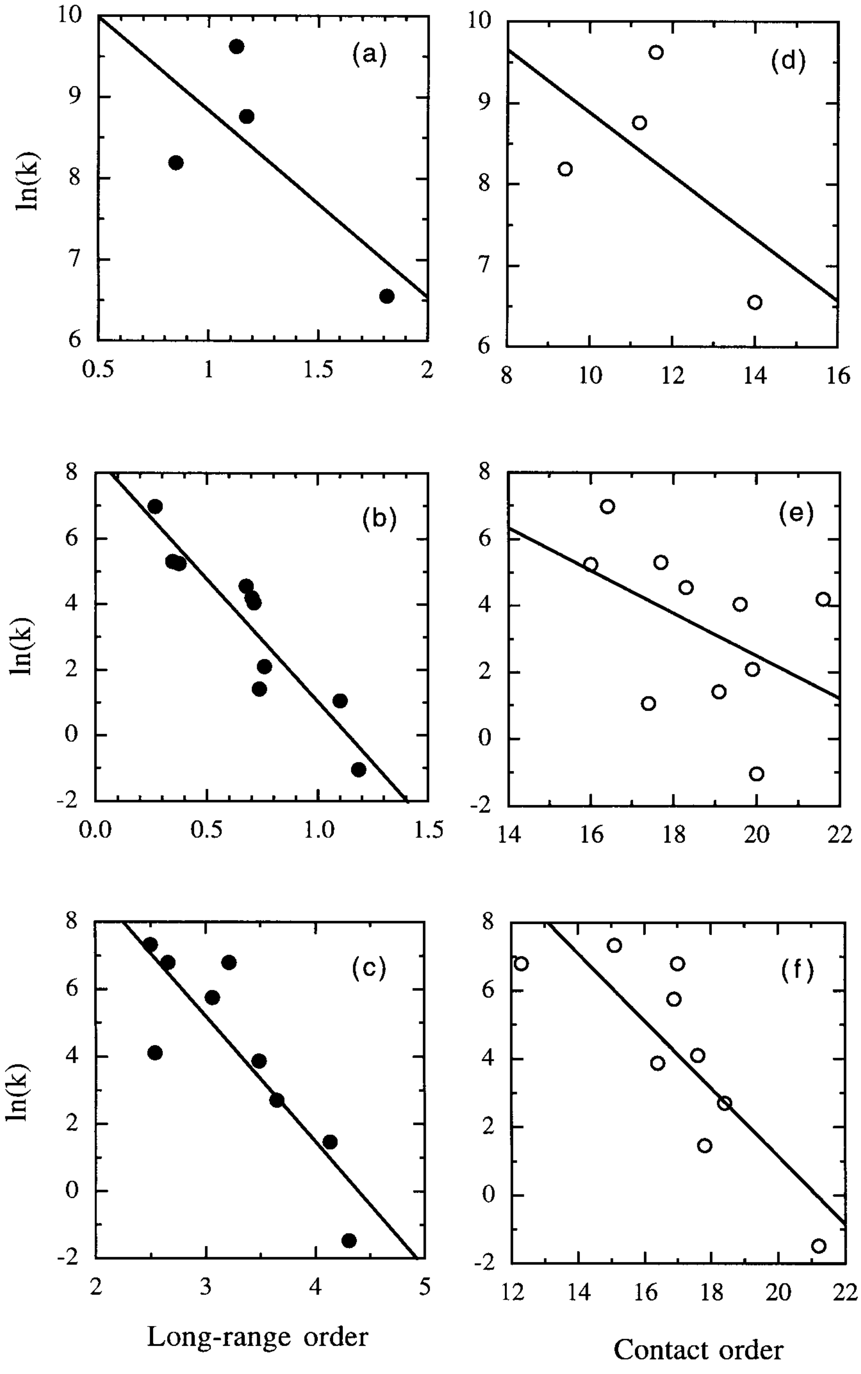

Figure 2. (a)-(c) Relationship between *LRO* and folding rate in different structural classes: (a) all-α proteins ($r = -0.72$); (b) all-β proteins ($r = -0.92$) and (c) mixed-class proteins ($r = -0.86$). (d)-(f) Relationship between *CO* and folding rate: (d) all-α proteins ($r = -0.56$); (e) all-β proteins ($r = -0.46$); and (f) mixed-class proteins ($r = -0.82$).

all-α and all-β proteins, in agreement with experimental observations that more residues have high phi values (>0.8) in mixed-class proteins than all-α and all-β proteins.[10,14,15] As the information about the structural class is available for a protein of known three-dimensional structure, this study reveals the necessity of grouping proteins into different structural classes for improved prediction of folding rates, similar to protein secondary structure prediction.[16]

Comparison with contact order

The relationship between *CO* and folding rate of protein showed that the *CO* has a strong negative correlation for mixed-class proteins ($r = -0.82$) as shown in Figure 2(f) and the correlation is much weaker for all-α ($r = -0.56$) and all-β ($r = -0.46$) proteins (Figure 2(d) and (e)). This is probably due to the same treatment of all structural classes in defining the parameter. On the other hand, our

Table 1. Long-range order, contact order and folding rate of 23 globular proteins

PDB code	*LRO*	*CO*	ln(*k*)
A. *All-α proteins*			
1LMB	1.126	9.400	8.190
1HRC	2.212	11.200	8.760
2ABD	2.302	14.000	6.550
1YCC	2.214	11.600	9.620
B. *All-β proteins*			
1CSP	3.045	16.400	6.980
1TEN	3.888	17.400	1.060
1SHF	2.847	18.300	4.550
2AIT	4.135	21.600	4.200
3MEF	2.957	17.700	5.300
1MJC	2.986	16.000	5.240
1AEY	3.000	19.900	2.090
1SHG	3.018	19.100	1.410
1SRL	3.107	19.600	4.040
1PKS	3.842	20.000	−1.050
C. *Mixed-class proteins*			
1UBQ	2.368	15.100	7.330
1CIS	3.333	16.400	3.870
1PCA	2.553	17.000	6.800
2PTL	2.231	17.600	4.100
1HDN	3.459	18.400	2.700
1APS	4.184	21.200	−1.480
1URN	2.917	16.900	5.760
1FKB	3.963	17.800	1.460
1VIK	2.970	12.300	6.800

LRO was computed using equation (1); contact order (*CO*) and folding rate, ln(*k*) are taken from Jackson.[17]

Table 2. Experimentally observed and theoretically predicted folding rates for 23 two-state globular proteins

			Prediction		
PDB	*LRO*	Experimental ln(*k*)	Back-check	Jack-knife test	DG
A. *All-α proteins*					
1LMB	0.851	8.190	9.181 (−0.991)	10.404 (−2.214)	6.56
1HRC	1.173	8.760	8.444 (0.316)	8.323 (0.437)	
2ABD	1.814	6.550	6.976 (−0.426)	11.114 (−4.564)	3.78
1YCC	1.126	9.620	8.551 (1.069)	8.132 (1.488)	
B. *All-β proteins*					
1CSP	0.269	6.980	6.503 (0.477)	6.281 (0.699)	1.31
1TEN	1.101	1.060	0.297 (0.763)	−0.053 (1.113)	
1SHF	0.678	4.550	3.452 (1.098)	3.327 (1.223)	
2AIT	0.703	4.200	3.266 (0.934)	3.158 (1.042)	0.60
3MEF	0.348	5.300	5.914 (−0.614)	6.106 (−0.806)	
1MJC	0.377	5.240	5.698 (−0.458)	5.822 (−0.582)	
1AEY	0.759	2.090	2.848 (−0.758)	2.935 (−0.845)	
1SHG	0.737	1.410	3.012 (−1.602)	3.193 (−1.783)	
1SRL	0.714	4.040	3.184 (0.856)	3.084 (0.956)	
1PKS	1.184	−1.050	−0.323 (−0.727)	0.164 (1.214)	−2.97
C. *Mixed-class proteins*					
1UBQ	2.500	7.330	7.070 (0.260)	6.974 (0.356)	0.92
1CIS	3.485	3.870	3.386 (0.484)	3.324 (0.546)	
1PCA	2.660	6.800	6.472 (0.328)	6.385 (0.415)	
2PTL	2.538	4.100	6.928 (−2.828)	7.966 (−3.866)	0.28
1HDN	3.647	2.700	2.780 (−0.080)	2.800 (−0.100)	−0.50
1APS	4.306	−1.480	0.316 (−1.796)	1.554 (3.034)	−1.93
1URN	3.063	5.760	4.964 (0.796)	4.857 (0.903)	5.77
1FKB	4.131	1.460	0.970 (0.490)	0.754 (0.706)	2.46
1VIK	3.212	6.800	4.407 (2.393)	4.110 (2.690)	

The Back-check is the prediction made when all proteins were included in the regression equations. The regression equations are: (i) all-α proteins, $\ln(k) = -2.29\ LRO + 11.13$; (ii) all-β proteins, $\ln(k) = -7.46\ LRO + 8.51$; (iii) mixed-class proteins, $\ln(k) = -3.74\ LRO + 16.42$. The minimum limit of residue separation for computing *LRO* is 27 ($|i - j| > 27$ in equation (1)) for all-α proteins, 44 for all-β proteins and ten for mixed-class proteins.

The jack-knife test in the prediction made when the protein was excluded in the regression equations.

DG, Debe & Goddard.[4]

LRO performs extremely well ($r = -0.72$ for all-α, $r = -0.92$ for all-β and $r = -0.86$ for mixed-class proteins) in determining the folding rate of two-state proteins belonging to all structural classes (Figure 2(a)-(c)). Although the direct comparison of correlation coefficients obtained for *CO* and *LRO* with the folding rate of proteins is not appropriate, the empirical relationships derived for different structural classes predict the folding rates with greater accuracy.

Prediction of folding rate based on long-range order

We have made an attempt to predict the folding rate of two-state proteins using the concept of long-range order. We set up regression equations for all-α, all-β and mixed class proteins by relating the folding rate and long-range order obtained with the minimum distance of separation of 27, 44 and ten residues, respectively. A back-check test was carried out to verify the self-consistency of the analysis; it entails calculating coefficients of multiple regression using all of the proteins and computing their folding rates by re-substituting the values. The calculated *LRO* and predicted ln(*k*) values are presented in Table 2. We found an excellent agreement between the predicted folding rates and experimental observations as seen in Figure 3(a). The average deviation for the set of 23 proteins is 1.13.

We have also performed the jack-knife test to examine the validity of the present method and the results are also included in Table 2. This test validates the present method by determining the coefficients of multiple regression using ($n - 1$) data (i.e. omitting one protein at a time) and then computing the folding rate of the omitted protein. We found that more than 50 % of the considered proteins agreed very well with the experiment (Figure 3(b)) and the deviation is less than one unit. Only five of the 23 proteins deviate significantly from the experiment and the average deviation for all the 23 proteins is 1.782.

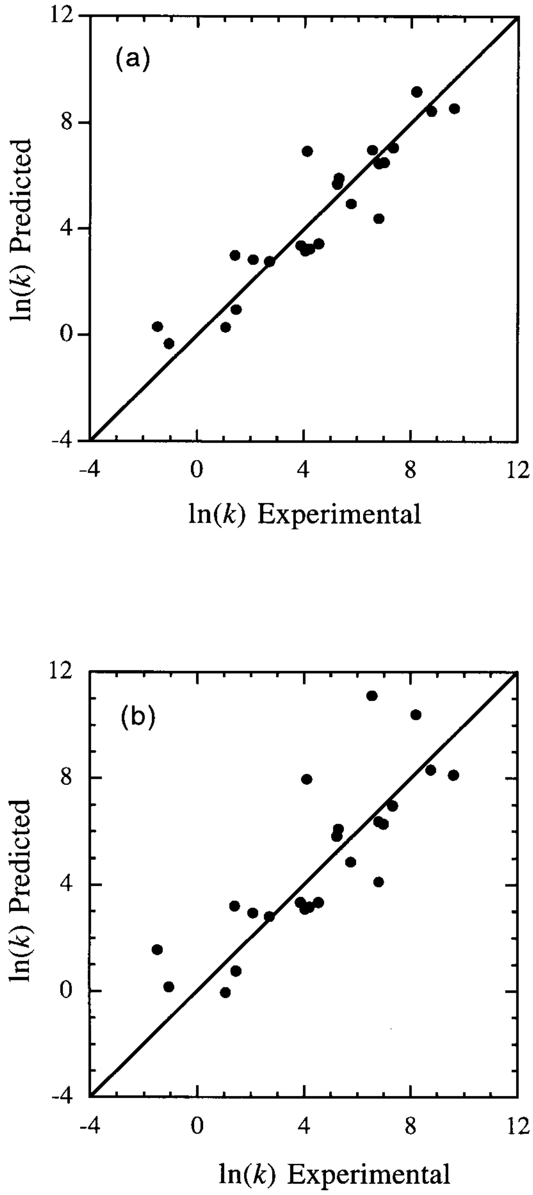

Figure 3. Scatter plot of the experimental and predicted folding rates in 23 small, two-state proteins. (a) Back-check prediction; (b) jack-knife test.

Comparison with other methods

Debe & Goddard[4] proposed a method to predict the folding rate of proteins based on the first principles of protein folding and reported that the average deviation is 3.7 for a set of 22 proteins. The present method using the *LRO* parameter predicted the folding rate of 23 two-state proteins with great accuracy and the deviations are only 1.13 and 1.78, respectively, for the back-check and jack-knife predictions (Table 2 and Figure 3). Further inspection of the data set revealed that 11 proteins have been used in both these methods for predictions. The deviation of predicted ln(*k*) values from the experimental folding rates for these 11 proteins by the method of Debe & Goddard[4] are given in Table 2, and we observed that the deviation is 3.20. For the same set of proteins, our method based on *LRO* predicted the folding rates within the deviation of 1.16 and 2.22 by back-check and jack-knife predictions, respectively.

Munoz & Eaton[5] proposed a simple statistical method to calculate the folding rates for 22 proteins using the parameter, contact order and obtained a correlation of 0.83 between theory and experiment. The present method shows the correlation of 0.92 and 0.83 between experimental and computed folding rates for the back-check and jack-knife tests, respectively. These comparisons reveal the superior performance of the *LRO* method for predicting the folding rate of proteins.

Summarizing, the concept of long-range order has been proposed and it predicts the folding rate of proteins successfully. This study reveals the necessity of protein structural classification to understand the folding mechanism of two-state proteins and the parameter *LRO* can be used for theoretical analysis. We have tested the significance of *LRO* in predicting the folding rate of proteins and observed an excellent agreement with experimental data, and the performance is comparable to other methods in the literature.

After submission of this manuscript for publication, D. E. Makarov *et al*. (unpublished) reported that the relative folding rates of simple, Gaussian models of proteins are controlled by the number of long-range native contacts. It is very interesting that both these observations demonstrate the empirical relationship between the number of contacts in native structures and the folding rate of two-state proteins.

Acknowledgements

We thank Dr K. W. Plaxco for constructive comments regarding this manuscript.

References

1. Eaton, W. A., Munoz, V., Hagen, S. J., Jas, G. S., Lapidus, L. J., Henry, E. R. & Hofrichter, J. (2000). Fast kinetics and mechanisms in protein folding. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 327-359.
2. Plaxco, K. W., Simons, K. T. & Baker, D. (1998). Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **277**, 985-994.
3. Fersht, A. R. (2000). Transition-state structure as a unifying basis in protein-folding mechanisms: contact order, chain topology, stability, and the extended nucleus mechanism. *Proc. Natl Acad. Sci. USA,* **97**, 1525-1529.
4. Debe, D. A. & Goddard, W. A. (1999). First principles prediction of protein folding rates. *J Mol Biol.* **294**, 619-625.
5. Munoz, V. & Eaton, W. A. (1999). A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl Acad. Sci. USA,* **96**, 11311-11316.
6. Miyazawa, S. & Jernigan, R. L. (1999). An empirical energy potential with a reference state for protein fold and sequence recognition. *Proteins: Struct. Funct. Genet.* **36**, 357-369.
7. Gromiha, M. M. & Selvaraj, S. (1999). Importance of long-range interactions in protein folding. *Biophys. Chem.* **77**, 49-68.
8. Gugolya, Z., Dosztanyi, Z. & Simon, I. (1997). Inter-residue interactions in protein classes. *Proteins: Struct. Funct. Genet.* **27**, 360-366.
9. Gromiha, M. M. & Selvaraj, S. (2000). Inter-residue interactions in the structure, folding and stability of proteins. *Rec. Res. Dev. Biophys. Chem.* **1**, 1-14.
10. Itzhaki, L. S., Otzen, D. E. & Fersht, A. R. (1995). The structure of the transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
11. Grantcharova, V. P., Riddle, D. S., Santiago, J. V. & Baker, D. (1998). Important role of hydrogen bonds in the structurally polarized transition state for folding of the src SH3 domain. *Nature Struct. Biol.* **5**, 714-720.
12. Manavalan, P. & Ponnuswamy, P. K. (1978). Hydrophobic character of amino acid residues in globular proteins. *Nature,* **275**, 673-674.
13. Gromiha, M. M., Oobatake, M., Kono, H., Uedaira, H. & Sarai, A. (1999). Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng.* **12**, 549-555.
14. Serrano, L., Matouschek, A. & Fersht, A. R. (1992). The folding of an enzyme. III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure. *J. Mol. Biol.* **224**, 805-818.
15. Lopez-Hernandez, E. & Serrano, L. (1996). Structure of the transition state for folding of the 129 aa protein CheY resembles that of a smaller protein, CI-2. *Fold Des.* **1**, 43-55.
16. Gromiha, M. M. & Selvaraj, S. (1998). Protein secondary structure prediction in different structural classes. *Protein Eng.* **11**, 249-251.
17. Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**, R81-R91.

Edited by P. E. Wright

(Received 20 March 2001; received in revised form 7 May 2001; accepted 15 May 2001)