doi:10.1016/j.jmb.2007.04.051 J. Mol. Biol. (2007) 370, 439–448

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

COMMUNICATION

What Determines Protein Folding Type? An Investigation of Intrinsic Structural Properties and its Implications for Understanding Folding Mechanisms

Bin-Guang Ma[1,2]*, Ling-Ling Chen[1] and Hong-Yu Zhang[1]*

[1]*Shandong Provincial Research Center for Bioinformatic Engineering and Technique Center for Advanced Study Shandong University of Technology, Zibo 255049 P. R. China*

[2]*College of Chemistry and Chemical Engineering Suzhou University Suzhou 215006, P. R. China*

**Corresponding authors*

Protein folding experiments demonstrate that the folding behaviors of many proteins can be roughly classified into two types: two-state kinetics and multi-state kinetics. Although the two types of protein folding kinetics have been observed for a long time, what determines the folding type of a protein is still largely unclear. The present work performed a comparative study based on a dataset of 43 two-state and 42 multi-state folders at different levels of proteins' intrinsic properties from the simplest sequence length to native structure topology. The results show that protein's amino acids composition and the long-range interaction-based topological complexity rather than secondary structure contents are the major determinants of protein folding type. Furthermore, a sequence-based folding type prediction achieved an accuracy of more than 80%. These findings implicate that there is no clear boundary between secondary and tertiary structure formation during the protein folding process and support the existence of a continuum of folding mechanism between the two ends of hierarchic and nucleation folding scenarios.

Keywords: folding intermediate; amino acid composition; topological complexity; folding mechanisms; folding type prediction

Two-state and multi-state folding kinetics

Protein folding is the process by which a protein transforms from its denatured state to its specific biologically active conformation. Revealing the mechanism of protein folding is a great challenge in molecular biology. It has been observed that the folding behaviors of proteins can be roughly classified into two types: two-state kinetics and multi-state kinetics.[1–5] The two-state folders have no visible intermediates in the course of folding, which therefore occurs as an "all-or-none" process under some experimental conditions,[1,3] while the proteins with multi-state folding kinetics fold *via* intermediates, which accumulate during the early stages of folding and show a stepwise assembly procedure.[1,6–9]

Although the two types of protein folding kinetics have been observed for a long time, what determines the folding type of a protein is still largely unclear. It has been reported that the folding behaviors are affected by the experimental conditions such as salt concentration or temperature,[1,10] while in the same environment, the main determinant information should be, in principle, reflected in the protein's intrinsic property.[11] There is evidence showing that point mutations on some crucial positions can switch the folding of some proteins from two-state to multi-state[8] or *vice versa*.[12] A recent study showed that the native topology affects the appearance of intermediates during the folding of proteins with a flavodoxin-like fold.[13] Our recent work also found the difference of folding-rate-determining amino acids between the two types of folders.[14] But untill now, there is no systematical investigation on the intrinsic determi-

Abbreviations used: Sn, sensitivity; Sp, specificity; Ac, accuracy; CO, contact order; CC, clustering coefficient; TCD, total contact distance; Abs_CO, absolute contact order; CTP, chain topological parameter; LRO, long range order; F_{LOCAL}, fraction of local contacts; *LR_CO*, long range contact order.

E-mail addresses of the corresponding authors: bgMa@sdut.edu.cn; zhanghy@sdut.edu.cn

nant factors of protein folding type, i.e. what determines whether or not the intermediate appears. Thanks to the efforts of experimental scientists, more and more data on folding kinetics become available, which provide unprecedented opportunities for exploring the determinant factors of folding types.

Here, we attempt to address the following question: what determines protein folding type? A comparative study was performed for a dataset of 43 two-state and 42 multi-state folders at different levels of protein structural properties from the simplest sequence length to native structure topology. The results show that protein's amino acid composition and the long-range interaction-based topological complexity are the major determinants for protein folding type.

Dataset and method

The dataset used in this study was compiled from recent literature, with particular reference to Kamagata *et al*.[4] and Ivankov & Finkelstein.[5] It includes 85 folders with known 3D structures (see Supplementary Data, for a detailed list); 43 of them are two-state folders and the others are multi-state folders. Frequency counting of amino acids and the calculation of topological parameters are implemented by python scripts based on the Bio.PDB module in the Biopython software package.[15]

For the evaluation of a classifier, the commonly adopted indexes used in the evaluation of gene identification algorithms were employed.[16] To use these indexes, we need two sets of samples: positive and negative. It is all right that we assign two-state folders as a positive sample and multi-state folders as a negative sample or *vice versa.* Here we take the group with larger average parameter values as the positive sample and the other group as negative. The efficiency of a classifier can be expressed as sensitivity (Sn) and specificity (Sp), where Sn represents the proportion of positive samples that have been correctly recognized as positive and Sp represents the proportion of negative samples that have been correctly recognized as negative (for the details of the definition of Sn and Sp, see Burset & Guigo[16]); and then the accuracy (Ac) of a prediction is simply defined as $Ac=(Sn+Sp)/2$. By letting Sn approximately equal Sp, a proper threshold of a parameter can be determined, and then the folding type of a protein is predicted as multi-state (or two-state) if it has a parameter value larger (or less) than the threshold.

Table 1. Average amino acid contents of two-state and multi-state folders

	Two-state folders	Multi-state folders	*p*-value
A	0.0749	0.0822	0.3862
C	**0.0050**	**0.0110**	**0.0918**
D	0.0535	0.0563	0.5799
E	0.0792	0.0785	0.9003
F	**0.0436**	**0.0361**	**0.0885**
G	**0.0867**	**0.0754**	**0.0908**
H	**0.0157**	**0.0217**	**0.0907**
I	0.0488	0.0564	0.1073
K	0.0919	0.0767	0.1175
L	**0.0715**	**0.0899**	**0.0035**
M	0.0192	0.0217	0.3893
N	0.0434	0.0438	0.9322
P	0.0387	0.0430	0.4913
Q	0.0449	0.0373	0.1405
R	**0.0390**	**0.0505**	**0.0805**
S	0.0606	0.0498	0.1178
T	0.0634	0.0598	0.6276
V	0.0740	0.0656	0.2204
W	0.0143	0.0154	0.6658
Y	0.0318	0.0287	0.4344
Length[a]	**86**	**139**	**5.61e-05**

A two-tailed *t*-test was performed (using statistics software R) to investigate whether or not the differences between the mean amino acid content of two-state and multi-state folders are significant. *p*-values in the fourth column indicate the significance level. If the significant level is set at 0.1, the mean contents of amino acids C, F, G, H, L, R are significantly different between two-state and multi-state folders; if the significance level is set at 0.01, only L is significant (0.0715 *versus* 0.0899). Bold font indicates that the difference is significant at a certain level in the two-tailed *t*-test.

[a] The average sequence lengths of two-state and multi-state folders.

Sequence length is a major determinant for folding type but not sufficient

Protein size greatly affects the folding kinetics. It has been revealed that chain length is an important factor for the determination of protein folding rates.[3,17,18] Here, we attempt to investigate whether the sequence length affects the appearance of intermediates. As shown in Table 1, the average length of the 43 two-state folders is 86, that is significantly ($p=5.61e\text{-}05<0.0001$) shorter than that of the 42 multi-state folders (139) in a two-tailed *t*-test, indicating the sequence length is a major determinant for a protein's folding type, which consists with the common understanding that small single-domain proteins (<~100 residues) usually exhibit apparently two-state folding. However, sequence length is not sufficient to determine the folding type of a protein, i.e. long proteins do not necessarily exhibit folding behaviors of multi-state. For example, among the 43 two-state folders, the longest protein is Lyme disease variable surface antigen VlsE of *Borrelia burgdorferi* (PDB code: 1L8W)[19] with a length of 341, which is far longer than the average sequence length of multi-state folders; while among the 42 multi-state folders, the shortest protein is leech carboxypeptidase inhibitor (PDB code: 1DTV)[20] with a length of 67, which is shorter than the average sequence length of two-state folders. Therefore, sequence length is not a sufficient indicator for a protein's folding type.

Amino acid content difference between two-state and multi-state folders

Amino acid content is one of the most basic properties of a protein sequence. As shown in Table 1, six amino acids have significant difference between the two types of folders on a relatively significant level of $p<0.1$, with F and G rich in two-state folders and C, H, L, R rich in multi-state folders. If viewed on a significant level of $p<0.01$, there is only L significant, rich in multi-state folders.

According to Chou–Fasman[21] statistics of amino acid composition for secondary structure prediction, the propensities (normalized frequencies) for α-helix, β-sheet and β-turn of the six relatively significant ($p<0.1$) amino acids are listed in Table 2. It can be seen that the two amino acids F and G rich in two-state folders have the largest propensity for β-sheet and β-turn, respectively, and that among the four amino acids rich in multi-state folders, C and L have the largest propensity for β-sheet while H and R have the largest propensity for α-helix and β-turn, respectively. What should be noticed is that, among these six significant amino acids, five of them have the largest propensities either for β-sheet or for β-turn and only one residue (H) has the most propensity for α -helix, which indicates that the composition difference between two-state and multi-state folders is mostly embodied by the residues who participate in relatively long range interactions as in β-sheet or β-turn rather than local interactions as in α -helix.

Moreover, for the amino acid L, which is different at a significant level of $p<0.01$ and rich in multi-state folders, it has the most preference for β-sheet, which may imply that the stabilization of intermediates in multi-state folding not only relies on the formation of helix[22] but also on an important contribution from β-sheet.[23]

Secondary structure content difference between two-state and multi-state folders

Secondary structures are defined as the backbone conformations of successive residues in protein sequence.[24] Secondary structure contents, especially the content of helix, are believed to reflect the strength of local interactions. To investigate the importance of local interactions in the determination of folding behavior, the secondary structures of the two types of folders are assigned by the DSSP program[24] and the average secondary structure contents over all the 43 two-state and 42 multi-state folders are calculated and listed in Table 3. From the *p*-values of *t*-test, it can be seen that there is no significant difference between them, although the helix content in multi-state folders is slightly larger than that in two-state folders, which means that local interactions between amino acids are not a determinant factor of protein folding type, at least for the totality of the proteins used in this study.

Table 2. The secondary structure propensities[a] for the amino acids whose contents are significantly different ($p<0.1$) between two-state and multi-state folders[b]

	C	F	G	H	L	R
Helix	0.70	1.13	0.57	1.00	1.21	0.98
Sheet	1.19	1.38	0.75	0.87	1.30	0.93
Turn	0.96	0.66	1.56	0.95	0.50	1.01

[a] The secondary structure propensity for α-helix, β-sheet and β-turn are extracted from the AAindex database.[57]
[b] The columns with a light gray background indicate amino acids whose contents are larger in multi-state folders than in two-state folders.

Table 3. Average secondary structure contents of 43 two-state and 42 multi-state folders

	Two-state folders	Multi-state folders	*p*-value[a]
Helix	0.2577	0.3224	0.1926
Sheet	0.2914	0.2554	0.3910
Loop	0.4509	0.4223	0.2413

Secondary structures are assigned from the PDB[58] coordinates by using the program DSSP[24] [http://www.sander.ebi.ac.uk/dssp]. DSSP defines seven types of secondary structure: H, α-helix; B, residue in isolated β-bridge; E, extended strand, participates in β ladder; G, 3-helix (3/10 helix); I, 5-helix (Pi helix); T, hydrogen bonded turn; S, bend. Here, H and G are assigned as helix; E and B are assigned as sheet; and others are assigned as loop. This assignment is consistent with the Q3 measure for the accuracy of secondary structure prediction.[59]
[a] The *p*-values of the two-tailed *t*-test performed in comparison of the difference between two-state and multi-state folders.

The difference of topological complexity of tertiary structures between two-state and multi-state folders measured with a variety of parameters

A variety of topological parameters were calculated and compared for the two types of folders. The results are listed in Table 4. Among these parameters (see Supplementary Data for their definitions), relative contact order (CO) is primary one, which has the meaning of the average sequence separation between all pairs of contacting residues per contact per residue and reflects topological complexity of a protein's tertiary structure.[25] From Table 4, it can be found that the average CO value of two-state folders is significantly ($p=4.175e-07$) larger than that of multi-state folders, indicating that the topology of two-state folders are more complex than those of multi-state folders. A later born sibling of CO is absolute contact order (Abs_CO) where the influence from the size of a protein on folding rate is considered. The average Abs_CO is also calculated and listed in Table 4, from

Table 4. The averages of a variety of topological indicators over the two-state and multi-state folders used in this study

Indicator[a]	Two-state folders	Multi-state folders	*p*-value[b]
CO	0.2422	0.1872	4.175e-07
Abs_CO	19.6160	23.8763	0.0124
TCD	1.2961	1.0414	0.0007
CTP	8.7151	9.7163	0.2677
LRO	1.3727	1.3780	0.9630
F_{LOCAL}	0.6542	0.6766	0.3997
CC	0.6445	0.6200	5.252e-05
LR_CO	0.4746	0.3664	3.568e-07

[a] The topological indicators are calculated according to the cutoff values of sequence separation (L_{cut}) and/or distance definition (R_{cut}) reported in their original publications. CO and Abs_CO are calculated with $R_{cut}=6$ (Å);[18,25] LRO is calculated with $L_{cut}=12$ (residue) and $R_{cut}=8$ (Å);[28] F_{LOCAL} is calculated with $L_{cut}=4$ (residue) and $R_{cut}=7$ (Å);[29,30] TCD is calculated with $L_{cut}=2$ (residue) and $R_{cut}=6$ (Å);[26] CTP is calculated with $R_{cut}=4$ (Å);[27] CC is calculated with $R_{cut}=4.5$ (Å);[31] *LR_CO* is calculated with $L_{cut}=12$ (residue) and $R_{cut}=8$ (Å).
[b] The *p*-values of the two-tailed *t*-test performed in comparison of the difference between two-state and multi-state folders.

which it can be found that the Abs_CO value of multi-state folders is larger than that of two-state folders although the p-value (= 0.0124) is far larger than that of CO but still significant.

Two parameters derived from CO are total contact distance[26] (TCD) and chain topological parameter[27] (CTP). TCD differs from CO in its pre-factor where the contact number has been replaced by the sequence length, i.e. the sequence length has been squared in TCD. TCD also shows a significant ($p = 0.0007$) difference between the two types of folders although it does not pronounce as significant as CO. The difference between CTP and CO is that the sequence separation has been squared in CTP. The average CTP shows no significant ($p = 0.2677$) difference between the two groups of folders.

Another two parameters that have been proposed inspired by CO are long range order[28] (LRO) and fraction of local contacts[29,30] (F_{LOCAL}). LRO is defined as the total number of long range contacts (sequence separation longer than a cutoff value) normalized by the sequence length. F_{LOCAL} is the ratio of the number of local contacts to the total number of contacts. Both of them exhibit no significant difference (Table 4) between the two groups of folders. These results are consistent with the previous work by Kamagata *et al.* where no significant difference of the number of sequence-distance native pairs is found between the two types of folders,[4] which may suggest that merely the number (or fraction) of long range (or local) contacting pairs is not a good indicator for characterizing the topological difference between the two folding types.

Another parameter defined from another viewpoint is the clustering coefficient (CC), which is also called cliquishness, resulting from a network analysis of protein structure.[31] As CO, CC also has an intuitive meaning: it provides a measure of the extent to which different residues interacting with residue i are also interacting with each other. CC also has a significant difference (Table 4) between the two groups of folders, with the larger value in two-state folders. CC is believed to be able to reflect the cooperativity of the folding process. The larger CC of two-state folders indicates a higher cooperativity in two-state folding than in multi-state folding.

The best classifier for folding type based on the topology of protein structure

As mentioned above, the average CO value of two-state folders is significantly larger than that of multi-state folders, indicating that the topology of two-state folders is more complex than multi-state folders. However, when measured by Abs_CO, the topology of multi-state folders seems more complex than two-state folders. Here arises a question: which one should be used to measure the topology of a protein structure in terms of classification of the two types of folders? To address this problem, the aforementioned parameters were compared as classifier for the two groups of folders by seeking a proper threshold and the results are listed in Table 5. It can be seen that among these parameters, CO achieves the best classification accuracy, with Sn, Sp and Ac to be 72.09%, 71.43% and 71.76%, respectively. All the other parameters get lower accuracy than CO except for TCD, which achieves as good an accuracy as CO. Abs_CO gets a much lower classification accuracy due to its exclusion of the sequence length in its definition.

Moreover, a modified version of contact order, called long range contact order (*LR_CO*), can improve the classification accuracy further. *LR_CO* adopts the formulism of CO[25] whereas contacts are defined according to LRO,[28] that is:

$$LR\_CO = \frac{1}{L \cdot N} \sum^{N} \Delta S_{i,j} \qquad (1)$$

where N is the total number of contacts, $\Delta S_{i,j}$ is the sequence separation , in residues, between contacting residues i and j, and L is the total number of residues in the protein.[25] Slightly different from contact order, here contacts are defined as two residues with the C^{α}–C^{α} distance less than a cutoff distance R_{cut} and separated by at least a cutoff value of residue separation L_{cut}. This definition of contacts is the same to that of long range order.[28] *LR_CO* also shows a significant difference between the two types of folders (Table 4) and achieves the best classification efficiency with accuracy to be 74.11% (Table 5).

The difference of *LR_CO* between two-state and multi-state folders increases with the increase of the values of sequence separation and distance definition

Now we concentrate on the best classifier *LR_CO*. As a combination of CO and LRO, *LR_CO* is of the same meaning to CO but with more power to distinguish short and long-range interactions. The adjustable parameter L_{cut} in *LR_CO* allows us to investigate the topological complexity of protein structures with respect to the variation of the sequence separation between two contacting residues. This feature makes *LR_CO* more effectively characterize the relative importance of short and

Table 5. The threshold, sensitivity, specificity and accuracy of different topological parameters as classifier for two-state and multi-state folders

	CO	Abs_CO	TCD	CTP	LRO	F_{LOCAL}	CC	*LR_CO*
Threshold	0.2137	20.8980	1.0935	8.5328	1.4379	0.6537	0.6271	0.4313
Sensitivity	0.7209	0.6190	0.7209	0.5238	0.4762	0.5476	0.6977	0.7442
Specificity	0.7143	0.6279	0.7143	0.5116	0.4651	0.5581	0.6905	0.7381
Accuracy	0.7176	0.6235	0.7176	0.5177	0.4707	0.5529	0.6941	0.7411

long-range interactions in a protein structure. As shown in Figure 1, there is a statistically significant ($p<0.0001$) difference between *LR_CO* values of the two-state and multi-state folders. It is observed that the *LR_CO* of the two-state folders is significantly larger than that of the multi-state folders with $R_{cut}=8$ Å (as in the definition of long range order[28]) and L_{cut} varying from 1 to 20 residues, indicating that the topology of two-state folders are more complex than that of multi-state folders; and furthermore, the difference of *LR_CO* values between the two types of folders increases monotonically with the increase of the cutoff value of sequence separation (L_{cut}), which suggests that the topological complexity is mainly attributed to the long range contacts.

Similarly, we can fix the L_{cut} value to be the minimum possible value, i.e. let $L_{cut}=1$, and examine how the topological complexity depends on the variation of distance definition (R_{cut}) in 3D space. As shown in Figure 2, the topological complexity (indicated by *LR_CO* values) of two-state folders is significantly ($p<0.001$) larger than that of multi-state folders in a wide range of R_{cut} definition from 4 Å to 14 Å, and the difference increases monotonically with the increase of R_{cut} value, suggesting that the topological complexity mainly originates from far distance contacts in tertiary structures. The two figures together inform us that two-state folders are more complex than multi-state folders in tertiary topology and this complexity

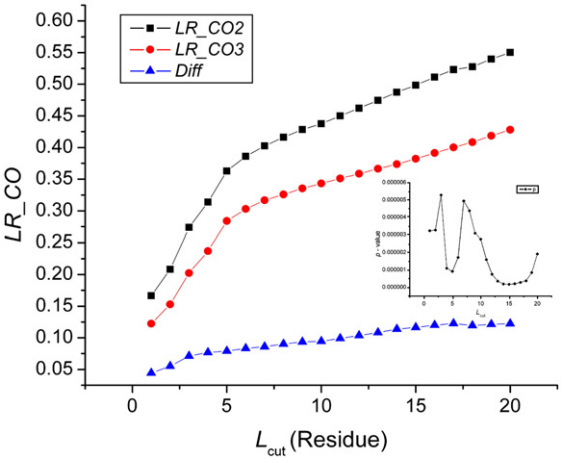

Figure 1. The averaged long range contact order (*LR_CO*) as a function of the variation of the cutoff value of sequence separation (L_{cut}). *LR_CO* values are calculated with $R_{cut}=8$ (Å) and L_{cut} varying from 1 residue to 20 residues. Squares represent the average *LR_CO* values of two-state folders; circles represent the average *LR_CO* values of multi-state folders; triangles represent the difference between the averaged *LR_CO* values of two-state and multi-state folders. The inset represents the *p*-values of two-tailed *t*-tests in the comparison of the *LR_CO* values of the two types of folders. All the *p*–values for L_{cut} from 1 to 20 residues are less than 0.00001, indicating the topological complexity of two-state folders is significantly larger than that of multi-state folders in a wide range of L_{cut} definitions.

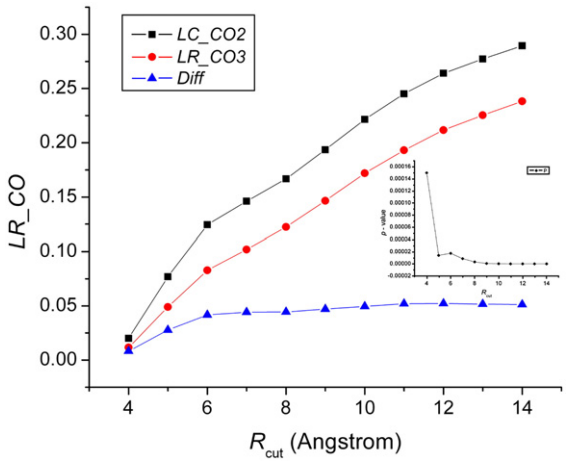

Figure 2. The averaged long range contact order (*LR_CO*) as a function of the variation of the cutoff value of C^{α}–C^{α} distance. *LR_CO* values were calculated with $L_{cut}=1$ (residue) and R_{cut} varying from 4 to 14 (Å). Squares represent the average *LR_CO* values of two-state folders; circles represent the average *LR_CO* values of multi-state folders; triangles represent the difference between the averaged *LR_CO* values of two-state and multi-state folders. All the *p*-values of the two-tailed *t*-test are less than 0.001 (inset), indicating that the topological complexity of two-state folders are significantly larger than that of multi-state folders in a wide range of R_{cut} definitions.

mainly comes from long-range (in primary sequence) and far-distance (in tertiary structure) contacts. Therefore, long-range interaction-based topological complexity is a determinant for the folding type of a protein.

Composition-based folding type prediction

Considering the significant difference of sequence length between the two types of folders, the folding type of a protein may be predicted by its length. On the present dataset, the proper threshold of sequence length is 95 where the Sn (69.05%) and Sp (69.77%) of the prediction are approximately equal to each other and the accuracy is 69.41% (Table 6). However, when the difference of residue composition between the two types of folders is taken into account, the prediction accuracy can be improved further.

Based on the difference of amino acid content between the two types of folders, coupled with the difference between the average chain length, we can define a folding type predictor which has two components: the sequence length L and the content sum (*Csum*) of the amino acids significantly ($p<0.1$) rich in multi-state folders, i.e.

$$Csum = \frac{1}{L}(N_C + N_H + N_L + N_R) \qquad (2)$$

where N_C, N_H, N_L and N_R are the occurrence numbers of amino acid cysteine, histidine, leucine and arginine in the protein sequence, respectively; and L is the sequence length.

Table 6. The sensitivity, specificity and accuracy for the folding type prediction of two-state and multi-state folders by the "length-threshold" approach and the composition-based predictor (Cp)

Method	Sensitivity	Specificity	Accuracy
Length threshold[a]	0.6905	0.6977	0.6941
Back-check (Cp)[b]	0.8095	0.8140	0.8117
Jack-knife (Cp)[c]	0.7970	0.8200	0.8085

[a] False negative, 1DTV (67), 1QQV (67), 2CRO (71), 1UZC (71),1UBQ (76), 2ABD (86), 1CEI (87), 1TIT (89), 1BRS (89), 1BTA (89), 1FNF (90), 1GXT (91), 1TTG (94); false positive, 1APS (98), 2ACY (98), 1RIS (101), 1URN (102), 1HRC (104), 256B (106), 1D6O (107), 1FKB (107), 1YCC (108), 2VIK (126), 1AZU (128), 1LOP (164), 1L8W (341). The number in parenthesis is the length of the corresponding protein sequence. The threshold of length that best classifies the two groups of folders is 95.

[b] False negative, 1ADW (−0.0057), 1BNI (−0.0045), 1CEI (−0.0117), 1EAL (−0.0119), 1IFC (−0.0040), 1TTG (−0.0182), 1UZC (−0.0256), 2ABD (−0.0205); false positive, 1AZU (0.0001), 1D6O (0.0006), 1ENH (0.0139), 1FKB (0.0006), 1FNF (0.0107), 1L8W (0.0284), 1RIS (0.0197), 1YCC (0.0028). The number in parenthesis is the Cp value of the corresponding protein.

[c] The procedure of the jack-knife test is as follows: in each round, one protein is omitted and the (n−1) proteins remained are used to derive Cp, and then prediction is made on the omitted one. Totally, 1000 rounds are performed and the listed Sensitivity, Specificity and Accuracy are averages.

Then the composition-based folding type predictor (Cp) is defined as follows:

$$\mathrm{Cp} = a \times Length + b \times Csum + c \quad (3)$$

where a, b, c are three parameters to be estimated based on a protein dataset with known folding types.

Based on the proteins used in this study, Cp can be determined as:

$$\mathrm{Cp} = 0.000199 \times Length + 0.257186 \times Csum - 0.061498 \quad (4)$$

With this predictor, predicting the folding type of a protein is simply to see the value (sign) of Cp: if Cp>0, then the folder is multi-state; otherwise, the folder is two-state.

The prediction accuracy of Cp for the two folding types on the present dataset is listed in Table 6. As what is shown, the Sn is 80.95% and the Sp is 81.40%. As a result, the accuracy of back-check prediction is improved from 69.41% (predicted only by length) to 81.17% (predicted by Cp), with an increase of more than 10%. The proteins, which constitute the false positive and the false negative of the prediction, are listed in the notation of Table 6. Figure 3 shows the distribution of these folders in the prediction by Cp. A jack-knife test was performed as a leave-one-out procedure. From Table 6, it can be seen that an almost equally good prediction is achieved in the jack-knife test, showing the robustness of this approach.

Classification accuracy can be further improved by combining Cp and *LR_CO*

Since both sequence composition and topological complexity are of significant difference between the two types of folders, which are characterized by Cp and *LR_CO*, respectively, it can be reasonably expected that a combination of the two parameters may further increase the classification accuracy. To test this expectation, a complex indicator of protein folding type is defined as:

$$\mathrm{I_{FT}} = a \times Length + b \times Csum + c \times LR\_CO + d \quad (4)$$

which includes the contribution of *LR_CO*.

On the present dataset, the coefficients of $\mathrm{I_{FT}}$ can be determined:

$$\mathrm{I_{FT}} = 3.4986\text{e-}005 \times Length + 0.300146 \times Csum - 0.169505 \times LR\_CO + 0.021659 \quad (5)$$

Likewise, the folder with $\mathrm{I_{FT}}>0$ is identified as multi-state, otherwise as two-state.

By using $\mathrm{I_{FT}}$, the classification results on the present dataset are as follows: Sn=0.8571, Sp=0.8372 and Ac=0.8472, with the false negative to be 1CEI (−0.0442), 1EAL (−0.0074), 1UBQ (−0.0044), 1UZC (−0.0294), 2ABD (−0.0319), 2CRO (−0.0063), and the false positive to be 1AZU (0.0008), 1ENH (0.0128), 1FNF (0.0165), 1L8W (0.0105), 1RIS (0.0097), 256B (0.0132), 2VIK (0.0060), respectively. The numbers in parentheses are the corresponding $\mathrm{I_{FT}}$ values.

Compared with the results obtained only by using Cp (Ac=81.17%) or *LR_CO* (Ac=74.11%), the classification accuracy by $\mathrm{I_{FT}}$ (Ac=84.72%) is indeed higher. The false negative and the false positive are reduced by two and one proteins, respectively, in comparison with the prediction result by Cp. The increase of the classification accuracy suggests that not only do there exist significant differences between the two folding types, both on primary sequence composition and on tertiary topology complexity, but also the differences on the two levels seem to be independent (to some extent) with each other, although the amino acids with significantly different contents between the two types of folders tend to take part in long-range interactions.

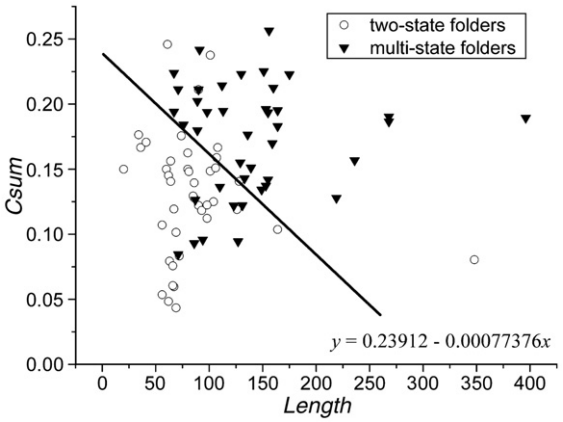

Figure 3. Folding type prediction of two-state and multi-state folders. Circles represent two-state folders; triangles represent multi-state folders. The continuous line is the linear boundary with equation to be $0.000199x + 0.257186y - 0.061498 = 0$ or $y = 0.23912 - 0.00077376x$. The prediction accuracy is 81.17%.

Implications for understanding folding mechanisms

Anfinsen's pioneer work on protein denaturing and refolding demonstrates that protein chains are capable of self-organization, i.e. they can form their native structures spontaneously in an appropriate environment,[32] which leads to a thermodynamic hypothesis that the native state has the lowest free energy.[11] Since then, deciphering the underlying mechanisms that govern protein folding becomes a main subject of molecular biology. In 1968, Levinthal raised an influential argument[33] that came to be known as the Levinthal paradox: an unfolded protein molecule needs "an eternity" to search zillions of possible conformations, yet it reaches the single native state rapidly.[34] Two major scenarios on the solution of this paradox have been proposed owing to theoreticians' efforts: the hierarchical model and nucleation mechanism of protein folding.[3,8,29]

The hierarchic model of protein folding (also known as framework[35,36] or diffusion-collision model[37,38]) postulates that the folding process starts from the formation of local secondary structures, which serves as preformed units for subsequent stages of folding, i.e. the folding proceeds in a stepwise manner.[2,39] Hierarchic folding usually populates on-pathway intermediates although they are not always detectable. This folding scenario has been exemplified by many proteins such as barnase protein,[40] colicin E7 immunity protein[41] and the chemotactic protein CheY,[42] etc. In comparison, nucleation mechanism (here means nucleation-condensation[43,44] not classical nucleation[45,46]) suggests that local secondary elements and global tertiary topologies form simultaneously, i.e. the folding proceeds in a parallel manner. Nucleation-condensation folding usually follows single-exponential kinetics and does not populate intermediates. This folding scenario has also been instantiated by proteins chymotrypsin inhibitor 2[47] (CI2), cyclophilin A[48] and *Borrelia burgdorferi* VlsE,[49] etc. Besides these experimental facts, the above two theoretical folding mechanisms have also been demonstrated in computer simulations.[50–52] Since the two folding scenarios have been manifested in experiment and simulation, now comes a question: which one answers for the folding processes of real proteins? Or alternatively, can one simple model describe the totality of the protein folding processes?

The present results may provide new clues to answering the above questions. According to the results of this study, there are statistically significant differences between amino acid composition and structure topological complexity of two-state and multi-state folders. The difference of residue composition can be utilized to predict the folding type of a protein. The accuracy of the prediction by composition-based predictor Cp is 10% higher than the case only predicted by sequence length, which, on the other hand, also proves the existence of the composition difference between the two types of proteins. The topological complexity of tertiary structure is found larger in two-state folders than in multi-state folders, when measured by CO or *LR_CO*. Someone may argue that the significant difference between the *LR_CO* values of the two types of folders is merely a reflection of the difference between their average lengths. However, it seems not the case. Firstly, only using sequence length, the classification accuracy (69.41%) is not as high as by *LR_CO* (74.11%). Secondly, the difference of sequence length cannot explain the monotonic increase of the difference of *LR_CO* values with the increase of the cutoff values of sequence separation (L_{cut}) and distance definition (R_{cut}), because the average sequence lengths of the two types of folders do not change with the change of these cutoff values. Thirdly, the cliquishness (CC), which is defined from another viewpoint (namely, from the network analysis of protein structure) and measures the interdependence of native contacts, also indicates that the topology of two-state folders are more complex than multi-state folders. The three lines of evidence suggest that sequence length is not the sole factor that characterizes the difference of the two types of proteins and that the topological difference also exists to some extent between the two folding types.

The existence of the difference of sequence composition and topological complexity between the two types of folders implies that one simple model may not be sufficient to describe all protein folding processes. Some proteins may incline to adopt the hierarchic folding due to its relatively simple native topology, while some others may adopt a nucleation mechanism due to a more complicated native topology. The difference of folding behavior not only relies on the environmental conditions of a protein but also roots in its intrinsic property, i.e. the difference of intrinsic properties may render some proteins to prefer (or exhibit more) two-state folding and others to multi-state folding, which exhibits the diversity and complexity of biomolecular behaviors.

On the other hand, the results of this study show that there is no statistically significant difference of secondary structure contents between two-state and multi-state folders, which implies that merely local interactions are not sufficient to determine a protein's folding type and that long-range interaction-based "cooperativity" is of particular importance in the protein folding process. This finding provides new clues to understanding the relative importance of secondary *versus* tertiary interactions in the determination of protein folding behavior and helps reconcile the controversy on folding mechanisms. As stated by Baldwin & Rose, a hierarchic folding (framework model) demands that secondary structures can be sufficiently determined by local sequence information.[2] However, the so-called "chameleon sequence" experiment by Minor & Kim provided a prominent example of the secondary structure formation depending on the context within the total protein.[53] A systematic investigation of a large dataset of non-homologous proteins recently reported by Kihara confirmed the effect of long-range interactions on the protein's secondary structure formation.[54] These findings, coupled with the results of the present analysis that long range interactions play dominative

roles in the determination of protein folding behavior, may imply that the formation of secondary structure cannot be completely independent of the formation of tertiary structure, just as the latter cannot be completely free of the former. The interdependence of secondary and tertiary structure formations blurs the boundary between them in the protein folding process and supports the existence of a continuum of protein folding mechanism between the two ideal folding scenarios of hierarchic and nucleation.[55] For the real proteins, there is the probability that one protein adopts the two folding mechanisms concurrently, which has been verified in a recent report on the folding of cellular myeloblastis protein (c-Myb) conspired by simulation and experiment.[56]

No matter a protein adopts which folding mechanism, hierarchic or nucleation or even their mixture, it can conquer the difficulty described by Levinthal paradox. Hierarchic folding *via* intermediates can effectively reduce the conformational search space due to the decrease of entropy induced by the formation of preformed secondary structures. The nucleation folding possesses multiple pathways and can confine the folding on some specific pathway on a funnel-like energy landscape. The mechanism-mixed folding process of course will benefit from the above two aspects. In summary, the interplay between secondary and tertiary structure formation dominates protein folding.

Lastly, it is necessary to point out some limitations of the present work in the understanding of the protein folding mechanism. First, the result of the present work is a statistical one, which means that it cannot be denied that some individual proteins may not keep to the statistical rules stemmed from the current study. Particularly, the composition of primary sequence as an indicator of protein folding type may be unable to account for the effect of single amino acid mutation that can switch the folding type of a protein. Second, although this work reveals some determinant factors for protein folding behavior, it gives no further explanation on why the difference in intrinsic properties leads to the different folding type, which is a great challenge for further study. Although a unified mechanism of protein folding is more favored, such a mechanism must be able to account for the significant difference of intrinsic structural property observed between the two types of folders. Third, the current findings are grounded merely on the examination of protein sequence and native structures, i.e. static properties of proteins. Uncovering the details of the protein folding process also rests on the examination of the structures of intermediates or transition states.

Concluding remarks

Summarizing the above results and discussion, we may arrive at the following concluding remarks.

(1) The amino acid compositions (indicating sequence length and amino acid contents) of two-state and multi-state folders are significantly different, which can be used for protein folding type prediction.
(2) The secondary structure contents (characterizing local interactions to some extent) of two-state and multi-state folders are not significantly different.
(3) The topological complexity of tertiary structures of two-state and multi-state folders is of statistically significant difference.
(4) Considering (1), (2) and (3) simultaneously, we may reach the conclusion that amino acid composition and the long-range interaction-based topological complexity are the major determinants of protein folding type.
(5) One simple model may be not enough for describing all protein folding processes. The differences of intrinsic properties may render proteins to prefer (or exhibit more) either two-state or multi-state folding behaviors, while a clear-cut boundary between secondary and tertiary structure formation is difficult (if not impossible) to be marked out, which supports the existence of a continuum of folding mechanism between the two ends of hierarchic and nucleation folding scenarios in the protein universe.

The limited amount of available data and the discrepant experimental conditions of data acquisition prevent us to draw a very solid conclusion. The results of this study are expected to be tested on a larger and more consistent dataset.

Acknowledgements

This work was supported by the National Basic Research Program of China (2003CB114400) and the National Natural Science Foundation of China (30600119 and 30570383). Thanks go to all those who deposited their experimental data in public databases, and to those who maintain these databases.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2007.04.051

References

1. Jackson, S. E. (1998). How do small single-domain proteins fold? *Fold. Des.* **3**, R81–R91.
2. Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? II. Folding intermediates and transition states. *Trends Biochem. Sci.* **24**, 77–83.
3. Finkelshtein, A. V. & Galzitskaya, O. V. (2004). Physics of protein folding. *Phys. Life Rev.* **1**, 23–56.
4. Kamagata, K., Arai, M. & Kuwajima, K. (2004). Unification of the folding mechanisms of non-two-state and two-state proteins. *J. Mol. Biol.* **339**, 951–965.

5. Ivankov, D. N. & Finkelstein, A. V. (2004). Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl Acad. Sci. USA*, **101**, 8942–8944.
6. Maity, H., Maity, M., Krishna, M. M., Mayne, L. & Englander, S. W. (2005). Protein folding: the stepwise assembly of foldon units. *Proc. Natl Acad. Sci. USA*, **102**, 4741–4746.
7. 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.
8. 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.
9. Feng, H., Zhou, Z. & Bai, Y. (2005). A protein folding pathway with multiple folding intermediates at atomic resolution. *Proc. Natl Acad. Sci. USA*, **102**, 5026–5031.
10. Maxwell, K. L., Wildes, D., Zarrine-Afsar, A., De Los Rios, M. A., Brown, A. G., Friel, C. T. *et al.* (2005). Protein folding: defining a "standard" set of experimental conditions and a preliminary kinetic data set of two-state proteins. *Protein Sci.* **14**, 602–616.
11. Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223–230.
12. Viguera, A. R. & Serrano, L. (2003). Hydrogen-exchange stability analysis of Bergerac-Src homology 3 variants allows the characterization of a folding intermediate in equilibrium. *Proc. Natl Acad. Sci. USA*, **100**, 5730–5735.
13. Bollen, Y. J. & van Mierlo, C. P. (2005). Protein topology affects the appearance of intermediates during the folding of proteins with a flavodoxin-like fold. *Biophys. Chem.* **114**, 181–189.
14. Ma, B. G., Guo, J. X. & Zhang, H. Y. (2006). Direct correlation between proteins' folding rates and their amino acid compositions: an ab initio folding rate prediction. *Proteins: Struct. Funct. Genet.* **65**, 362–372.
15. Hamelryck, T. & Manderick, B. (2003). PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
16. Burset, M. & Guigo, R. (1996). Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
17. Galzitskaya, O. V., Garbuzynskiy, S. O., Ivankov, D. N. & Finkelstein, A. V. (2003). Chain length is the main determinant of the folding rate for proteins with three-state folding kinetics. *Proteins: Struct. Funct. Genet.* **51**, 162–166.
18. Ivankov, D. N., Garbuzynskiy, S. O., Alm, E., Plaxco, K. W., Baker, D. & Finkelstein, A. V. (2003). Contact order revisited: influence of protein size on the folding rate. *Protein Sci.* **12**, 2057–2062.
19. Eicken, C., Sharma, V., Klabunde, T., Lawrenz, M. B., Hardham, J. M., Norris, S. J. & Sacchettini, J. C. (2002). Crystal structure of Lyme disease variable surface antigen VlsE of *Borrelia burgdorferi*. *J. Biol. Chem.* **277**, 21691–21696.
20. Mondragon, A., Wolberger, C. & Harrison, S. C. (1989). Structure of phage 434 Cro protein at 2.35 Å resolution. *J. Mol. Biol.* **205**, 179–188.
21. Chou, P. Y. & Fasman, G. D. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47**, 45–148.
22. Khandelwal, P., Seth, S. & Hosur, R. V. (2000). Stepwise formation of helical structure and side-chain packing in a peptide from scorpion neurotoxin support hierarchic model of protein folding. *Biophys. Chem.* **87**, 139–148.
23. Forge, V., Hoshino, M., Kuwata, K., Arai, M., Kuwajima, K., Batt, C. A. & Goto, Y. (2000). Is folding of beta-lactoglobulin non-hierarchic? Intermediate with native-like beta-sheet and non-native alpha-helix. *J. Mol. Biol.* **296**, 1039–1051.
24. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
25. 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.
26. Zhou, H. & Zhou, Y. (2002). Folding rate prediction using total contact distance. *Biophys. J.* **82**, 458–463.
27. Nolting, B., Schalike, W., Hampel, P., Grundig, F., Gantert, S., Sips, N. *et al.* (2003). Structural determinants of the rate of protein folding. *J. Theor. Biol.* **223**, 299–307.
28. Gromiha, M. M. & Selvaraj, S. (2001). 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.* **310**, 27–32.
29. Mirny, L. & Shakhnovich, E. (2001). Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* **30**, 361–396.
30. Kuznetsov, I. B. & Rackovsky, S. (2004). Class-specific correlations between protein folding rate, structure-derived, and sequence-derived descriptors. *Proteins: Struct. Funct. Genet.* **54**, 333–341.
31. Micheletti, C. (2003). Prediction of folding rates and transition-state placement from native-state geometry. *Proteins: Struct. Funct. Genet.* **51**, 74–84.
32. Anfinsen, C. B., Haber, E., Sela, M. & White, F. H., Jr (1961). The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl Acad. Sci. USA*, **47**, 1309–1314.
33. Levinthal, C. (1968). Are there pathways for protein folding? *J. Chim. Phys.* **65**, 44–45.
34. Bai, Y. (2003). Hidden intermediates and levinthal paradox in the folding of small proteins. *Biochem. Biophys. Res. Commun.* **305**, 785–788.
35. Ptitsyn, O. B. (1973). Stages in the mechanism of self-organization of protein molecules. *Dokl. Akad. Nauk. SSSR*, **210**, 1213–1215.
36. Ptitsyn, O. B. (1994). Kinetic and equilibrium intermediates in protein folding. *Protein Eng.* **7**, 593–596.
37. Karplus, M. & Weaver, D. L. (1976). Protein-folding dynamics. *Nature*, **260**, 404–406.
38. Karplus, M. & Weaver, D. L. (1994). Protein folding dynamics: the diffusion-collision model and experimental data. *Protein Sci.* **3**, 650–668.
39. Baldwin, R. L. & Rose, G. D. (1999). Is protein folding hierarchic? I. Local structure and peptide folding. *Trends Biochem. Sci.* **24**, 26–33.
40. Matouschek, A., Kellis, J. T., Jr, Serrano, L., Bycroft, M. & Fersht, A. R. (1990). Transient folding intermediates characterized by protein engineering. *Nature*, **346**, 440–445.
41. Ferguson, N., Capaldi, A. P., James, R., Kleanthous, C. & Radford, S. E. (1999). Rapid folding with and without populated intermediates in the homologous four-helix proteins Im7 and Im9. *J .Mol. Biol.* **286**, 1597–1608.
42. Munoz, V., Lopez, E. M., Jager, M. & Serrano, L. (1994). Kinetic characterization of the chemotactic protein from Escherichia coli, CheY. Kinetic analysis of the inverse hydrophobic effect. *Biochemistry*, **33**, 5858–5866.

43. Fersht, A. R. (1997). Nucleation mechanisms in protein folding. *Curr. Opin. Struct. Biol.* **7**, 3–9.
44. Galzitskaya, O. V., Ivankov, D. N. & Finkelstein, A. V. (2001). Folding nuclei in proteins. *FEBS Letters*, **489**, 113–118.
45. Wetlaufer, D. B. (1973). Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA*, **70**, 697–701.
46. Wetlaufer, D. B. (1990). Nucleation in protein folding–confusion of structure and process. *Trends Biochem. Sci.* **15**, 414–415.
47. Jackson, S. E. & Fersht, A. R. (1991). Folding of chymotrypsin inhibitor 2. 1. Evidence for a two-state transition. *Biochemistry*, **30**, 10428–10435.
48. Ikura, T., Hayano, T., Takahashi, N. & Kuwajima, K. (2000). Fast folding of Escherichia coli cyclophilin A: a hypothesis of a unique hydrophobic core with a phenylalanine cluster. *J. Mol. Biol.* **297**, 791–802.
49. Jones, K. & Wittung-Stafshede, P. (2003). The largest protein observed to fold by two-state kinetic mechanism does not obey contact-order correlation. *J. Am. Chem. Soc.* **125**, 9606–9607.
50. Srinivasan, R. & Rose, G. D. (1995). LINUS: a hierarchic procedure to predict the fold of a protein. *Proteins: Struct. Funct. Genet.* **22**, 81–99.
51. Dokholyan, N. V., Buldyrev, S. V., Stanley, H. E. & Shakhnovich, E. I. (2000). Identifying the protein folding nucleus using molecular dynamics. *J. Mol. Biol.* **296**, 1183–1188.
52. Snow, C. D., Sorin, E. J., Rhee, Y. M. & Pande, V. S. (2005). How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.* **34**, 43–69.
53. Minor, D. L., Jr & Kim, P. S. (1996). Context-dependent secondary structure formation of a designed protein sequence. *Nature*, **380**, 730–734.
54. Kihara, D. (2005). The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci.* **14**, 1955–1963.
55. Tan, Y. J., Oliveberg, M. & Fersht, A. R. (1996). Titration properties and thermodynamics of the transition state for folding: comparison of two-state and multi-state folding pathways. *J. Mol. Biol.* **264**, 377–389.
56. White, G. W., Gianni, S., Grossmann, J. G., Jemth, P., Fersht, A. R. & Daggett, V. (2005). Simulation and experiment conspire to reveal cryptic intermediates and a slide from the nucleation-condensation to framework mechanism of folding. *J. Mol. Biol.* **350**, 757–775.
57. Kawashima, S., Ogata, H. & Kanehisa, M. (1999). AAindex: amino acid index database. *Nucl. Acids Res.* **27**, 368–369.
58. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H. *et al.* (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235–242.
59. Rost, B. (2001). Review: protein secondary structure prediction continues to rise. *J. Struct. Biol.* **134**, 204–218.

Edited by K. Kuwajima

(Received 23 February 2007; received in revised form 8 April 2007; accepted 18 April 2007)
Available online 4 May 2007