Protein Science (1999), *8*:1210–1217. Cambridge University Press. Printed in the USA.

Diversity of functions of proteins with internal symmetry in spatial arrangement of secondary structural elements

KENGO KINOSHITA, AKINORI KIDERA, AND NOBUHIRO GO

Division of Chemistry, Graduate School of Science, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto 606-8502, Japan

(RECEIVED November 2, 1998; ACCEPTED February 19, 1999)

Abstract

We carry out a systematic analysis of the correlation between similarity of protein three-dimensional structures and their evolutionary relationships. The structural similarity is quantitatively identified by an all-against-all comparison of the spatial arrangement of secondary structural elements in nonredundant 967 representative proteins, and the evolutionary relationship is judged according to the definition of superfamily in the SCOP database. We find the following *symmetry rule*: a protein pair that has similar folds but belong to different superfamilies has (with a very rare exception) certain internal symmetry in its common similar folds. Possible reasons behind the symmetry rule are discussed.

Keywords: evolution of protein; spatial arrangement of secondary structural elements; symmetry of protein folds; three-dimensional structure comparison

Structural comparison and classification are the most fundamental procedures in the studies of protein three-dimensional (3D) structures in the database (Alexandrov & Go, 1994; Murzin et al., 1995; Holm & Sander, 1996; Orengo et al., 1997). Similarity, in the 3D structures is found a distant evolutionary relationship beyond the twilight zone of the sequence comparison. An accumulation of such similar structures will be a basis of structural classification, which will provide us an overview of evolutionary and functional relationships among proteins (Orengo et al., 1994; Brenner et al., 1997). However, we sometimes encounter pairs of proteins that have very similar 3D structures with no clear evolutionary and/or functional relationship. Is there any reason for such similarities?

To address this question, we have classified all entries in the Protein Data Bank (PDB) (Bernstein et al., 1977) by an all-against-all comparison. The structural comparison is performed at the level of the spatial arrangement of secondary structural elements (SSEs) (Mizuguchi & Go, 1995). Judgment of evolutionary relationships is a difficult task to carry out for sure. We employ in this paper a practical method of judgment, i.e., we judge an evolutionary relationship to exist between a pair of proteins if they are classified into the same superfamily in the SCOP database (Murzin et al., 1995). Necessity to refine this judgment will be discussed later.

In this study, by combining these two sets of information, structural similarity and evolutionary relationship, we tried to find a characteristic feature that differentiates the two types of protein pairs having similar 3D structures, one sharing a common superfamily and the other belonging to different superfamilies. In the former case, the structural similarity should definitely show its evolutionary origin. Thus, we focus our attention in this paper on the similarity in the latter case.

Reprint requests to: Akinori Kidera, Department of Chemistry, Graduate School of Science, Kyoto University, Kitashirakawa, Sakyo-ku, Kyoto 606-8502, Japan; e-mail: kidera@qchem.kuchem.kyoto-u.ac.jp.

Abbreviations: PDB, Protein Data Bank; SSE, secondary structural element; TIM, triosephosphate isomerase.

Results and discussion

The 967 representative protein chains in the database are compared in the all-against-all manner by the program COSEC2 (Mizuguchi & Go, 1995), which detects similar spatial arrangement of the SSEs in a pair of protein structures. In this study, a pair is defined to be similar when the following two criteria are met. (1) A high significance level of similarity, i.e., $Z(A,B) \geq 5$ [Eq. (10)] and (2) a sufficiently large size of similar structure, i.e., the number of corresponding SSEs ≥ 6. For each pair of similar proteins, their superfamily names are assigned after the definition of the SCOP (Murzin et al., 1995). In this procedure, it is observed that many pairs of proteins, each having the TIM barrel fold, are detected to have similar structure and to belong to different superfamilies. Because the TIM barrel fold is a highly symmetric fold, we have carried out further analysis to find any internal symmetry in other protein pairs that have similar structure but belong to different superfamilies. A protein fold is defined to have a internal symmetry when it has a nontrivial self-similar correspondence (see Materials and methods).

The result is summarized in Table 1, where each similar pair is classified into one of the three types of the structural similarity relationships, Type 1 (an overall correspondence), Type 2 (one containing the other), and Type 3 (only a part of the structure shared), as illustrated in Figure 1. The results in each type will be discussed in turn.

Table 1. *Number of pairs for each similarity type*[a]

Type	Common superfamily[b]	Different superfamilies[c]		Total
		Symmetric[d]	Nonsymmetric[e]	
Type 1	385	32	2	422
Type 2	368	185 (128)[f]	50 (8)[f]	503
Type 3	654	833 (570)[f]	422 (95)[f]	1,909
Total	1,407	1,050	474	2,834

[a]A similar pair satisfies the two criteria: (a) $Z(A,B) \geq 5$ (Equation 10) and (b) the number of corresponding SSEs ≥ 6.
[b]The number of pairs belonging to the same superfamily.
[c]The number of pairs belonging to different superfamilies.
[d]The number of pairs, which belong to different superfamilies and whose similar substructures are symmetric.
[e]The number of pairs, which belong to different superfamilies and whose similar substructures are nonsymmetric.
[f]The numbers in the parentheses are the number of corresponding SSEs ≥ 8 for criterion b.

Type 1 relation—overall correspondence

Figure 2 shows the number of Type 1 protein pairs belonging to the same superfamily (common SF) and to different superfamilies (further classified into symmetric and nonsymmetric folds) against the number of corresponding SSEs (Fig. 2A), or against the *Z*-score of Equation 10 (Fig. 2B). Sum of all entries in Figure 2 is given in the Table 1. We see that more than 90% of Type 1 protein pairs belong to the same superfamily. This fact remains valid, even when we shift the threshold values for the criteria of judging structural similarity as can be seen in Figure 2.

As mentioned, we will focus more on protein pairs belonging to different superfamilies. The result is so striking that we describe it as a rule.

Symmetry rule: Similar protein pairs belonging to different superfamilies have (with a very rare exception) certain internal symmetry in their common folds.

One of the examples of such symmetric folds is shown in Figure 3, where we can see a C_3 symmetry in both proteins. There are only two exceptions to the above symmetry rule; (1) histidine-rich actin-binding protein (1hce) vs. interleukin-1 (1i1b), and (2) c-Raf1, Ras-binding domain (1guab) vs. ubiquitin (1ubi). The folds of the former pair are found to be almost symmetric by visual inspection, but due to their skewed arrangements of the SSEs, computer program COSEC2 could not detect it. In the latter case, however, the folds (β-grasp) are clearly nonsymmetric as shown in Figure 4. Therefore, the number of exceptions to the symmetry rule is just 1 rather than 2 as in Table 1.

The exception of the β-grasp was found because this pair has six corresponding SSEs more than the threshold. The SCOP database contains five entries (1guab, 1ubi, 1tif, 1igd, and 2ptl) belonging to the β-grasp fold. Two of them contain five SSEs, and the other three entries have six SSEs. Owing to the threshold in the number of corresponding SSEs, and due to their structural variety, our method identified significant similarities only for the pair between 1guab and 1ubi.

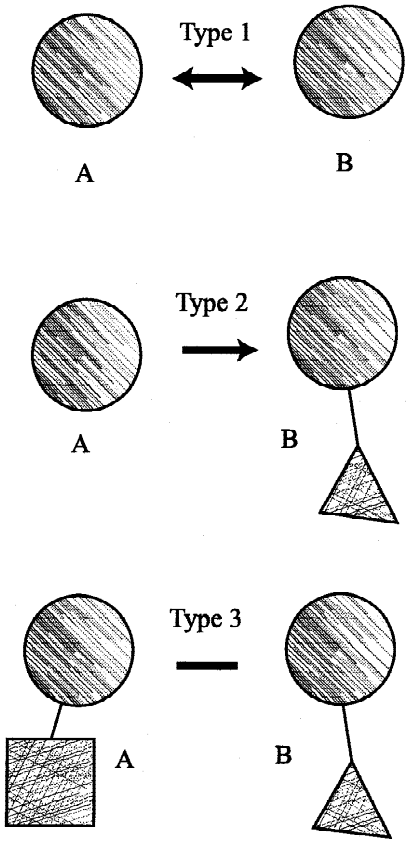

Fig. 1. Schematic representation of the three types of structural similarity: Type 1 (an overall correspondence), Type 2 (one containing the other), and Type 3 (only a part of the structure shared).

Type 2 relation (one containing the other) and Type 3 relation (only a part of the structure shared)

Figures 5 and 6 are the same as Figure 2 but for Type 2 and Type 3 protein pairs, respectively. The symmetry rule still holds in these two types of similarity but with a slightly weaker form. However, exceptions occur only when the number of corresponding SSEs < 8 in both Type 2 and Type 3 relations. Therefore, the rule can be made to hold more strictly, even in Type 2 and Type 3 relations, if we employ SSEs ≥ 8 in criterion (2) for judging structural similarity. Because the symmetry rule holds strongly in all the three types of similarity relations, we want to examine the exceptional cases more carefully.

In Figure 7, examples of these exceptions are illustrated. In Figure 7A, thioredoxin (1thx) (Saarinen et al., 1995) and phosducin C-terminal domain (2trc chain P) (Gaudet et al., 1996) show significant structural similarity, but their equivalent SSEs have no internal symmetry. These proteins are assigned to different superfamilies, probably because phosducin does not have the two cysteine residues conserved in the thioredoxin active site. However, as shown by Gaudet et al. (1996), the structurally corresponding parts show weak homology (22% identity but 52% similarity) (Gaudet et al., 1996). Also in the Type 3 relation shown in Figure 7B, structurally corresponding parts in insecticidal toxin CRYIA (A) (1ciy) and PNGas F (glycosylasparaginase from *Flavobacterium menigosepticum*; 1pgs) show 14% identity and 49% similarity.

We cannot make a definite judgment about evolutionary relationship in the above two cases, because the significance of the sequence similarity is very subtle. However, the above examples

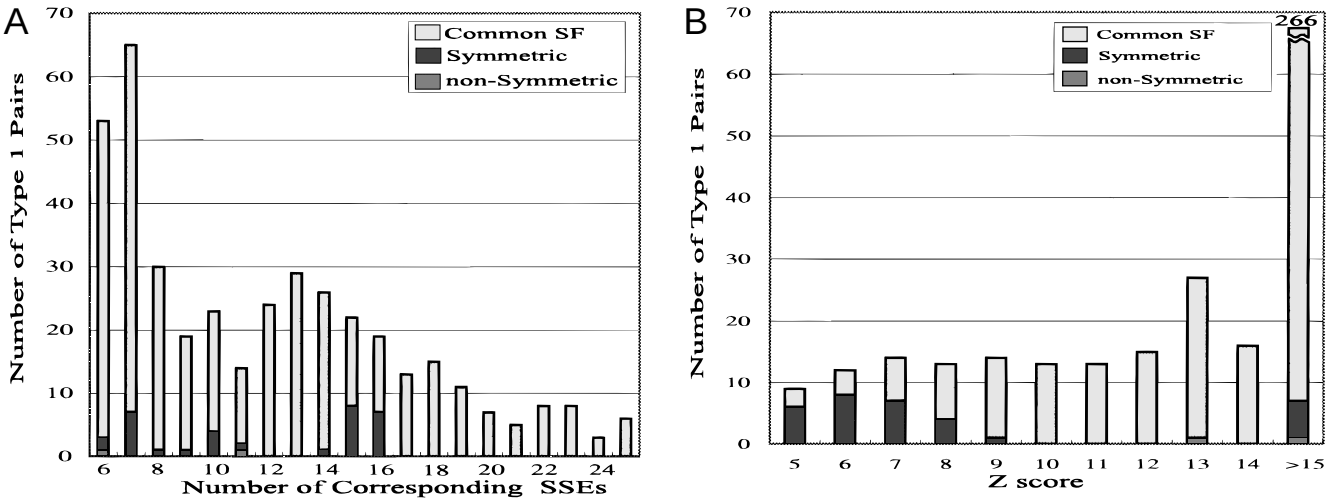

Fig. 2. (**A**) The distribution of the number of Type 1 protein pairs classified against the number of corresponding SSEs and (**B**) against the *Z*-score. The protein pairs are classified into those belonging to a common superfamily and those belonging to different superfamilies. The latter is further classified by the folds, symmetric fold, and nonsymmetric fold.

suggest that a certain number of cases classified in Table 1 to "different superfamilies and nonsymmetric" may actually be evolutionarily related. They are the cases judged in the SCOP database to belong to different superfamilies, but have in fact a weak evolutionary relationship. When these possible cases are removed from the fourth column in Table 1, and when we understand it as indicating "no evolutionary relation and nonsymmetric," its entries should be even smaller. Then the symmetry rule becomes more striking.

Frequently observed symmetric folds

The number of folds of protein domains is known to have an extremely skewed distribution; that is, a small number of folds are very dominant in the database (Orengo et al., 1994; Brenner et al., 1997). Here, we try to classify the symmetric folds detected in the comparison. There are three very dominant folds. Figure 8 shows a distribution of the number of symmetric folds, which appeared in all three types of similarity relations against the number of corresponding SSEs. The number of cases belonging to the dominant folds is also indicated. Two of the dominant folds are well-known ones, TIM barrel and Ig fold. Another dominant one basically consists of four parallel β-strands and four or more intervening α-helices with the topology $-1x,2x,1x$ (in Richardson's nomenclature) (Richardson, 1977), which we propose to call R-motif, because it is contained in the well-known Rossmann folds (Rao & Rossmann, 1973). Some examples of the R-motifs appearing in various folds are summarized in Figure 9. Since the functions of proteins in Figure 9 are diverse, they are classified into various folds in SCOP (Murzin et al., 1995). However, as far as the structural similarity concerns, they should be classified into a single structural group, R-motif.

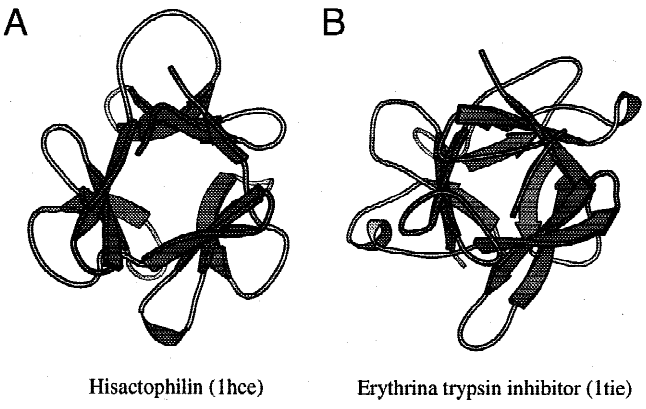

Fig. 3. (**A**) Ribbon representation of a protein pair having a symmetric fold, hisactophilin (1hce), and (**B**) erythrina trypsin inhibitor (1tie). According to the SCOP classification, these two proteins belong to different superfamilies, histidine-rich actin-binding protein, and Kunitz (STI) inhibitors, respectively. This and succeeding figures were drawn by MOLSCRIPT (Kraulis, 1991).

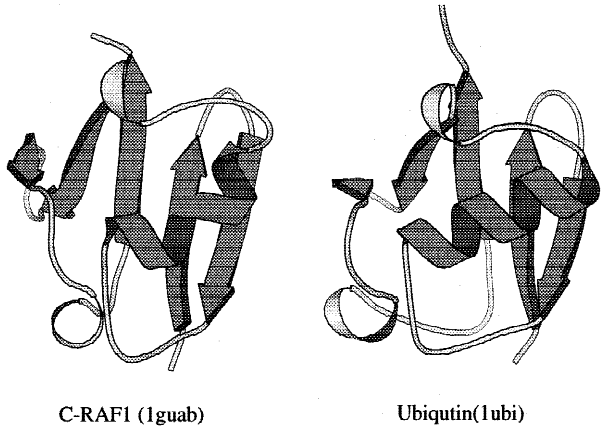

Fig. 4. Ribbon representation of a protein pair having a nonsymmetric fold, C-RAF1 (1guab) and ubiquitin (1ubi). Due to a central α-helix, these folds do not have any internal symmetry.

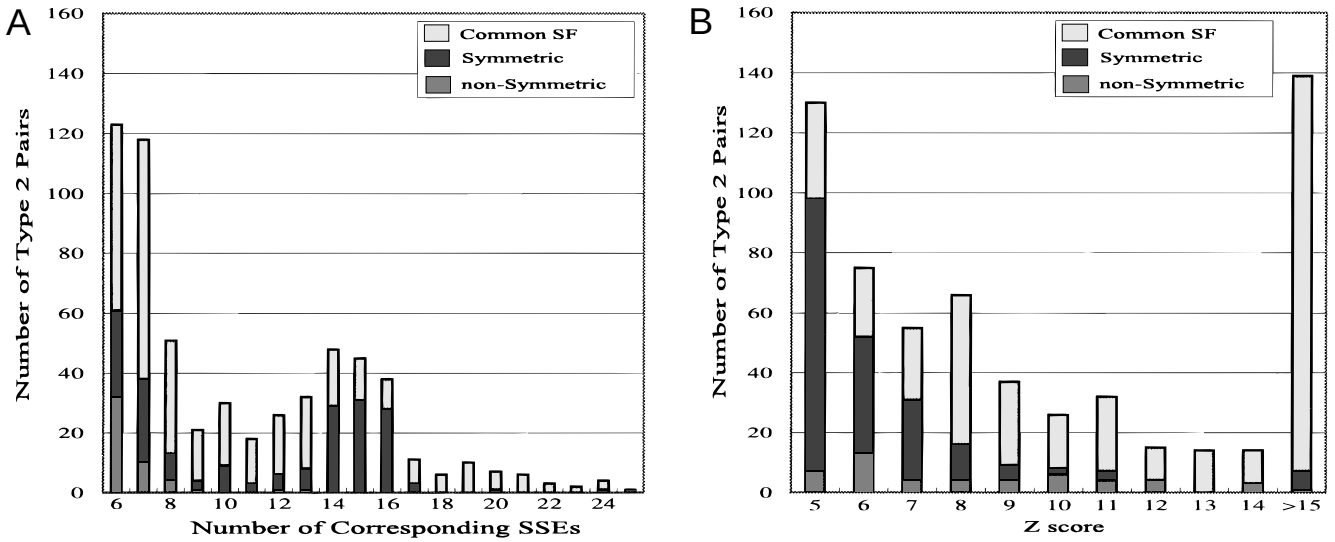

Fig. 5. (**A**) The distribution of the number of Type 2 protein pairs against the number of corresponding SSEs and (**B**) against the *Z*-score. The meaning of each bar is the same as in Figure 2.

Possible reasons behind the symmetry rule

Because the symmetry rule holds so strongly in all three types of similarity relations, we have to look for reasons behind it. We think that there is not enough evidence to exclude either one of the convergence or the divergence mechanism behind it.

In the divergence point of view, we are forced to accept that proteins with internal symmetry in their spatial arrangement of SSEs have been able to create different functions in their history of molecular evolutions, while those with no internal symmetry have been able to sustain only one function. Because different functions mean generally grossly different amino acid sequences, this should mean that proteins with internal symmetry have a larger allowance to its amino acid mutations. An amino acid replacement would be accepted when the new mutant proteins are reasonably stable and fast folding. We will later discuss possible relations between internal symmetry and foldability.

In the convergence point of view, we assume that different proteins with the same internally symmetric folds have emerged independently in the history of molecular evolutions, while proteins with no internal symmetry have had very small chances of independent emergence. Emergence of a new fold should, of course, be influenced by the stability and foldability of such a fold.

We think that there are two reasons why proteins with internal symmetry are fast folding. The first is the one pointed out by Wolynes (1996). This is based on the argument that a symmetric

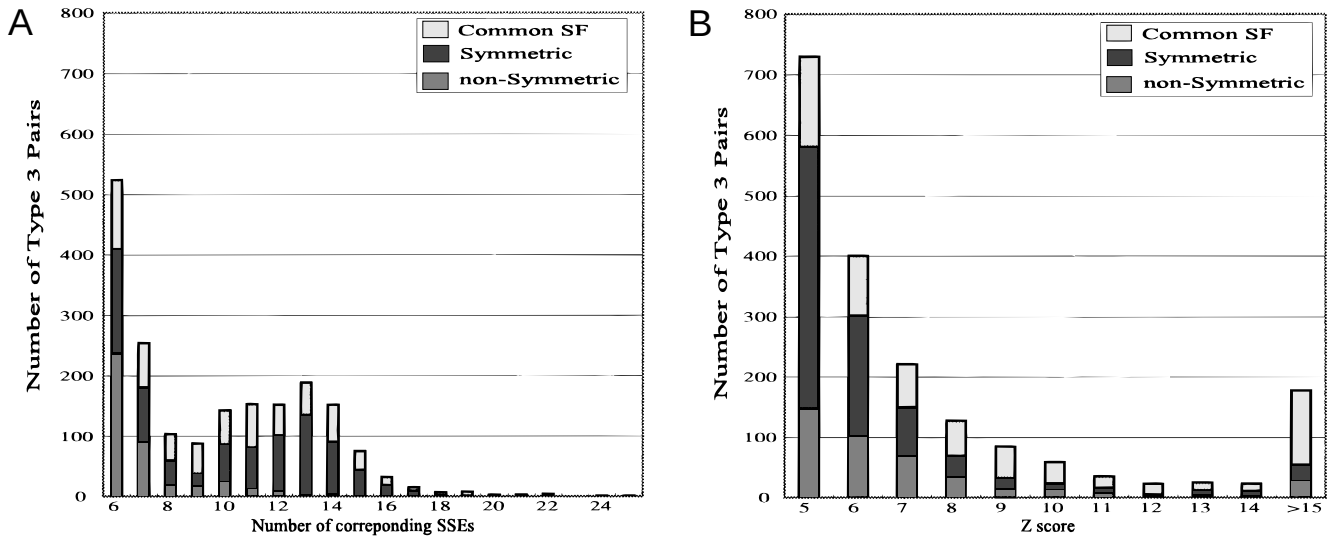

Fig. 6. Same as Figure 2 or 4, but for Type 3 protein pairs.

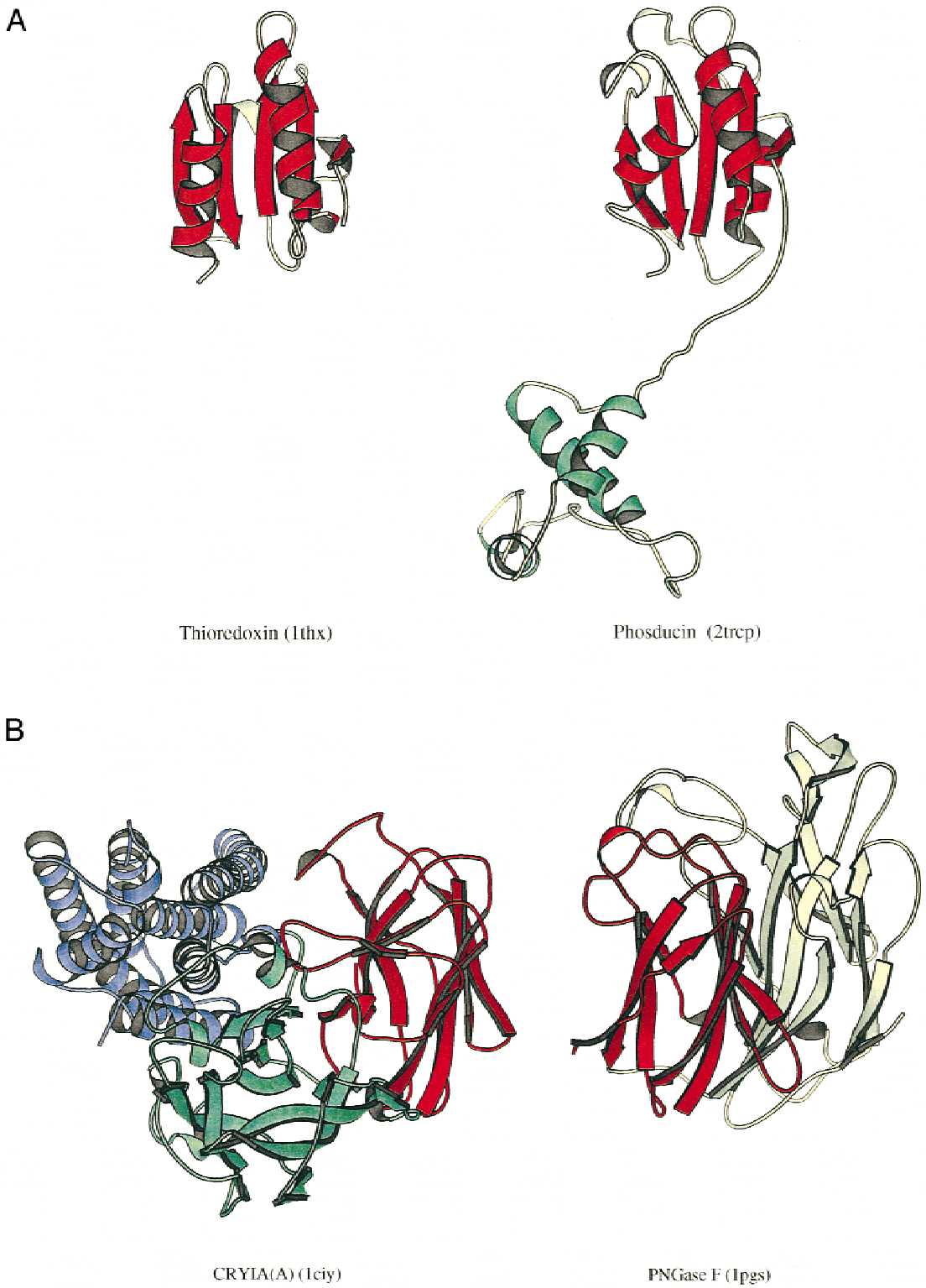

Fig. 7. Ribbon representation of protein pairs of (**A**) Type 2 and (**B**) Type 3 relations, respectively, belonging to different superfamilies and having nonsymmetric folds. Corresponding SSEs are shown in red. **A:** Thioredoxin (1thx) and phosducin C-terminal domain (2trc chain P) show significant similarity ($Z = 10.7$ and the number of corresponding SSEs = 8) but have nonsymmetric folds. They belong to thioredoxin-like and phosducin superfamilies, respectively. **B:** Insecticidal toxin CRYIA(A) and PNGas F (glycosylasparaginase from *F. menigosepticum*) show significant similarity ($Z = 16.6$ and corresponding SSEs = 9) but have no symmetry in their similar part. The sequences in the corresponding region show 14% identity and 49% similarity. They belong to galactose-binding domain-like and glycosyl-asparaginase superfamily, respectively.

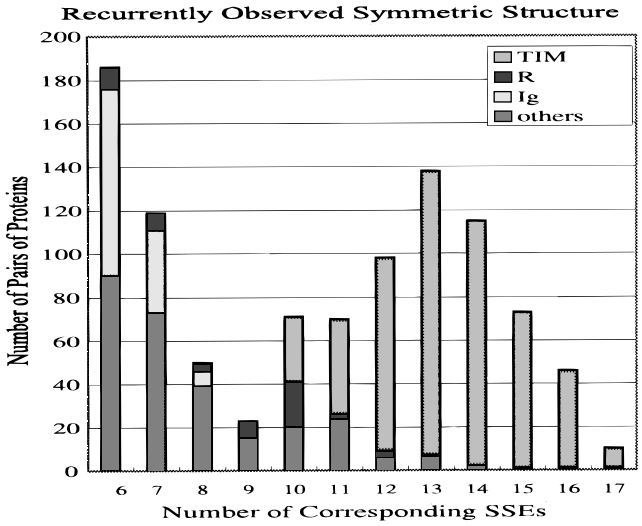

Fig. 8. The distribution of the number of symmetric folds, which appeared in all the three types of similarity relations, plotted against the number of corresponding SSEs. The number of cases belonging to the three dominant folds, TIM barrel, Ig fold, and R-motif, is also indicated.

fold would have multiple folding pathways; that is, any symmetry related part can have an equal probability to form the initial nucleus of the folding process. The second comes from our impression that the symmetric folds appear to have simpler structures than nonsymmetric ones, i.e., in symmetric folds neighboring SSEs along the sequence tend to be closer to each other in the 3D structure. This impression becomes more obvious if we regard the N-terminus to follow the C-terminus. Because SSEs near the chain can easily interact, formation of the folding nucleus in proteins with such simpler structures should be easy, leading to an enhanced foldability in such proteins (Chothia & Finkelstein, 1990).

Materials and methods

Structural comparison method

We compared protein structures with program COSEC2 (Mizuguchi & Go, 1995) to improve the computational performance and the reasonableness of the definition of structural similarity. In this program, a protein structure is described by a set of vectors representing SSEs. The optimal superposition for a given pair of proteins, say *A* and *B*, is calculated by searching for similar spatial arrangements of the vectors that maximize the following similarity score:

$$S(A,B) = \sum_{\text{SSEpair}} \sum_{k=1}^{6} \left(w_k - \frac{(q_k^A - q_k^B)^2}{w_k} \right) \qquad (1)$$

where q_k^A is the k^{th} index ($k = 1,\ldots,6$) describing the mutual position of a SSE pair in protein *A*, and w_k is the normalization factor. The summation is taken over all pairs of SSEs and the six indices. The six indices correspond to all degrees of freedom to describe the mutual arrangement of two vectors, i.e., direction, lengths, tilt angles, and mutual distance of the vectors are considered at the same time. The definition of the six indices and the normalization factors is described in Mizuguchi and Go (1995). In this study considering the relation between structural similarity and evolutionary relationship, sequence order dependent similarities are forced in the definition of similar pair. We assigned the secondary structural code to each residue according to the definition of DSSP (Kabsch & Sander, 1983) with the following modifications required for the vector representation of SSE: (1) A bent strand is a strand satisfying the following conditions: N_{res} (the number of residues in a SSE) ≥ 7 and D_{res} (a distance between the terminal C_α atoms of a SSE divided by N_{res}) < 2.0 Å. A bent strand is divided into two successive strands at the residue, whose distance from a line between the terminal C_α atoms is maximum. This procedure is repeated until all strands become nonbent. (2) A bent helix is a helix satisfying the following conditions: $N_{res} < 12$ and $D_{res} < 1.35$ Å. A bent helix is divided into two successive helices at the residue whose (ϕ,ψ) angles deviate maximally from the mean value of a typical α-helix (64.5, 39.7). A resulting helix having less than four residues will be re-assigned as coil. (3) Splitting helices are two successive helices satisfying the following conditions: the number of intervening residues is two or less, the distance between the first and the last C_α of the first helix, and the first C_α of the second helix atoms, is less than 7.0 Å, and the angle between vectors for the two helices is less than 30°. A pair of splitting helices is merged into a helix by converting the assignments of the intervening residues into helix.

For Type 3 relation with a small number of corresponding SSEs ($\leq$10), we further imposed a criterion of compactness to the definition of similarity. This is to avoid the case, where some nonsimilar SSEs are situated in the middle of similar SSEs. The compactness is defined by the condition, $d_{max}(N) \leq \langle d_{max}(N)\rangle + 2\sigma(N)$, where $d_{max}(N)$ is the maximum value of the midpoint distances among N SSEs in the corresponding pair, and $\langle d_{max}(N)\rangle$ and $\sigma(N)$ are the mean value and the standard deviation of $d_{max}(N)$ in the monomeric globular proteins.

We compared protein structures in the PDB select 35% list, October 1997 release (Hobohm et al., 1992; Hobohm & Sander, 1994). But entries with only C_α coordinates are ignored. The total number of the proteins compared in this study is 967, in which an oligomer protein is counted by the number of its chains.

Definition of internal symmetry

We define a protein fold to have an internal symmetry when it has a nontrivial (i.e., nonidentical) self-similar correspondence with f (the fraction of SSEs) ≥ 0.7. In Type 2 and Type 3 relations, an internal symmetry can be defined not for a whole protein, but for the part of SSEs assigned to be similar. This definition covers the rotatory symmetries and the screw symmetries with many repeating units. In principle, we should also examine the mirror symmetry and the screw symmetry with a small number of repeating units. To search for a mirror symmetry fold, we compared all protein structures with each of their mirror images. It was confirmed that there is no mirror symmetric fold in the database. It may be because the influence of chirality in the peptide structure remains even in the level of the vector representation, e.g., the right-handed twist of a β-sheet. The screw symmetry fold with a small number of repeating units can be detected by the following method. When nontrivial self-similar parts are found by allowing them not to satisfy the condition, $f \geq 0.7$, such similar parts in the protein are superimposed to give a translation vector and a rotation matrix. This protein is regarded to have the screw symmetry if the

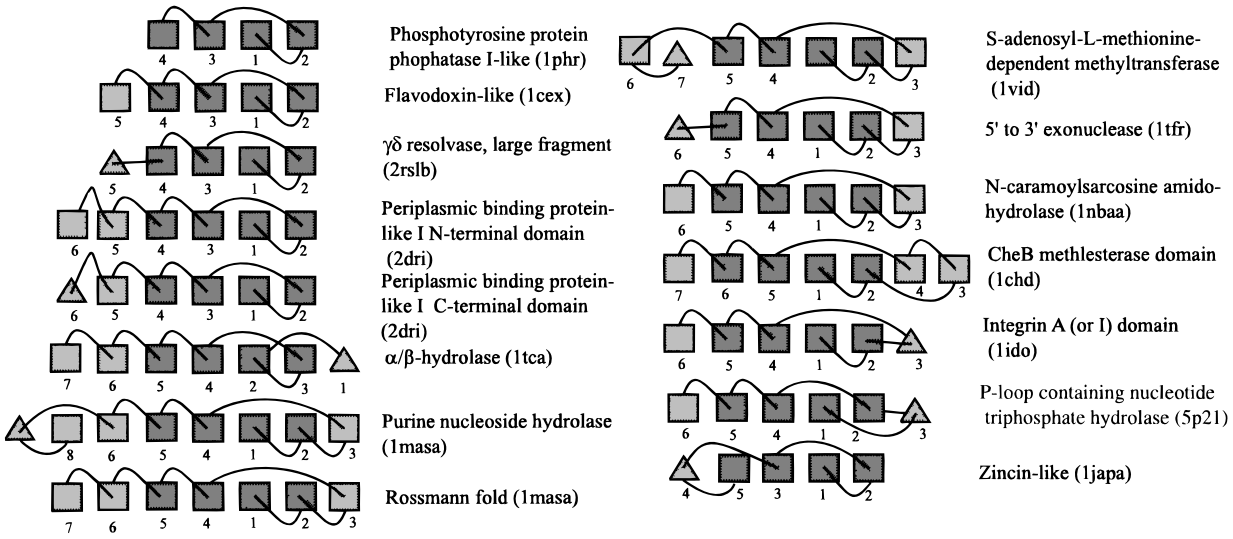

Fig. 9. R-motifs appearing in various protein folds. Each square box represents a parallel strand and a triangle indicates an strand antiparallel to the rest. A set of four deeply shadowed squares represent the R-motif. Intervening helices between parallel strands are not written in the figure. The names of the folds given in the SCOP database along with PDB ID are shown.

translation vector and the rotation axis are parallel to each other. As a result of such an analysis, it turned out that all screw symmetric folds are already found with the condition, $f \geq 0.7$.

Significance of structural similarity

The comparison method, COSEC2, does not require identification domains prior to the comparison. It automatically detects the similar portions in the query proteins even when their sizes are largely different. Therefore, when one compares protein *A* with protein *B*, there should be the following three cases, as schematically shown in Figure 1:

1. Type 1: an overall similarity where both *A* and *B* have large *f*-values, i.e., $f_A \geq 0.7$ and $f_B \geq 0.7$.
2. Type 2: *A* contains *B* where $f_A < 0.7$ and $f_B \geq 0.7$.
3. Type 3: only a part of the structures is shared where $f_A < 0.7$ and $f_B < 0.7$.

When we find a pair of similar structures, the significance is evaluated by calculating how seldom such a pair can be found in a random set of protein structures. Here, we adopted the following assumption for the random set. The random distribution of the similarity score for protein *A*, $p_A(S)$, can be calculated from the probability of occurrence of any protein *B* having the score, $S(A,B)$ (Equation 1), in the database consisting of 967 proteins. In other words, the database used here is assumed the random set. It is further assumed that $p_A(S)$ is given by a Gaussian function of a mean and a standard deviation defined for protein *A*, $\langle S\rangle_A$, and σ_A, by

$$p_A(S) = \frac{1}{\sqrt{2\pi}} \exp(-z^2/2) \qquad \text{with } z = \frac{S - \langle S\rangle_A}{\sigma_A}. \tag{2}$$

Therefore, when a comparison of *A* and *B* gives a similarity score $S_0(A,B)$, the significance of the similarity can be assessed by the probability of finding a score $S \geq S_0(A,B)$ as

$$p[S \geq S_0(A,B)] = \text{erfc}[z_0(A,B)] \tag{3}$$

with

$$z_0(A,B) = \frac{1}{2}\left[\frac{S_0(A,B) - \langle S\rangle_A}{\sigma_A} + \frac{S_0(A,B) - \langle S\rangle_B}{\sigma_B}\right], \tag{4}$$

where erfc is an error function, and $z_0(A,B)$ is the average of the two *z* values defined for *A* and *B*, respectively. Since the similar structures *A* and *B* will give similar values of $\langle S\rangle$ and σ, the averaging in Equation 4 is actually trivial.

Note, however, that the criterion of Equation 3 is applicable only to Type 1, or an overall similarity from N- to C-terminus. In the Type 2 or Type 3 relation, protein *A* contains not only the SSEs similar to protein *B*, but also the remaining SSEs, both of which would influence the parameters, $\langle S\rangle_A$ and σ_A. The influence from the latter part of SSEs would make the value of $p(S \geq S_0)$ inappropriate for assessing the significance of the similarity between proteins *A* and *B*. To solve this problem, we adopted the following procedure. When a pair of similar proteins, *A* and *B*, are found, these similar portions, *a* (ϵ A) and *b* (ϵ B), excluding all the rest of the proteins, are subject to the comparison against all the other proteins in the database. Then, the parameters $\langle S\rangle_a$, σ_a, $\langle S\rangle_b$, and σ_b are calculated for these similar portions, *a* and *b*, to evaluate the significance, $p(S \geq S_0)$. Here, we have to take into account the effects of the original size of proteins *A* and *B*. Such a size effect is based on the fact that the chance of finding a fragment having a certain structure should increase with the size of the protein, or the significance should decrease with the size. To incorporate such an effect in $p(S \geq S_0)$, we adopt the following function for the significance of the structural similarity, instead of Equation 3:

$$p[S \geq S_0(A,B)] = \frac{n_A(n_A-1)n_B(n_B-1)}{m(m-1)(\langle n^2\rangle - \langle n\rangle)} \text{erfc}[z_0(a,b)], \tag{5}$$

where n_A and n_B are the number of SSEs in protein *A* and *B*, respectively; *m* is the number of the corresponding SSEs in the two

proteins, and the averages $\langle n^2 \rangle$ and $\langle n \rangle$ are the square average and the average of the number of SSEs in proteins in the database, respectively.

Equation 5 was derived by an analogy of *p*-value of the sequence comparison (Karlin & Altschul, 1990). The significance of similarity between two sequences of length n_A and n_B is measured for a large S_0 by

$$p(S \geq S_0; n_A, n_B) = K n_A n_B \exp(-\lambda S_0), \quad (6)$$

where n_A and n_B are explicitly written in the arguments of p, and K and λ are constants. Here, the prefactor $n_A n_B$ is the number of ways to compare two sets of amino acids in the proteins *A* and *B*. In the structural comparison in this study, the element defining the similarity score of Equation 1 is a pair of SSEs in a protein. Thus, the corresponding prefactor should be

$$\frac{n_A(n_A-1)}{2} \frac{n_B(n_B-1)}{2}$$

instead of $n_A n_B$. This analogy results in

$$p(S \geq S_0; n_A, n_B) = K \frac{n_A(n_A-1)}{2} \frac{n_B(n_B-1)}{2} \exp(-\lambda S_0). \quad (7)$$

This is the criterion for the significance. According to our procedure written above, m corresponding SSEs in two proteins are compared against all other proteins in the database. This comparison would result in the following *p*-value:

$$\int p(S \geq S_0; l, n) f(n)\, dn = K \int \frac{m(m-1)}{2} \frac{n(n-1)}{2} \times \exp(-\lambda S_0) f(n)\, dn, \quad (8)$$

where n is the number of SSEs of a protein in the database, and $f(n)$ is the distribution function of n in the database. Equation 8 simply becomes

$$p(S \geq S_0; l, \langle n \rangle) = K \frac{m(m-1)}{2} \frac{(\langle n^2 \rangle - \langle n \rangle)}{2} \exp(-\lambda S_0), \quad (9)$$

where

$$\langle n^2 \rangle = \int n^2 f(n)\, dn \quad \text{and} \quad \langle n \rangle = \int n f(n)\, dn.$$

Comparing Equation 9 with Equation 3, the right-hand side of Equation 9 should be equal to Equation 3. Therefore, we finally have Equation 5. The assessments of the significance in this study were performed by Equation 5 with the values of $\langle n^2 \rangle$ and $\langle n \rangle$ being 285.4 and 16.6, respectively, which were calculated from the database of the 967 proteins.

This definition was inspired by the definition in the database VAST, which also uses an analogy of *p*-value, but defines the random set of protein structures by those randomly generated in the Cartesian space (Gibrat et al., 1996). The difference is explained as follows. According to our definition, the structural similarity for a protein frequently occurring in the database tends to be less significant than that of a protein rarely seen in the database. On the other hand, VAST adopts a pure geometrical definition of the structural similarity (Gibrat et al., 1996).

In the text, for the illustrative purpose, $p[S \geq S_0(A,B)]$ defined in Equation 5 is rescaled into the following *Z*-score:

$$\mathrm{erfc}[Z(A,B)] = p[S \geq S_0(A,B)]. \quad (10)$$

This *Z*-score contains the influence of the original size of the query protein while the *z* value defined in Equation 4 does not.

Acknowledgments

We thank Dr. S.H. Bryant for helpful discussions. This work has been supported by a grant from MESC to A.K. and N.G. The computations were done in the Computer Center of the Institute for Molecular Science, Center for Promotion of Computational Science, Engineering of JAERI, and Data Processing Center, Kyoto University, Kyoto, Japan.

References

Alexandrov NN, Go N. 1994. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci 3*:866–875.

Bernstein FC, Koetzle TF, Williams GJ, Meyer EE, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Brenner SE, Chothia C, Hubbard TJP. 1997. Population statistics of protein structures: Lesson from structural classification. *Curr Opin Struct Biol 7*:369–376.

Chothia C, Finkelstein AV. 1990. The classification and origins of protein folding patterns. *Annu Rev Biochem 59*:1007–1039.

Gaudet R, Bohm A, Sigler BP. 1996. Crystal structure at 2.4 Å resolution of the complex of transducin beta-gamma and its regulator, phosducin. *Cell 87*:577–588.

Gibrat JF, Madej T, Bryant SH. 1996. Surprising similarities in structure comparison. *Curr Opin Struct Biol 6*:377–385.

Hobohm U, Sander C. 1994. Enlarged representative set of protein structures. *Protein Sci 3*:522–524.

Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci 1*:409–417.

Holm L, Sander C. 1996. Mapping the protein universe. *Science 273*:595–602.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of proteins structures. *J Appl Cryst 24*:946–950.

Mizuguchi K, Go N. 1995. Comparison of spatial arrangements of secondary structural elements in proteins. *Protein Eng 8*:353–362.

Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*:536–540.

Orengo C, Michie A, Jones S, Swindells M, Thornton J. 1997. CATH: A hierarchic classification of protein domain structures. *Structure 5*:1093–1108.

Orengo CA, Jones DT, Thornton JM. 1994. Protein superfamilies and domain superfolds. *Nature 372*:631–634.

Rao ST, Rossmann MG. 1973. Comparison of supersecondary structures in proteins. *J Mol Biol 76*:241–256.

Richardson JS. 1977. Beta-sheet topology and the relatedness of proteins. *Nature 268*:495–500.

Saarinen M, Gleason KF, Eklund H. 1995. Crystal structure of thioredoxin-2 from Anabaena. *Structure 3*:1097–1108.

Wolynes PG. 1996. Symmetry and the energy landscapes of biomolecules. *Proc Natl Acad Sci USA 93*:14249–14255.