doi:10.1006/jmbi.2001.5250 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2002) **315**, 887–898

JMB

Protein Fold Similarity Estimated by a Probabilistic Approach Based on C^{α}-C^{α} Distance Comparison

Oliviero Carugo[1,2]* and Sándor Pongor[1]

[1] *Protein Structure and Function Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, Padriciano 99 34012, Trieste, Italy*

[2] *Department of General Chemistry, University of Pavia Viale Taramelli 12 27100, Pavia, Italy*

**Corresponding author*

The distribution of the C^{α}-C^{α} distances between residues separated by three to 30 amino acid residues is highly characteristic of protein folds and makes it possible to identify them from a straightforward comparison of the distance histograms. The comparison is carried out by contingency table analysis and yields a probability of identity (PRIDE score), with values between zero and 1. For closely related structures, PRIDE is highly correlated with the root-mean-square distance between C^{α} atoms, but it provides a correct classification even for unrelated structures for which a structural alignment is not meaningful. For example, an analysis of the CATH database of fold structures showed that 98.8 % of the folds fall into the correct CATH homologous superfamily category, based on the highest PRIDE score obtained. Structural alignment and secondary-structure assignment are not necessary for the calculation of PRIDE, which is fast enough to allow the scanning of large databases.

Keywords: protein structure; protein fold; protein domain; protein-protein similarity; protein classification

Introduction

Understanding protein structure is central to the post-genomic era. In order to understand the data, the structures need to be categorized according to their recurrent local motifs or folds, which is not an easy task, since nearly all proteins have structural similarities to other proteins.[1] In order to find the known folds in a protein structure one needs, on the one hand, a comprehensive collection of fold structures and, on the other, a suitable search program that can compare the new structure with the entries of the fold collection. Even though the construction and maintenance of a fold database is a formidable task, several such collections have been developed, such as FSSP, SCOP and CATH.[2–4] However, the quantitative assessment of structural similarity is problematic in many respects. First of all, the comparison of three-dimensional (3D) structures is very computer-intensive, which is partly due to the nature of structural alignments. Second, the root-mean-square distance (rmsd) value (which is calculated between the C^{α} atoms of a pair of optimally superposed structures) works well as an indicator of similarity only if the structures are closely related. Clearly, distantly related structures may only share a small segment that can be structurally aligned, and the magnitude of rmsd values determined within that small segment may not reflect the similarity of otherwise divergent structures. The categorization of unrelated structures depends at present on knowledge-based fold-classification schemes that are different in many subtle details, such as the assignment of secondary structures and domain boundaries, and the differences may, in some cases, lead to classification conflicts.[5] Furthermore, structural-genomics initiatives are set to produce a large amount of data,[6–8] so there is a clear need for novel, fast data-analysis strategies to extract biologically relevant information.[9]

The present work aims to define a simple numerical measure of fold similarity that is sufficiently fast to compute so as to allow the scanning of large databases. The underlying assumption is that if a fold of a protein can be reconstructed from the distances measured between all pairs of its C^{α} atoms, then folds can also be compared in terms of their C^{α}-C^{α} distances. We used histograms of the C^{α}-C^{α} distance distributions as a simplified representation of the structure. The comparison of distance histograms can be carried out using the standard statistical method of contingency table analysis that yields a probability of identity value for the two structures that we call the PRIDE score.

Abbreviations used: 3D, three-dimensional; rmsd, root-mean-square deviation(s).

E-mail address of the corresponding author: carugo@icgeb.trieste.it

0022-2836/02/040887–12 $35.00/0

On closely related 3D structures the PRIDE score is well correlated with rmsd distance. It is also shown, using the CATH database[4] as a model, that PRIDE scores correlate well with the similarity relations of the more distantly related fold groups. As the calculation of the score does not require either a structural alignment or knowledge-based decisions, such as the assignment of secondary structures, we believe that it can be useful in automated fold-classification programs.

Results and Discussion

The PRIDE score: probability of identity between two protein structures

For the calculation of similarity, we represent the protein structure by the length distribution of its C^α distances. For the sake of simplicity, let us suppose that we describe a protein fold by a set of its $C^\alpha(i)$-$C^\alpha(i+8)$ distances, i.e. by the distances separating the C^α atoms eight residues apart, and prepare a histogram of the length distribution. Identical structures will give rise to identical histograms and, more importantly, the probability of identity between two histograms derived from two different structures can be simply assessed by contingency-table analysis based on the χ^2 test. Contingency-table analysis is a robust statistical method that can also be applied if there are no analytical models describing the distribution of the population.[10] The details of the calculation are described in Data and Methods. Naturally, there is no reason to limit the description of a structure to $C^\alpha(i)$-$C^\alpha(i+8)$ distances. For the calculation of the fold similarity, we described a protein structure by the distributions of its $C^\alpha(i)$-$C^\alpha(i+n)$ distances, where n is an integer ranging from 3 to 30, i.e. by 28 histograms. The probability of identity (*PRIDE*) of two protein structures was then calculated by comparing each of the 28 histogram pairs, and calculating the average of the resulting 28 probability values.

An example of the computation is given in Figure 1. The crystal structure of the N-terminal domain of human serum albumin has been reported in two different space groups, *C*2 (PDB file 1bj5) and $P2_1$ (1uor).[12,13] There are only minor differences between the two structures (Figure 1(a)) due to different crystal-packing interactions and to different ligands complexed to the protein. It is apparent, for example, that the last two C-terminal helical segments are quite different in 1bj5 and 1uor. The histograms showing the distribution of the $C^\alpha(i)$-$C^\alpha(i+n)$ distances for $n=8$, 15, 22, and 29 are shown in Figure 1(b). The *PRIDE* values, computed for each of the 28 histogram pairs with $3 \leqslant n \leqslant 30$ (Figure 1(c)), show that the two structures differ essentially because of the spatial arrangement of residues close to each other in the protein sequence. The probability values oscillate around 0.5 for $n<10$ while they approach 1.0 for higher n values. This is in fact expected, since the two structures differ only in the local stereochemistry of a few polypeptide segments, while their global conformation is quite similar. The average of the 28 probability values (the PRIDE score) is 0.73, i.e. the probability that the two structures are identical is 73 %.

The upper and lower limits of n, 30 and 3, were chosen on an empirical basis; other values may be chosen as the method is adapted for specific tasks. We found that n values below 3 carry little information, probably because the covalent nature of the peptide bond makes the distances between residues close in sequence nearly constant and independent of the protein conformation. In contrast, n values of 3 or slightly higher are able to distinguish, for example, the helical backbone conformation from the extended one. Finally, n values higher than 30 are also less informative because, especially for small structural domains, there are too few distances available (data not shown).

As *PRIDE* represents the probability of two structures represented by their C^α-distance distributions being identical, it follows that its value is between 0 and 1 ($0 \leqslant PRIDE \leqslant 1$) and that $1-PRIDE$ is, by definition, the probability of the two structures being different.

The metric properties of a structural similarity measure are important for clustering and for evolutionary studies. For M to be a metric distance between two structures, X and Y, the following criteria have to be fulfilled: (i) $M(X,Y) \geqslant 0$, the equality holding if, and only if, $X=Y$. The equality is by definition true both for *PRIDE* and for $1-PRIDE$. (ii) $M(XY)=M(YX)$ (symmetry). In fact, the symmetry property should follow from the definition of *PRIDE* but the calculation (in particular, the merging of histogram bins that contain less then 5 % of the data) depends on the serial order of the structures compared. This formula was checked for *PRIDE* (and $1-PRIDE$) on the same randomly selected dataset and was found to hold with an accuracy of 0.0001. (iii) $M(XY)+M(YZ) \leqslant M(XY)$ (triangular inequality). This inequality was checked for $1-PRIDE$ on the same dataset and found to hold with a maximal deviation of 0.00013. As the PRIDE values are given with a precision of two digits, we conclude that the PRIDE score conforms to a metric distance within the limits of its numerical accuracy.

The PRIDE score as a measure of close and distant similarities

The degree of similarity between a pair of protein structures is routinely measured with the rmsd between equivalent atoms computed after optimal superposition of the two three-dimensional structures. The relationship between *PRIDE* and rmsd was determined by an analysis of 230 NMR protein structure ensembles deposited in the Protein Data Bank (see Data and Methods).[14] This dataset was selected in order to represent close structural similarities, and also because the ensembles of

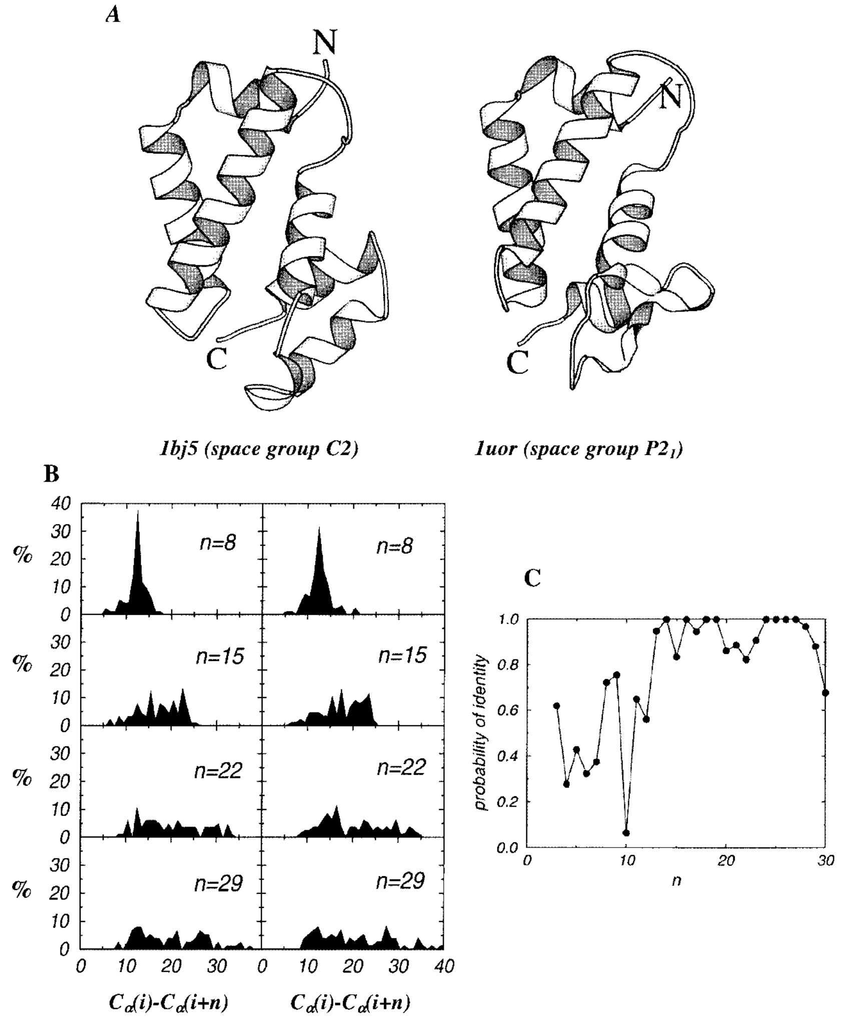


Figure 1. The crystal structure of the N-terminal domain of human serum albumin has been determined in different space groups. (a) MOLSCRIPT[11] view of the structures in the $C2$ and $P2_1$ space groups. (b) Distributions of the $C^{\alpha}(i)$-$C^{\alpha}(i+n)$ distances; n is the number of residues intercalated between the residue pair for which the distance is computed; values of $n = 8$, 15, 22 or 29 are given as examples. (c) Probability of identity for the 28 histogram pairs ($3 \leqslant n \leqslant 30$) as a function of n. The *PRIDE* value is the arithmetic average of the 28 values.

closely related structures can be aligned without difficulties. Care was taken to exclude high-value sequence similarities that might bias the results. As expected, *PRIDE* values decreased as rmsd increased, and a strong correlation was apparent (Figure 2); the relation between *PRIDE* and rmsd is apparently not linear. We note that by definition, *PRIDE* = 1.00 if rmsd = 0.00.

It is a well-known problem that optimal rmsd values cannot be easily determined for structures that are not closely related, because regions that can be structurally aligned may comprise only a part of the structures compared. Such an example is the classification of 28 NAD-binding domains.[15] While the similarity of these NAD-binding domains is apparent to the human eye, the regions that can be structurally aligned can be as low as 20 %.[15] As the rmsd values are known to vary with the length of the structural alignment,[16–18] an alternative strategy, based on a computer-intensive dynamic structural-alignment algorithm and a novel structural-similarity measure, had to be developed.[15] The folds were thus classified into four groups that coincide with their biological function (Figure 3(b)) such as the dihydrofolate group (1dlr, 8dfr, 1drh, and 3dfr), the glutathione group (1geu, 1lvl, 2npx, and 1typ), the glyceraldehyde group (1gd1, 1hdg, and 1gga), and the alcohol group (the remaining 17 domains in Figure 3(b)).[15] Using the PRIDE scores, on the other hand, one could produce a virtually identical classification (Figure 3(a)), without any need of structural alignments, indicating that the PRIDE score might be used in a broader range of similarities than rmsd values. The PSD index (protein structural distance) recently published by Yang & Honig[19] is also applicable in a broad range of similarities; however, PSD, as with other protein similarity measures,[1] is based on structural alignments.

In order to test the PRIDE score on a wider range of similarities we used the CATH database[4] as a model system. In the CATH fold-classification scheme, seven hierarchically organized labels (C, A, T, H, S, N and I) are assigned to each domain. The highest and broadest level of classification is the Class (label C), defined by the secondary-structure composition. The second, the Architecture (label A), is determined by the overall arrangement of the secondary-structural elements. The connectivity between the secondary-structural elements is

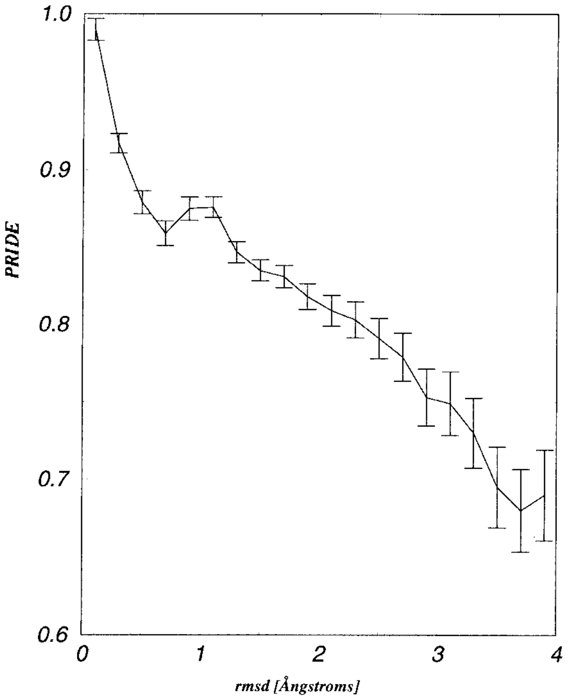


Figure 2. Relationship between the PRIDE and rmsd values. A total of 54,533 rmsd and *PRIDE* comparisons were made between the members of the 230 NMR structure assemblies listed in Table 3.

taken into consideration at the third level, the Topology (label T), while the fourth, the Homologous Superfamily (label H), has a definition based on a combination of three-dimensional and sequence similarity and clusters together homologous folds that share a common ancestor. The fifth classification level, the Sequence Family (label S), discriminates structures within each H level according to their sequence similarity (higher than 35 %) and therefore clusters together protein folds associated with highly similar functions. The last two hierarchical classification levels, the Nearly Identical Family (label N) and the Identical Family (label I), further subdivide each S-level according to the sequence similarity, which reaches 95 % and 100 %, respectively. The domain definitions and the classification are available at ftp.biochem.ucl.ac.uk/pub/cathdata/v2.0/(files domall.v2.0 and cath-s.list.v2.0). As it is apparent from the foregoing description, the CATH system is a full scheme of classification based on a vast amount of human knowledge, while PRIDE is a simple similarity measure, presently not optimised for fold classification. In fact, the goal of our analysis was to establish whether or not the PRIDE score is in agreement with the general principles of the CATH classification scheme. For this analysis we carried out an "all against all" comparison on the CATH database using the PRIDE score (68,817,241 structural comparisons).

Table 1 shows a summary of these comparisons organized according to the CATH labels. The folds grouped into the same I group (identical sequence, i.e. all the seven CATH labels, C, A, T, H, S, N and I, are identical) produce a *PRIDE* average score of 0.97, while those within the same N group but within a different I group (nearly identical sequences, i.e. six CATH labels, C, A, T, H, S and N, are identical and the seventh, I, is different) produce an average score of 0.93. In the same way, the broader the structural category, the lower the average PRIDE score calculated between the members, and the higher the variance of the score. In other words, *PRIDE* is in a good general agreement with the CATH categories throughout the entire dataset, i.e. throughout the entire range of fold similarities.

Another way to evaluate the PRIDE score is to use it as a "nearest neighbour classifier" for the CATH categories, i.e. to check whether or not the highest PRIDE score obtained for a given domain

Table 1. PRIDE values observed within various cluster groups of the CATH database

CATH label	CATH category	PRIDE score computed with members of the same category (average ± variance (no. of comparisons))	Nearest PRIDE neighbour within the same group[a] (%)
a	*b*	*c*	*d*
I	Identical representatives	0.97 ± 0.06 (72,517)	69.3
N	Nearly-identical representatives	0.93 ± 0.12 (75,429)	87.3
S	Sequence family	0.56 ± 0.22 (800,060)	98.5
H	Homologous superfamily	0.44 ± 0.19 (1,071,841)	98.8
T	Topology	0.35 ± 0.17 (2,568,350)	99.0
A	Architecture	0.26 ± 0.19 (9,028,969)	98.9
C	Class (α, β, α/β or few sec. str.)	0.26 ± 0.20 (24,547,471)	99.5
None	Different Class	0.11 ± 0.12 (68,817,241)	100.0

The fold entries of the CATH database were compared with each other within each group that belongs to the category. Taking line C as an example, the comparisons were made separately in the groups "α", "β", "α/β" or "few secondary structures", average and variance (column *c*) were computed with the exclusion of the PRIDE scores of domain pairs with identical A label.

[a] The nearest neighbour (column *d*) denotes the domain having the highest similarity to a given domain, based on the PRIDE score. A score in the Class = α group was considered positive if the highest PRIDE score pointed to a protein which was also member of the Class = α group.

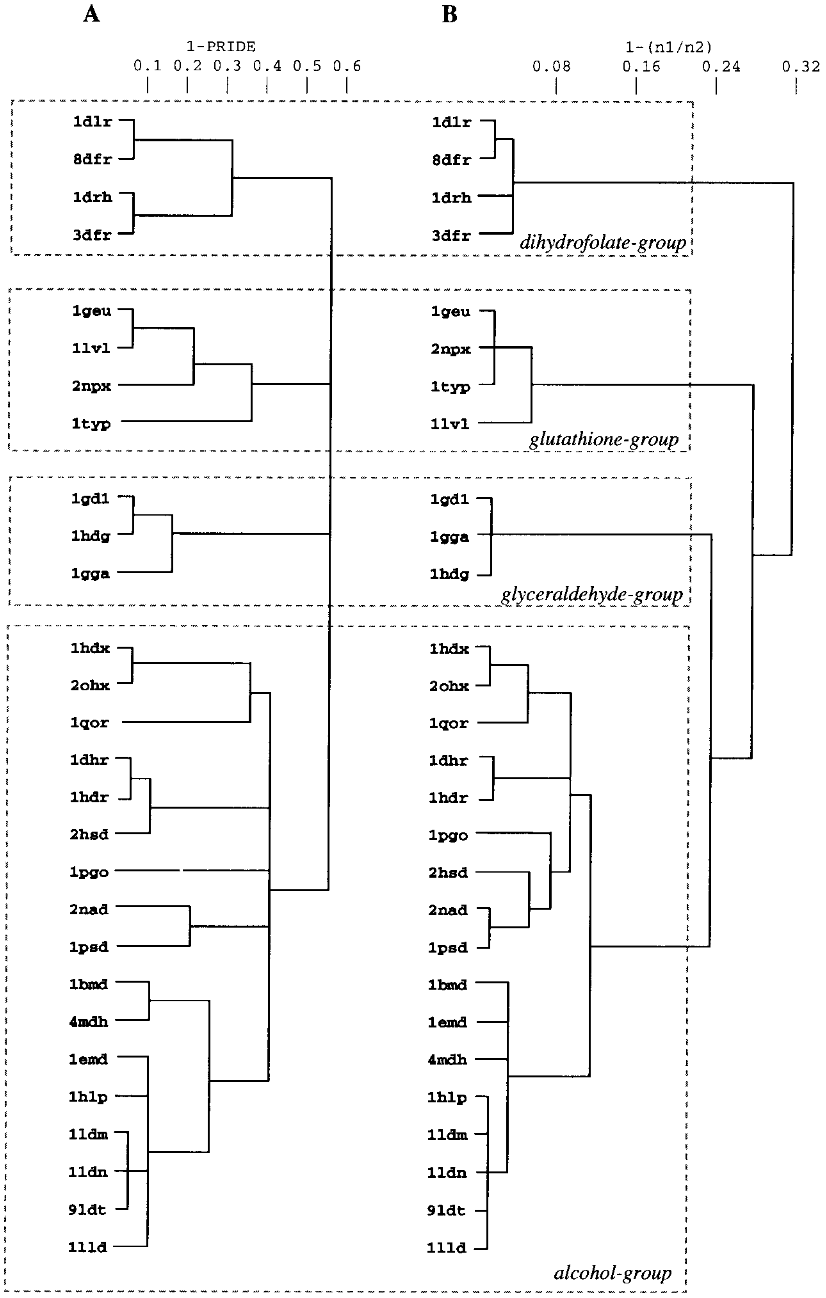


Figure 3. Graphs describing the cluster analyses of 28 NAD-binding domains carried out by a hierarchical agglomerative algorithm coupled with a single linkage similarity criterion.[30] (a) Cluster analysis based on PRIDE. The proximity matrix elements measuring the similarity between each pair of domains were equal to $1 - PRIDE$, where $PRIDE$ was the probability of identity between the two domains. (b) Cluster analysis based on the optimal superposition of the C^{α} atoms of each pair of domains; each proximity matrix element was defined as $100 - 100(n1/n2)$, where $n1$ is the number of residues structurally equivalent and $n2$ is the maximum number of residues that can be structurally equivalent, i.e. the number of residues in the structure with the shorter sequence over the two compared. The NAD-binding domains were taken from reference 15: 1dlr, dihydrofolate reductase (*Homo sapiens*); 8dfr, dihydrofolate reductase (*Gallus gallus*); 1drh, dihydrofolate reductase (*Escherichia coli*); 3dfr, dihydrofolate reductase (*Lactobacillus casei*); 1geu, glutathione reductase (*E. coli*); 1lvl, dihydrolipoamide dehydrogenase (*Pseudomonas putida*); 2npx, NADH peroxidase (*Streptococcus faecalis*); 1typ, trypanothione reductase (*Crithidia fasciculata*); 1gd1, D-glyceraldehyde-phosphate dehydrogenase (*Bacillus stearothermophilus*); 1hdg, D-glyceraldehyde-phosphate dehydrogenase (*Thermotoga maritima*); 1gga, D-glyceraldehyde-phosphate dehydrogenase (*Trypanosoma brucei brucei*); 1hdx, alcohol dehydrogenase (*H. sapiens*); 2ohx, alcohol dehydrogenase (*Equus caballus*); 1qor, quinone oxidoreductase (*E. coli*); 1dhr, dihydropteridine reductase (*Rattus norvegicus*); 1hdr, dihydropteridine reductase (*H. sapiens*); 2hsd, 3-α,20-β-hydroxysteroid dehydrogenase (*Streptomyces hydrogenans*); 1pgo, 6-phosphogluconate dehydrogenase (*E. coli*); 2nad, formate dehydrogenase (*Methylotrophic bacterium pseudomonas*); 1psd, D-3-phosphoglycerate dehydrogenase (*E. coli*); 1bmd, malate dehydrogenase (*Thermus flavus*); 4mdh, malate dehydrogenase (*Sus scrofa*); 1emd, malate dehydrogenase (*E. coli*); 1hlp, malate dehydrogenase (*Haloarcula marismortui*); 1ldm, lactate dehydrogenase (*Squalus acanthias*); 1ldn, lactate dehydrogenase (*B. stearothermophilus*); 9ldt, lactate dehydrogenase (*S. scrofa*); 1lld, lactate dehydrogenase (*Bifidobacterium longum*).

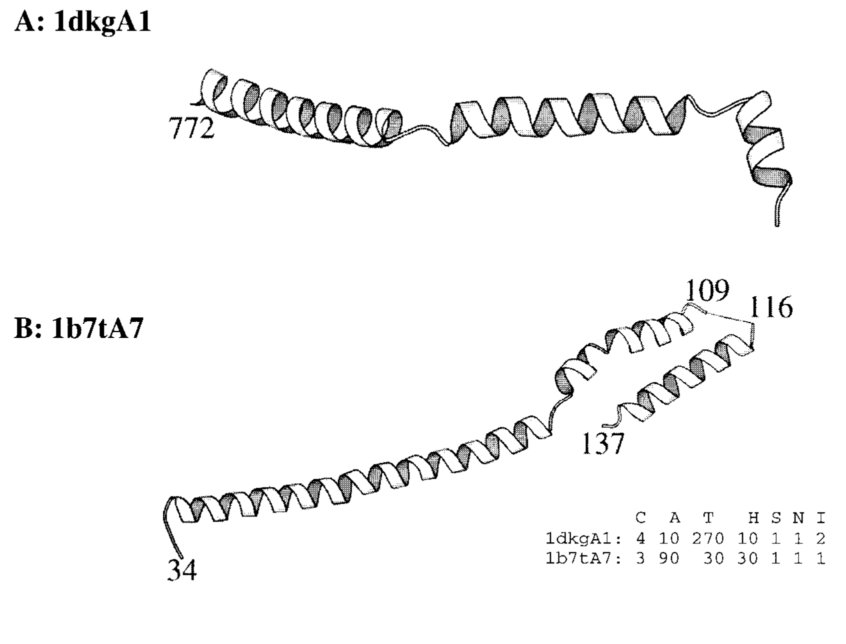
structure points to the correct category. The data in column *d* of Table 1 show that in 69 % of the cases the closest structure was labelled with the same seven classes, C, A, T, H, S, N and I. Most of the incorrect similarities were within the next level of similarity, i.e. with domains sharing only the first six CATH labels. This value increased to 87 % by allowing the last label to vary, and to 98 % by considering only the first five classes, C, A, T, H, and S, and ignoring the identity/difference of the last two. In other words, PRIDE can distinguish the S and H classes with an almost 99 % accuracy. We consider this result remarkable for two reasons. First, PRIDE was not optimised or tuned in any respect for recognizing the categories. Second, the 99.5 % success rate in the broadest categories (C) points to differences between the automatic categorization based on PRIDE nearest neighbours and the knowledge-based categorization of the CATH database. For example, out of the 67 CATH structures differently categorized by PRIDE, 45 (67 %) were preliminary annotations and seven (10 %) fell into the "few secondary structures" ($C = 4$) category. An examination of all the cases would go beyond the scope of this paper, but Figure 4 shows a typical example. The two domains 1b7tA7 (myosin heavy chain from *Aequipecter irradians*, residues 771-835) and 1dkgA1 (nucleotide-exchange factor from *Escherichia coli*, residues 34-137) have different CATH annotations (their C, A, T, H, S, N and I labels are 4, 10, 270, 10, 1, 1, 2, and 3, 90, 20, 20, 1, 1, 1, respectively); on the other hand, their overall structures are apparently quite similar (Figure 4) which is also reflected by the relatively high *PRIDE* value (0.31) determined between them.

A cluster-analysis dendrogram of 45 domains randomly selected from the CATH database is shown in Figure 5. The domains were selected so as to fall into three groups (with 15 domains in each group), represented by the following labels: group 1: $C = 1$, $A = 10$, and $T = 150$; group 2: $C = 2$, $A = 30$, and $T = 30$; and group 3: $C = 3$, $A = 10$, and $T = 20$. The dendrogram produced on the basis of $1 - PRIDE$ shows that the domains having different secondary structure (C label) are clearly distinguished. There is a clear discrimination also at the lower levels. For example, the domains with $C = 1$, $A = 10$, $T = 150$, $H = 20$, $S = 1$, and $N = 3$ are separated from those having identical C, A, T, H and S labels and $N = 1$.

Conclusions

The distribution of the $C^{\alpha}(i)$-$C^{\alpha}(i + n)$ distances seems to be a simple and useful description of fold geometry. The distance distributions can be compared *via* a straightforward and robust statistical technique, contingency-table analysis. This comparison yields a probability value for the two distributions being identical, the PRIDE score, the value of which is intrinsically normalized between 0 and 1. The PRIDE scores possess the basic metric properties, and were shown to produce the correct classification of folds over a wide range of similarities. For closely related protein 3D structures PRIDE is highly correlated with the rmsd values calculated between the main-chain C^{α} atoms. For more distant similarities, such as those included in the CATH database, PRIDE scores produced classifications that were in very good agreement with the known categories. PRIDE scores can be calculated for proteins of any size, but in order to test for the presence of a fold within a larger protein, the query structure needs to be divided into domains.

The PRIDE approach is clearly based on a simplified representation of the three-dimensional structure. Several alternative, simplified representations have been reported (such as the C^{α}-C^{α} distance plot) as well as various methods based on such simplified representations.[20–22] The method of Sippl[21] is conceptually related to that presented here, since it considers the distances between C^{α} atoms separated by a variable number, n, of residues. Nevertheless, PRIDE values are obtained by a completely different procedure, i.e. the distributions of these distances rather than their actual values are used. As a consequence, PRIDE can be much more easily applied to protein structures with a very modest degree of similarity. More recently, several fast methods have been proposed for the detection of protein-protein similarity.[23] These methods represent a protein structure as an ensemble of secondary-structural elements, and the orientation and the topology of the elements are then compared with different techniques. The



	C	A	T	H	S	N	I
1dkgA1:	4	10	270	10	1	1	2
1b7tA7:	3	90	30	30	1	1	1

Figure 4. Two domains from the CATH database: 1b7tA7 (myosin heavy chain from *Aequipecter irradians*, residues 771-835) and 1dkgA1 (nucleotide-exchange factor from *E. coli*, residues 34-1367) 1dkgA1 (b) is the domain most similar to 1b7tA7 (a), on the basis of the PRIDE value ($PRIDE = 0.31$). Despite the fact that they are differently classified in the CATH database (their C, A, T, H, S, N, and I labels are 4, 10, 270, 10, 1, 1, 2, and 3, 90, 20, 20, 1, 1, 1, respectively), the close similarity of these two domains is apparent.

advantage of the PRIDE approach over these alternative methods is that for the calculation of PRIDE, the secondary structure does not need to be determined. This is an important difference, since secondary structures can be defined in various ways,[24] and since the secondary structure assignments may strongly depend on the experimental method (e.g. results from NMR and X-ray diffractions are often very different) as well as on the crystallographic resolution.

A potential limitation of PRIDE could be that proteins with similar secondary-structural arrangements but different fold might be confused, due to the fact that (at least in the present version of PRIDE) only C^{α}-C^{α} distances between residues less than 30 residues are considered. In our experience, the resulting "background noise" is very limited; on the other hand it can also be extremely useful. For example, two apparently unrelated proteins, the PH domain of the human insulin receptor substrate 1 (1qqg, residues 12-114 of chain A)[25] and a segment containing residues 136-567 of chain A of *Paracoccus denitrificans* cytochrome cd1 nitrite reductase (1e2r)[26] were found to exhibit a PRIDE score of 0.66, which is indicative of considerable structural similarity (Table 1). The latter protein domain (1e2r) adopts a β-propeller fold, with eight blades formed by four-stranded antiparallel β-sheets, so it is clearly much larger and, at a first glance, different from the small PH domain. A closer inspection reveals nevertheless that each β-sheet of the β-propeller is intercalated between the preceding and the subsequent β-sheet in such a manner that one pair of four-stranded β-sheets in 1e2r is in fact very similar to the two β-sheets of the PH domain 1qqg (Figure 6). The PH domain could thus be considered as a subunit of the β-propeller fold.

The most important advantages of the PRIDE score are as follows. (i) As it requires no structural alignment, it is very fast to compute and therefore it can be used for scanning and/or classifying even very large structural databases. (ii) It provides a probability value that can also be appreciated by

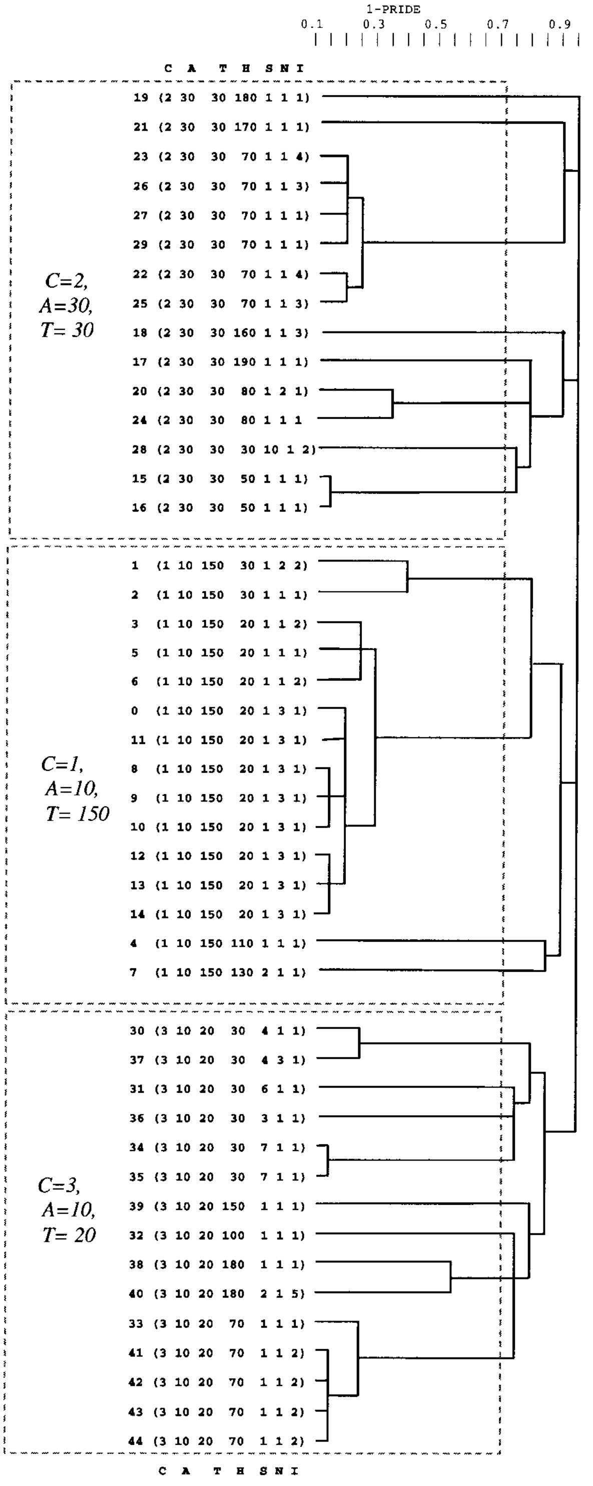


Figure 5. Classification of 45 three-dimensional domains based on the PRIDE scores. Each domain is identified by a number and, in parentheses, by its C, A, T, H, S, N, and I labels taken from the CATH database. The domains, numbered from 0 to 44, are defined below with their identification codes from the Protein Data Bank and their residue ranges. 0, bpx(92A-148A); 1, 1bvs(66D-148D); 2, 1cuk(67-142); 3, 1rpl(95-148); 4, 1zqf(9A-91A); 5, 1zqu(91-148); 6, 2bpf(92A-148A); 7, 5crx(20B-130B); 8, 7ici(92A-148A); 9, 7icp(92A-148A); 10, 7ict(92A-148A); 11, 8icf(92A-148A); 12, 8icn(92A-148A); 13, 8icr(92A-148A); 14, 9ica(92A-148A); 15, 1ahj(111B-212B); 16, 1ahj(111H-212H); 17, 1bcm(492B-559B); 18, 1bi1(175-222); 19, 1bia(271-317); 20, 1bkb(4-74); 21, 1c0 m(216C-269C); 22, 1d0y(33A-80A); 23, 1d1b(33A-80A); 24, 1eif(4-73); 25, 1lvk(33-80); 26, 1mmg(33-80); 27, 1mne(33-80); 28, 1qqg(159A-262A); 29, 1vom(33-80); 30, 1alo(1-74); 31, 1c4c(1A-76A); 32, 1div(60-149), 33, 1lgr(1-103); 34, 1qla(1E-106E); 35, 1qlb(1E-106E); 36, 2pia(226-321); 37, 1qj2(3A-79A); 38, 1bml(12C-147C); 39, 1bml(289C-371C); 40, 1bml(151D-284D); 41, 1f52(1A-103A); 42, 1f52(1C-103C); 43, 1f52(1E-103E); and 44, 1f52(1G-103G). The classification was based on a hierarchical agglomerative algorithm coupled with a single linkage similarity criterion.[30] The proximity matrix elements measuring the similarity between each pair of domains were equal to 1 – *PRIDE*, where *PRIDE* was the probability of identity between the two domains.

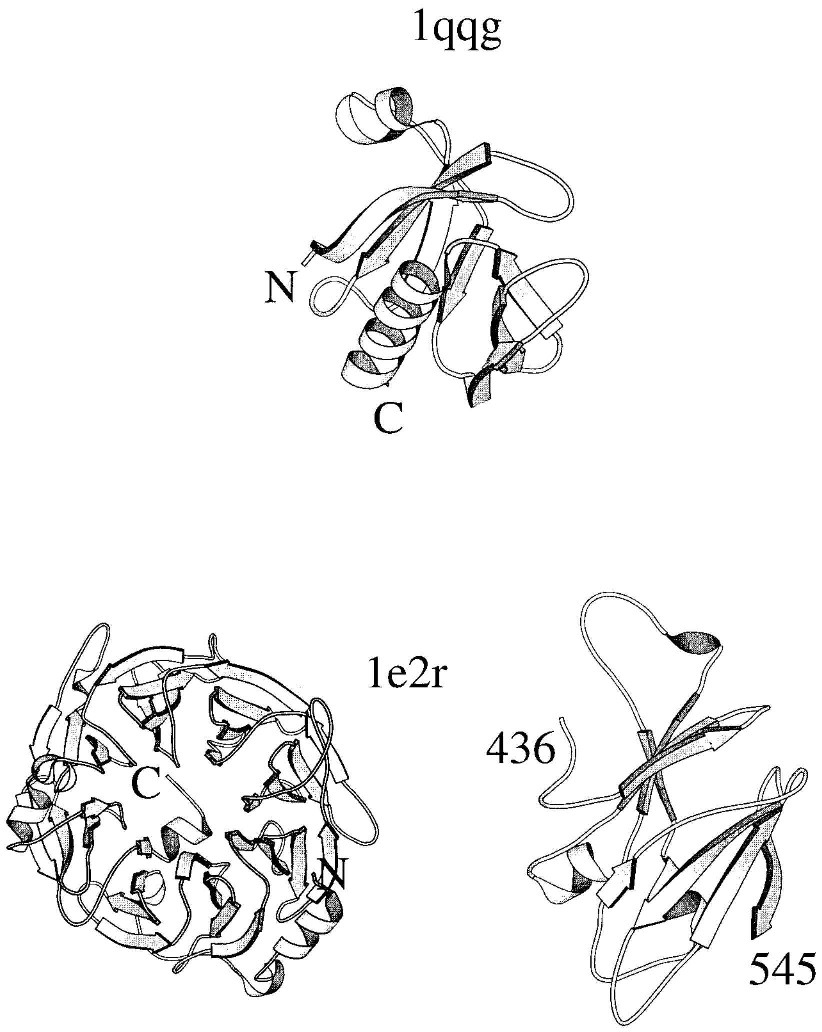


Figure 6. The PH domain 1qqg is surprisingly similar to a β-propeller fold (1e2r) on the basis of the PRIDE value (0.66). Actually, two adjacent β-sheets of the latter fold (residues 436-545) compare quite well with the sandwich of the two β-sheets of the PH domain, which can therefore be seen as a subunit of the larger β-propeller fold.

non-specialist users. (iii) For the purposes of classification, PRIDE scores provide a useful alternative to rmsd scores, since they provide a meaningful classification even for distantly related structures that cannot be unequivocally superposed. (iv) The only information needed to compute the PRIDE values is given by the positional parameters of the C^{α} atoms. Knowledge-based qualitative decisions, such as the assignment of secondary structure, are not necessary, and the sequence information is not used in the classification. The latter feature allows one to study the conservation of structure and sequence separately.

In the present, non-optimised implementation of PRIDE, the fold-database (CATH, about 10,000 domains) is pre-computed into distance histograms, which are stored on the disk. Each domain query is first translated into a set of distance histograms, and then PRIDE is calculated between the query and all members of the fold database; finally, the results of the comparison are written on the disk. Much of the real time is thus spent on input/output operations, and the complete time requirement will depend on the size of the database. Representative figures are given for a Silicon Graphics workstation equipped with an SGI R10000 CPU (Figure 7). If there are ten structures, a time of 0.15 second is required for computing and storing 100 *PRIDE* values. For 200 structures, a total time of 37.13 seconds is sufficient to calculate and store 40,000 *PRIDE* values, i.e. the time requirement is proportional to the square of the number of structures. For a database of 10,000 domain structures, about half a day is thus required to compute and store the *PRIDE* values associated with 49,995,000 structure comparisons. Once the database is stored on the disk, the evaluation of individual domain queries is quite fast: we estimate that about 1000 domains can be scanned in one second using an SGI R10000 CPU.

The limitations of the PRIDE score follow from its experimental nature. *PRIDE* values are presently calculated as a simple arithmetic average of the 28 probability values resulting from the comparison of all the histograms. However, it is likely that individual histograms carry different information, which might in turn influence the classification. Weighted averaging procedures can be

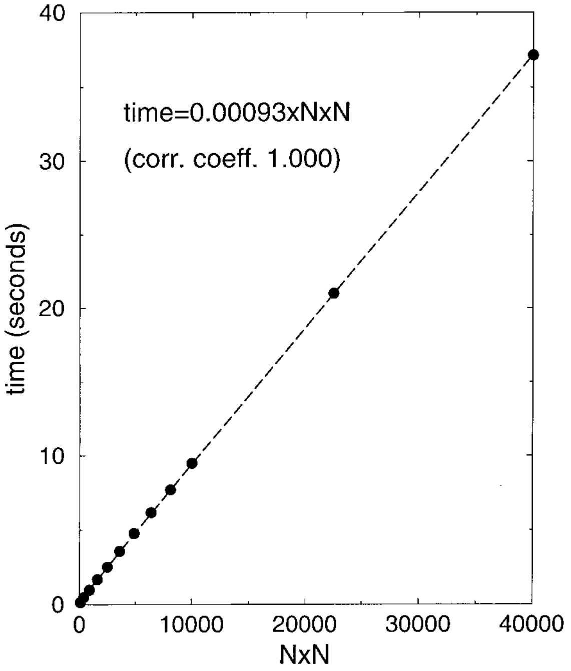


Figure 7. Time required with an SGI R10000 processor to read the 28 histograms of N domain structures, compute $N \times N$ *PRIDE* values, and store them in a file. The time increases linearly with the square of N (equation written at the top left).

developed depending on the goal of the analysis; furthermore, the entire probability *versus* n plot (Figure 1(c)) can be used as a graphic description of the similarity.

Materials and Methods

Structural data

The protein structures were taken from the Protein Data Bank[14] and from CATH.[4] A randomly selected, non-redundant dataset of protein domains was created for the testing of the metric properties of PRIDE as follows: 10 % of the PDB files were randomly selected by taking the 10th, 20th, 30th, etc., file from an alphabetical list. All structures were then subdivided into structural domains with the DomainParser algorithm[27] and the first domain in the protein sequence was retained if it was longer than 40 residues. A set of 869 structural domains (377,147 unique pairs) generated in this manner was used for checking the metric properties of PRIDE and/or (1 − *PRIDE*). For the determination of the *PRIDE versus* rmsd relationship, 230 NMR model ensembles were chosen from the PDB (Table 2). In order to avoid potential statistical biases, a maximal sequence identity of 25 % was allowed, as determined with PDB_SELECT.[28] Within each ensemble of NMR models, each model was compared to all the others. In total, 54,533 comparisons of model pairs were thus performed. The rmsd values were computed after optimal superposition of the equivalent C^{α} atom pairs carried out using the method of Kabsch.[29]

Calculation of the PRIDE score

The contingency-table analysis is a statistical method used to ascertain if two samples represented by their histograms come from the same population. One of its important features is that it can also be applied if there are no analytical models describing the distribution of the population.[10] The calculation is illustrated by the comparison of two histograms containing the $C^{\alpha}(i)$-$C^{\alpha}(i+8)$ distances of two structures, denoted as structure 1 and structure 2, respectively (Table 3). Given two histograms of m bins written as obs(1,1), obs(2,1), obs(3,1)...obs(m,1) for structure 1 and obs(1,2), obs(2,2), obs(3,1)...obs(m,2) for structure 2, one can calculate the expected value of each observation as:

$$\exp(i,j) = \frac{\mathrm{obs}(x,i)\mathrm{obs}(j,x)}{\mathrm{obs}(xx)} \qquad (1)$$

where obs(x,i) is the sum of the observations of structure 1 (column sum). In the example, obs(x,i) = 100, because of the use of percentages; the normalization is otherwise not a requirement. The obs(j,x) value is the sum of the ith observations in the two histograms (row sum, written in column g), and obs(x,x) is the total number of the m observations in the two histograms (in the example it equals 200 because of the normalization to percentages). The following χ^2 value can then be computed:

$$\chi^2 = \sum_{j=1}^{2}\sum_{i=1}^{m}\frac{[\mathrm{obs}(i,j) - \exp(i,j)]^2}{\exp(i,j)} \qquad (2)$$

and the probability of the two distributions being identical can be read from the corresponding χ^2 distribution of $m-1$ degrees of freedom. Care must be taken that none of the obs(i,j) values should fall below 5%. In the example (Table 3 and Figure 8), and throughout all the calculations, the initial histograms of the C^{α}-C^{α} $(i+n)$ distances were calculated with a bin width of 0.5 Å; then, starting from the smallest values, the bins were combined so that at least 5 % of the observations were included in each bin (Figure 8(c) and (d)). Table 3 shows the observed percentages for two structures after bin width adjustment as an example. Then the expected values are computed according to equation (1); for example, the expected percentage of distances within the 15.5-17.5 Å range for structure 1 is given by 15.4 × 100/200 = 7.7. The χ^2 value calculated according to equation (2) is 6.89, and this corresponds to a probability of identity of 0.72, since in this example there are ten degrees of freedom (the number of freedoms equals the number of histogram bins after bin-width adjustment). All 28 histogram pairs were processed in an analogous way, the 28 probability values were plotted as a function of n $(3 < n < 28)$, as shown in Figure 1(c) and the average was computed and given as the *PRIDE* score. A precision of two digits is used (e.g. 0.54 or 54 %).

Cluster analysis

The domains were classified by cluster analysis with a hierarchical agglomerative algorithm coupled with a single linkage similarity criterion;[30] the elements of the square proximity matrix were set to 1 − *PRIDE*. Each domain pair was thus discriminated by its *PRIDE* value. The classification in Figure 3(b) is taken from reference 15.

Table 2. Identification codes of the protein structures used here

```
1cg7(A)(30),1ej5(A)(20),1cfp(A)(25),1eit( )(10),1eio(A)( 5),1cfh( )(15),
1cfe( )(20),3leu( )(19),1bqv( )(28),1bqz( )(20),1ayg( )(20),1ef4(A)(20),
1ego( )(20),1rip( )( 6),1enw(A)(20),1eo0(A)(10),1erd( )(20),1rfa( )(30),
1rou( )(22),1rof( )(10),1ax3( )(16),1awo( )(20),1cf4(B)(20),3msp(A)(20),
3ncm(A)(20),1tba(A)(25),1suh( )(20),3ctn( )(30),3crd( )(15),1cis( )(15),
1e0l(A)(10),1chl( )( 7),1eci(B)(20),2cbh( )(41),1ed7(A)(30),1spf( )(20),
1sro( )(20),1b22(A)(30),3gcc( )(46),1fdm( )(20),1qmc(A)(42),1fct( )(27),
1aq5(A)(20),3rpb(A)(20),1ap7( )(20),1ap0( )(26),1aoy( )(23),1apf( )(20),
1apq( )(19),1apj( )(21),1fht( )(43),1qlo(A)(19),1aty( )( 9),1ce4(A)(20),
1qu5(A)(16),1auz( )(24),1esk(A)( 9),1auu(A)(10),1r63( )(20),1r2a(A)(17),
1ccv(A)(20),2def( )(20),1cdb( )(18),2pcf(B)(10),1qp6(A)(16),2ctn( )(30),
1qr5(A)(10),1wkt( )(20),1d8b(A)(15),1bf8( )(20),1d8j(A)(20),1wjb(A)(40),
1bgk( )(15),1bhi( )(20),1cn7(A)(20),1cmr( )(18),1cmo(A)(43),2sh1( )( 8),
1df6(A)(16),1bnx(A)(21),1dec( )(25),1ddb(A)(20),1bei( )(20),1beg( )(18),
2a93(B)(40),2a93(A)(40),1bk8( )(25),1bmw( )(38),1zto( )( 8),1cpz(A)(20),
2ptl( )(21),1cou(A)(18),1bm4(A)( 9),1cok(A)(18),2abd( )(29),1bmx( )( 8),
1cwx(A)( 4),1yuj(A)(50),1yua( )(26),1zfo( )(20),1zfd( )(45),1ztn( )( 8),
1zta( )(20),1bmy( )(10),1co4(A)(15),2prf( )(19),1zaq( )(12),1tpn( )(28),
1trl(A)( 8),1dv9(A)(21),2bi6(H)(18),2bds( )(42),1tvt( )( 6),1tsk( )(30),
1tnn( )(16),1dz1(A)(16),1dxz(A)(20),1tfi( )(12),1cjg(A)(11),1e01(A)(20),
1tle( )(14),3alc(A)(17),1b6f(A)(23),3bbg( )( 2),1dip(A)(10),2tbd( )(30),
1boe(A)(20),1bal( )(56),1bak( )(20),2bby( )(30),1dgz(A)(38),1vib( )(20),
1bc4( )(15),1cl4(A)(12),1dny(A)(21),1b8w(A)(20),2u1a( )(20),1dlx(A)(15),
1ba5( )(18),1ba6( )(10),1dk2(A)(25),1ckv( )(14),1b9u(A)(10),1b9r(A)(15),
1b9q(A)(19),1paa( )(10),1a6 s( )(20),1a66(A)(18),1a6b(B)(20),1a7 m( )(20),
1gyf(A)(16),1peh( )(10),1pba( )(20),1c4e(A)(20),1bvh( )(15),1hqi( )(12),
5znf( )(13),1hns( )(16),1ab3( )(26),2new( )(17),1pou( )(20),1pmc( )(36),
1pnh( )(25),1pnb(B)(10),1pnb(A)(10),1acz( )( 5),1prs( )(30),1abz( )(23),
1pft( )(25),1gnc( )(10),1pfl( )(20),4znf( )(41),2nbt(A)(10),1ghk( )(25),
1ghc( )(14),1hue(A)(25),1mut( )(15),1ncs( )(46),1c06(A)(16),1kjs( )(20),
1kla(A)(17),1ksr( )(20),1lre( )(20),2lef(A)(12),1by1(A)(20),1axh( )(20),
1bzg( )(30),1byv(A)(10),2jhb(A)(20),1iie(A)(20),1igl( )(20),1iml( )(48),
1irg( )(20),1imt( )(39),1c20(A)(21),2hp8( )(30),2hsp( )(20),1idz( )(20),
1iba( )(11),1ica( )(10),1jun(A)( 7),1jvr( )(20),1ngr( )(20),1joy(A)(21),
1khm(A)(20),2if1( )(29),1nkl( )(20),1qdp( )(20),1aiw( )(23),1fvl( )(18),
1qkh(A)(21),1akp( )(15),2fow( )(26),1agt( )(17),1pyc( )(15),2fmr( )(18),
1afo(A)(20),1qa5(A)( 2),1afp( )(40),1ah9( )(19),1qfq(B)(29),2nmb(A)(14),
1agg( )(24),1qk7(A)(20),1qk6(A)(10),1qky(A)(14),1qhk(A)(20),1qkl(A)(22),
1qey(A)(27),1adn( )(14)
```

Protein three-dimensional structures determined by NMR spectroscopy, deposited in the Protein Data Bank as model ensembles, used to determine the relationship between the *PRIDE* and the rmsd values. Each identification code is followed by the chain identifier and by the number of models in parentheses.

Table 3. Distribution of the C^{α}-$C^{\alpha}(i+8)$ distances for the two structures shown in Figure 1(a)

		Structure 1 (1bj5)		Structure 2 (1uor)		
Bin no. (i) *a*	Range (Å) *b*	Observed (%) *c*	Expected (%) *d*	Observed (%) *e*	Expected (%) *f*	Sum of observed percentages ($c+e$) *g*
7.2	14.4					
2	6.5-10.5	8.4	8.9	9.5	8.9	17.9
3	10.5-12.0	7.1	7.7	8.3	7.7	15.4
4	12.0-14.0	9.5	7.8	6.0	7.8	15.5
5	14.0-15.5	6.0	7.1	8.2	7.1	14.2
6	15.5-17.5	8.3	7.7	7.1	7.7	15.4
7	17.5-19.0	7.1	7.1	7.1	7.1	14.2
8	19.0-21.5	16.6	12.5	8.3	12.5	24.9
9	21.5-22.5	5.9	9.4	13.0	9.4	18.9
10	22.5-24.0	8.3	8.9	9.4	8.9	17.7
11	Over 24.0	15.6	15.8	15.9	15.8	31.5
	Sum:	100.0	-	100.0	-	200.0

The histograms were first computed with a bin width of 0.5 Å, then the bins were combined, wherever necessary, to ensure that they contained at least 5% of the observations (see Figure 6).

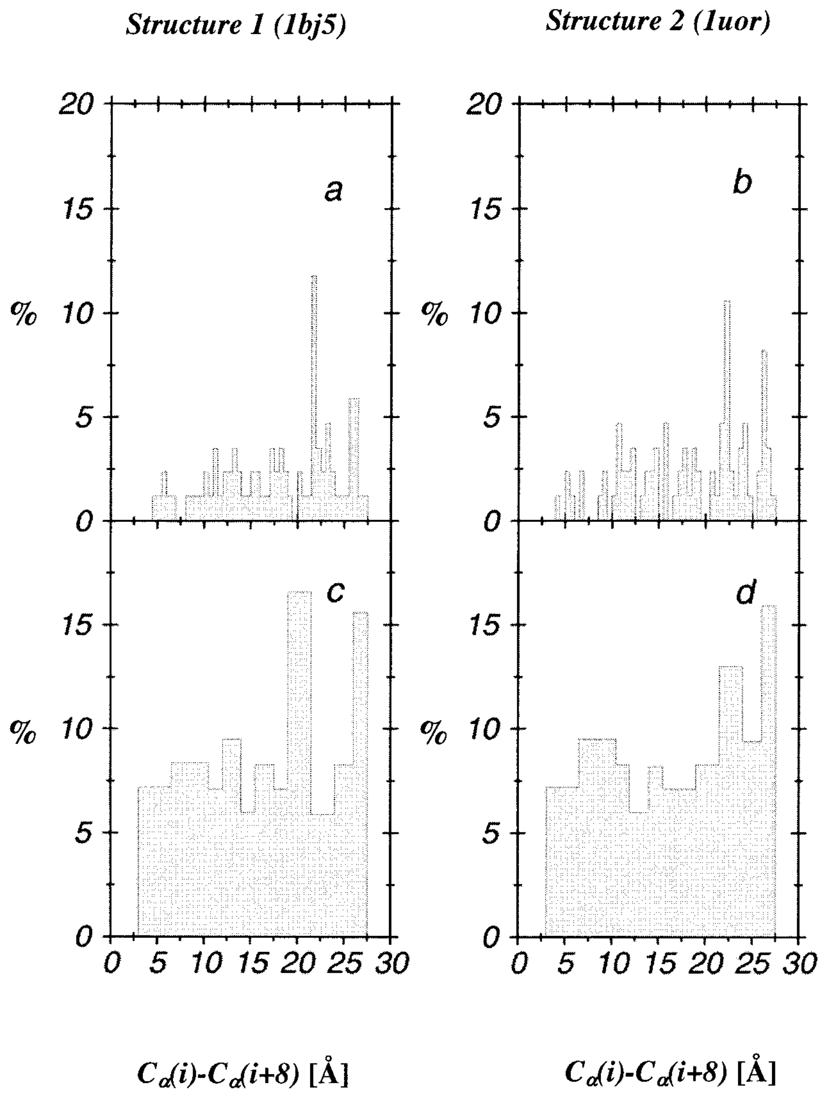


Figure 8. Histograms of $C^{\alpha}(i)$-$C^{\alpha}(i+n)$ distances determined for two structures of the N-terminal domain of human serum albumin, 1bj5 and 1uor (shown in Figure 1(a)). The histograms are first taken with a bin width of 0.5 Å, then the bins are combined so as to have at least 5% of the observations in each bin. The resulting histograms are numerically shown in Table 3.

Acknowledgements

The authors are grateful to Professor L. Patthy (Insitute of Enzymology, Budapest) or his comments.

References

1. Koehl, P. (2001). Protein structure similarities. *Curr. Opin. Struct. Biol.* **11**, 348-353.
2. Holm, L., Ouzunis, C., Sander, C., Tuparec, G. & Vriend, G. (1992). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691-1698.
3. Murzin, A., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP - a structural classification of protein databases for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.
4. Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH - a hierarichic classification of protein domain structures. *Structure,* **5**, 1093-1108.
5. Hadley, C. & Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure,* **7**, 1099-1112.
6. Terwilliger, T. C. (2000). Structural genomics in North-America. *Nature Struct. Biol.* **7**, 935-939.
7. Heinemann, U. (2000). Structural genomics in Europe: slow start, strong finish? *Nature Struct. Biol.* **7**, 940-942.
8. Yokoyama, S., Hirota, H., Kigawa, T., Yabuki, T., Shirouzu, M., Terada, T. *et al.* (2000). Structural genomics projects in Japan. *Nature Struct. Biol.* **7**, 943-945.
9. Gerstein, M. (2000). Integrative database analysis in structural genomics. *Nature Struct. Biol.* **7**, 960-963.
10. Dowdy, S. & Wearden, S. (1991). *Statistics for Research*, Wiley, New York.
11. Kraulis, P. J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.
12. Curry, S., Mandelkov, H., Brick, P. & Franks, N. (1998). Crystal structure of human serum albumin complexed with fatty acid reveals an asymmetric distribution of binding sites. *Nature Struct. Biol.* **5**, 827-831.

13. He, X. M. & Carter, D. C. (1992). Atomic structure and chemistry of human serum albumin. *Nature,* **358**, 209-213.
14. 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.
15. Carugo, O. & Argos, P. (1997). NADP-dependent enzymes. I: conserved stereochemistry of cofactor binding. *Proteins: Struct. Funct. Genet.* **28**, 10-28.
16. Carugo, O. & Pongor, S. (2001). A normalized root-mean-square distance for comparing protein three-dimensional structures. *Protein Sci.* **10**, 1470-1473.
17. Maiorov, V. N. & Crippen, G. M. (1995). Size-independent comparison of protein three-dimensional structures. *Proteins: Struct. Funct. Genet.* **22**, 273-283.
18. Irving, J. A., Whisstock, J. C. & Lesk, A. M. (2001). Protein structural alignment and structural genomics. *Proteins: Struct. Funct. Genet.* **42**, 378-382.
19. Yang, A. S. & Honig, B. (2000). An integrated approach to the analysis and modelling of protein sequences and structures. I. Protein structural alignment and a quantitative measure of protein structural distance. *J. Mol. Biol.* **301**, 665-678.
20. Nishikawa, K. & Ooi, T. (1974). Comparison of homologous tertiary structures of proteins. *J. Theor. Biol.* **43**, 351-374.
21. Sippl, M. J. (1982). On the problem of comparing protein structures. *J. Mol. Biol.* **156**, 359-388.
22. Orengo, C. A. (1992). A review of methods for protein structure comparison. In *Patterns in Protein Sequence and Structure* (Taylor, W. R., ed.), vol. 7, pp. 159-188, Springer-Verlag, Heidelberg.
23. Gibrat, J.-F., Madej, T. & Btyant, S. H. (1996). Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377-385.
24. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. & Mornon, J. P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantage of a consensus assignment. *Protein Eng.* **6**, 377-382.
25. Dhe-Paganon, S., Ottinger, E. A., Nolte, R. T., Eck, M. J. & Shoelson, S. E. (1999). Crystal structure of the pleckstrin homology-phosphotyrosine binding (PH-PTB) targeting region of insulin receptor substrate 1. *Proc. Natl Acad. Sci. USA,* **96**, 8378-8383.
26. Jafferji, A., Allen, J. W., Ferguson, S. J. & Fulop, V. (2000). X-ray crystallographic study of cyanide binding provides insights into the structure-function relationship for cytochrome cd1 nitrite reductase from *Paracoccus pantotrophus*. *J. Biol. Chem.* **275**, 25089-25094.
27. Xu, Y., Xu, D. & Gambow, H. N. (2000). Protein domain decomposition using a graph-theoretic approach. *Bioinformatics,* **16**, 1091-1104.
28. Hobohm, U. & Sander, C. (1994). Enlarged representative set of protein structures. *Protein Sci.* **3**, 522-530.
29. Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallog. sect. A,* **34**, 827-828.
30. Carugo, O. (1995). Use of the estimated errors of the data in structure-correlation studies. *Acta Crystallog. sect. B,* **51**, 314-328.

Edited by B. Honig

(Received 13 July 2001; received in revised form 12 October 2001; accepted 12 November 2001)