PROTEINS: Structure, Function, and Genetics 25:354–365 (1996)

Analysis of Topological and Nontopological Structural Similarities in the PDB: New Examples With Old Structures

Nickolai N. Alexandrov and Daniel Fischer*
Laboratory of Mathematical Biology, NCI, NIH, Frederick, Maryland 21701

ABSTRACT **We have developed a new method and program, SARF2, for fast comparison of protein structures, which can detect topological as well as nontopological similarities. The method searches for large ensembles of secondary structure elements, which are mutually compatible in two proteins. These ensembles consist of small fragments of C^{α}-trace, similarly arranged in three-dimensional space in two proteins, but not necessarily equally-ordered along the polypeptide chains. The program SARF2 is available for everyone through the World-Wide Web (WWW). We have performed an exhaustive pairwise comparison of all the entries from a recent issue of the Protein Data Bank (PDB) and report here on the results of an automated hierarchical cluster analysis. In addition, we report on several new cases of significant structural resemblance between proteins. To this end, a new definition of the significance of structural similarity is introduced, which effectively distinguishes the biologically meaningful equivalences from those occurring by chance. Analyzing the distribution of sequence similarity in significant structural matches, we show that sequence similarity as low as 20% in structurally-prealigned proteins can be a strong indication for the biological relevance of structural similarity.** © 1996 Wiley-Liss, Inc.*

Key words: classification of protein structures, method for structural comparison, porin, bacteriochlorophyll *a* protein, sequence-structure relationship, WWW

INTRODUCTION

Comparison of three-dimensional protein structures is an important tool in modern structural biology. Comparative analysis of protein structures can reveal hidden evolutionary relationships between nonhomologous proteins (TIM barrels[1]), may point to the possible location of the active site (α/β structures[2]), or hint at the origin of structural stability for proteins with a common core (globin fold in different proteins[3]). Methods for comparison of protein structures have been reviewed.[4,5] However, this problem of comparison remains open, since so far there is no universal method capable of finding all types of structural resemblance, mainly because of the ambiguous understanding of the term "similarity."

Different methods apply different measures of the significance of structural similarities. Several authors have measured significance of similarity as a function of the number of superimposable residues and the root mean square distance (rmsd) between them.[6–9] However, these similarity scores may not reflect the biologically-intuitive understanding of the significance of similarity. Two structural matches, having the same number of C^{α}-atoms and a similar rmsd, are not necessarily equally significant. For example, a significant similarity may consist of extended backbone fragments, while a less significant one may consist of relatively short segments or single residues. Another example can be observed when in one case large single β-sheets or two long α-helices (both are frequent motifs in protein structures) are matched, while in another case the same number of the C^{α}-atoms forms an interesting, unusual arrangement of backbone fragments. In this paper we use a more accurate statistical score to measure the significance of structural similarities.

From our previous experience in the comparison of protein structures,[10,11] we could see that the most interesting and biologically meaningful similarities are mainly composed from elements of secondary structure. Here we present an effective algorithm for detection of large ensembles of secondary-structure elements. The idea of representing a protein structure as a set of secondary-structure elements (SSEs) to search for their common spatial arrangement has been used in several methods of structural comparison.[12–14] However, the search in our method is carried out differently and sometimes leads to con-

Received October 30, 1995; revision accepted December 4, 1995.

Address reprint requests to Nickolai N. Alexandrov, Amgen, 1840 DeHavilland Dr., Thousand Oaks, CA 91320-1789.
*Current address: 201 Mol. Bio. Inst., P.O. Box 95170, University of California Los Angeles, Los Angeles, CA 90095-1570.

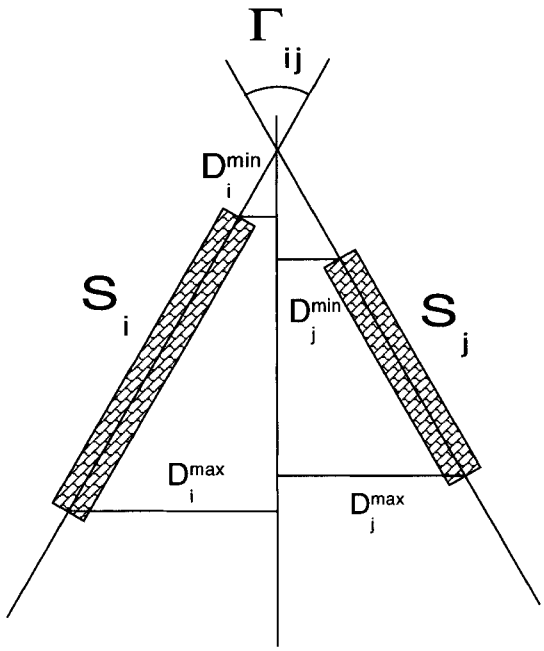

Fig. 1. Five parameters, determining compatibility of pairs of SSEs: angle Γ and four distances (see Methods, Search for Compatible Pairs of SSEs).

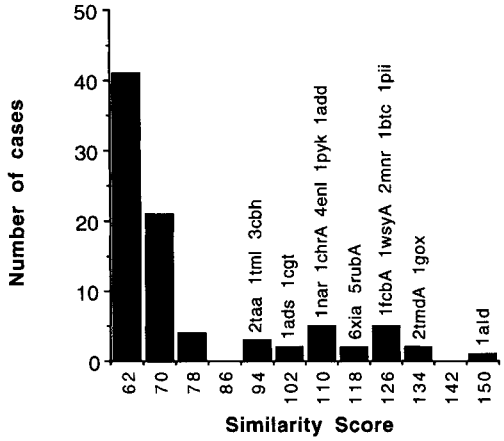

Fig. 2. Distribution of similarity score S between triosephosphate isomerase (5timA) and other proteins from our set. All other α/β-barrel structures are separated from the main peak of the distribution.

siderably different results. With our method we can detect "difficult" cases of structural similarities, when one of the SSEs is much larger than the corresponding one in another protein. We allow insertions and deletions in SSEs in structural alignment, and, as a result, we provide, if possible, a refined alignment with a large number of the C^{α}-atoms, superimposed with a small rmsd. Our program SARF2 (SARF stands for Spatial ARrangement of backbone Fragments) is available through the World-Wide Web (WWW) (URL address http:llwww-lmmb.ncifcrf.gov/~nicka/sarf2.html).

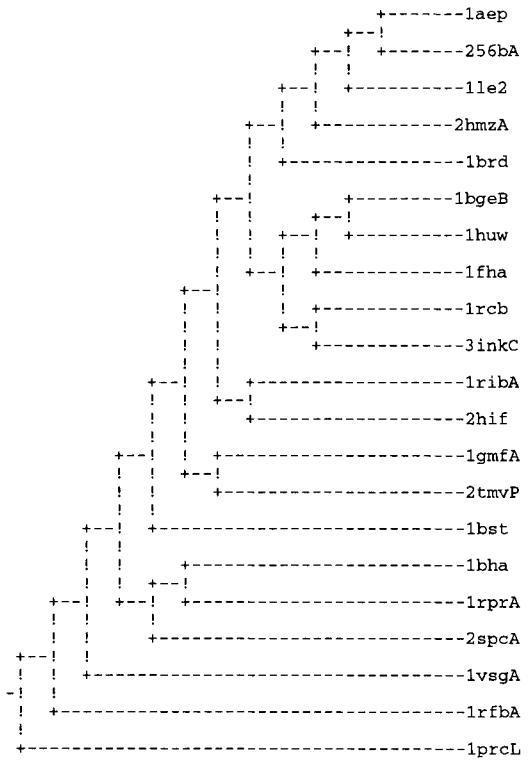

Fig. 3. α-helical bundles. This is one of the tightest subbranches in the tree. It contains the classical four-helical up-and-down bundles (cytochrome *c*, hemerythrin, and viral coat proteins), the four-helical cytokines, the ferritin-like bundles, the rop protein, the three-helical spectin repeat unit, five-helical apolipophorin-III, and others. Although the proteins in this subbranch have a significant structural similarity, in other classifications they are usually grouped in separate branches.

We have tested the program by comparing all the polypeptide chains from the Protein Data Bank (PDB), and have performed an automated hierarchical cluster analysis. We have compared our results to other classifications, including the semiautomatic structural classification of proteins (SCOP),[15] and two completely automatic classifications: Class-Architecture-Topology-Homology (CATH)[16] and fold classification based on structure-structure alignment of proteins (FSSP).[17] In general, these hierarchies are in good agreement, with several interesting differences. We have detected all previously reported cases of structural similarity in proteins, and have discovered a few new significant similarities.

METHODS

We designed our algorithm keeping in mind that the biologically meaningful similarities of protein structures consist mostly of secondary structure elements: α-helices and/or β-strands. Instead of operating on a large number of atoms or residues, we considered only secondary structure elements (SSEs), represented as vectors. Our goal was to find large sets of SSEs in two protein structures which

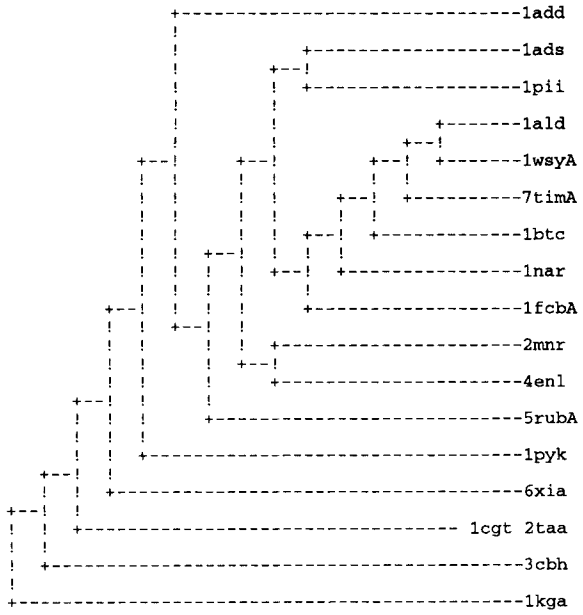

Fig. 4. TIM-barrels. The 18 TIM-barrel structures in the data set were clustered in one single branch. Although TIM-barrel proteins all have the same topology (with possible excursions), they are relatively difficult to superimpose. Some structural similarities between the members of this branch may include nontopological matches. Note that cellobiohydrolase I (3cbh) is a (7, 8) barrel.

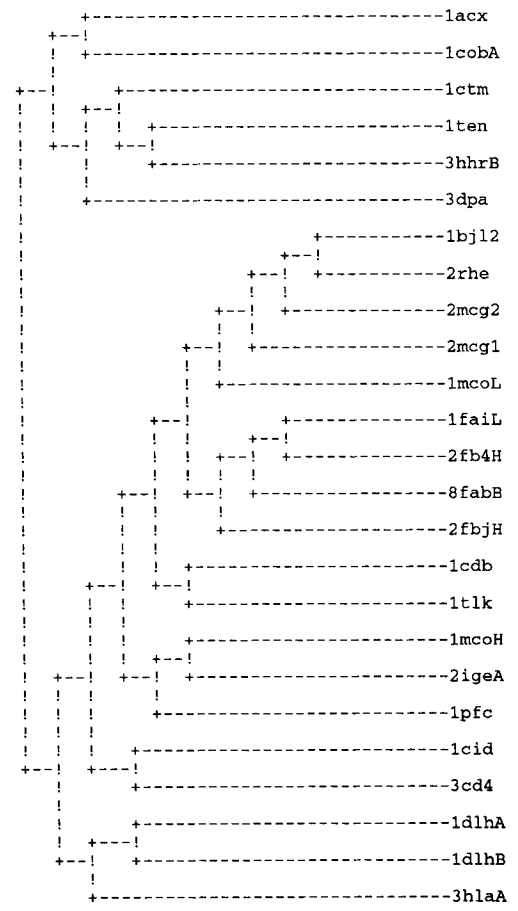

Fig. 5. Immunoglobulin-like fold. This subbranch contains all the immunoglobulins in the data set, plus other nonrelated proteins having an immunoglobulin-like fold. The latter include the fibronectin type III proteins, Cu,Zn superoxidedismutase, the major histocompatibility complex, CD4, and telokin. Note the two main divisions in this branch. The first contains immunoglobulin-like beta sandwiches (but no immunoglobulins), and the second contains all the immunoglobulins.

could be superimposed with a small rmsd. We declared that a pair of SSEs from one protein and a pair of SSEs from another protein are compatible if they satisfy certain distance and angle restraints. Usually such compatibility means that the significant portion of these pairs of SSEs can be superimposed with a relatively small rmsd. After enumerating all the compatible pairs of the SSEs of two proteins, we found several large ensembles of mutually compatible SSEs. Finally, to refine the match, we went back from the secondary structure representation to the C^{α}-trace representation, including other C^{α}-atoms (or other fragments of the C^{α}-trace) consistent with the initial match.

Thus, our algorithm consists of four steps:

1. Secondary structure assignment.
2. Search for compatible pairs of SSEs.
3. Search for large ensembles of mutually compatible pairs of SSEs.
4. Extension and refinement of the match.

Secondary Structure Assignment

Input for the algorithm is C^{α}-traces of two PDB entries of the proteins to undergo comparison. To assign the secondary structure to the C^{α}-trace we slide prototypes of an α-helix and a β-strand along the protein backbone. The fragment of C^{α}-trace is assigned a particular secondary structure if it superimposes with a small-enough rmsd to one of the prototypes. Usually we use a C^{α}-trace fragment size of 5 residues, an rmsd limit for α-helix of 0.4 Å, and an rmsd limit for β-strand of 0.8 Å. The typical helix and strand are both taken from the same PDB entry: 1bp2, residues 104–109 and 80–85, correspondingly, according to Unger et al.[18] The rmsd limits are chosen to make the number of residues in α-helical conformation and β-strand conformation approximately equal. Our assignment, based only on the conformation of the trace, but not on the pattern of the hydrogen bonds, usually fits well with the helix assignments in PDB files. For β-strands, however, our assignment can vary somewhat because we do not require hydrogen bond formation for the β-strand, considering even a single extended strand as being in β-conformation. This inaccuracy in secondary-structure assignment is not important for our procedure and does not affect results of the comparison.

The vectors representing each SSE are calculated as in MOLSCRIPT.[19]

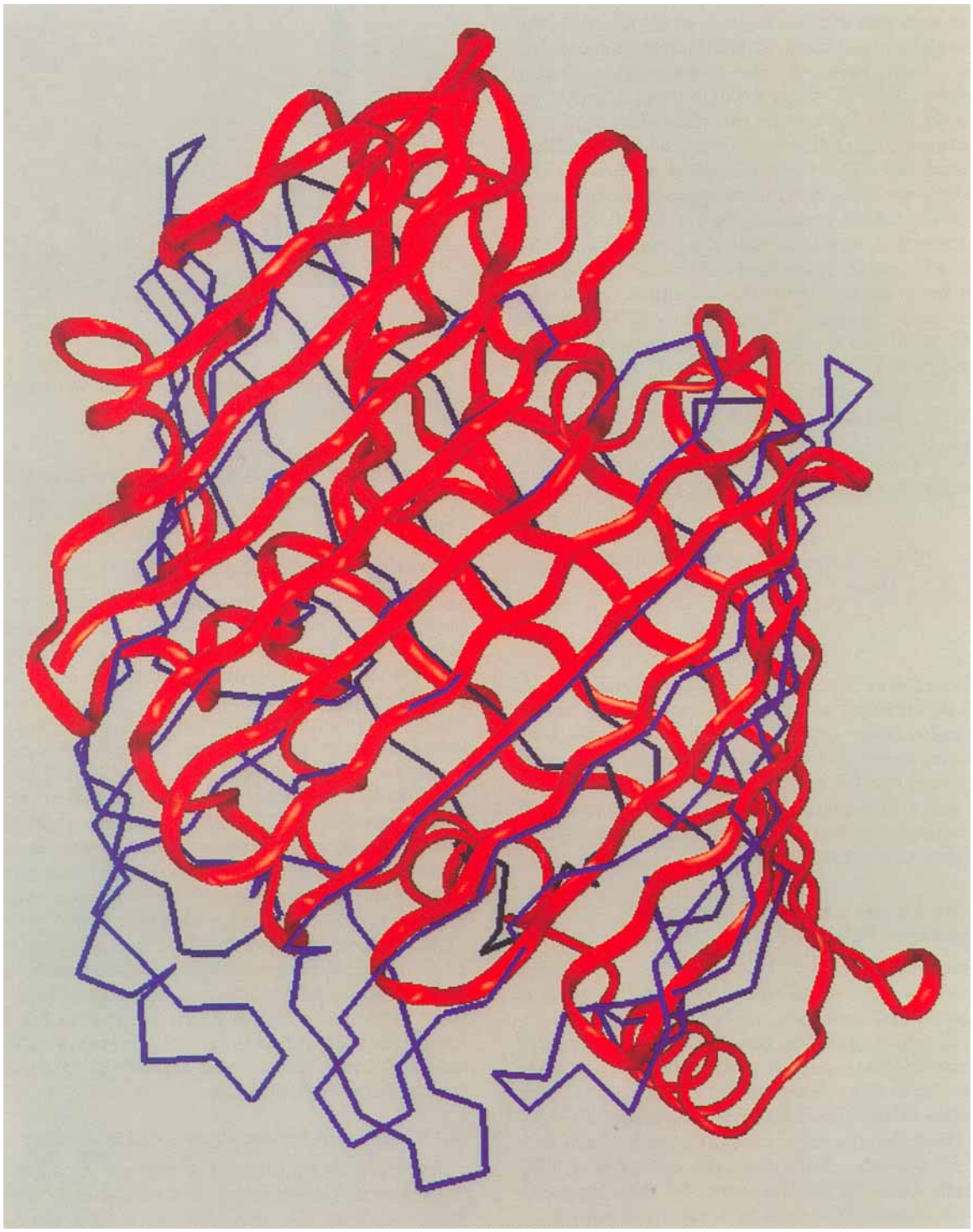

Fig. 6. Superimposed structures of bacteriochlorophyll *a* protein (PDB identifier 3bcl) and bacterial porin (PDB identifier 1omp). One hundred thirty-seven C^{α} atoms are matched with an rmsd of 2.35 Å, and a Z-score of 10.72. Corresponding residues are (with PDB enumerations):

1omf	3bcl
12–25, 39–46, 61–68, 73–79, 179–183, 184–189, 213–222, 225–235, 250–264, 269–280, 288–292, 294–316, 328–340	127–140, 202–209, 179–186, 157–163, 7–11, 19–24, 243–252, 229–239, 34–48, 50–61, 63–67, 68–90, 95–107

Search for Compatible Pairs of SSEs

For each pair of SSEs $\{S_i, S_j\}$, we calculate (1) the angle Γ_{ij} between them, (2) the shortest distance between their axes, (3) the corresponding closest points on the axes (C_i^j on the axis of the S_i and C_j^i on the axis of the S_j), and (4) the minimum (D_{min}) and maximum (D_{max}) distances from each SSE to their medium line. The medium line is defined by the middle point of the shortest segment between axes of the SSEs and the vector $(S_i/|S_i| + \delta_j^i * S_j/|S_j|)$, $\delta_i^j = \pm 1$. We chose the sign δ_i^j so that the projections of the vectors S_i and S_j to the medium line lie on one side from the projection of the closest points. In the case when vector S_i (and/or S_j) goes through the closest point, we divide it into two parts, separated by the closest point C_i^j (C_j^i). In total we have five parameters to determine the compatibility of two pairs of SSEs: angle Γ_{ij}, and two distances for each of the SSEs (Di_{min}, Di_{max}, Dj_{min}, Dj_{max}) (Fig. 1). We consider pairs of the SSEs (of the same type) $\{S_i^A, S_j^A\}$ from protein A and $\{S_k^B, S_l^B\}$ from protein B as compatible, if:

$$|\Gamma A_{ij} - \Gamma B_{kl}| < \Gamma_0 (\text{usually } \Gamma_0 = 50°)$$
$$Di_{min} - Dk_{max} < \varepsilon \; (\varepsilon = 1.5 Å)$$
$$Dk_{min} - Di_{max} < \varepsilon$$
$$Dj_{min} - Dl_{max} < \varepsilon$$
$$Dl_{min} - Dj_{max} < \varepsilon$$

These distance and angle restraints filter out differently arranged pairs of SSEs, leaving only superimposable SSEs. In contrast with the distance constraints, based on the distance between centers of the SSEs, our filtration procedure can detect difficult cases, when only relatively small segments of the SSEs are compatible, or when one of the SSEs is significantly larger than another.

Search for the Largest Ensemble of Compatible Pairs of SSEs

This problem is equivalent to the maximum clique problem and known to be NP-complete. To avoid a combinatorial explosion, we made a restriction on the maximum distances between SSEs, to be incorporated into the ensemble. This restriction automatically excludes meaningless cases of random matches of distant SSEs. We incorporate an SSE into the ensemble only if the distance between this and all the other SSEs from the ensemble is $<D_0$ (usually D_0 = 25 Å). The search for the largest ensemble is straightforward by a recursive scheme.

A new ensemble begins with a pair of compatible SSEs. In each step we check the next pair of compatible SSEs, to see if they are compatible with all the SSEs in the ensemble. If they are, we add these SSEs into the ensemble. When all pairs of compatible SSEs are checked, we remove the last one from the ensemble and repeat the procedure. Thus we test all possible combinations of compatible SSEs. The program keeps in memory not only the largest ensemble, but several (up to 3,000) different ensembles.

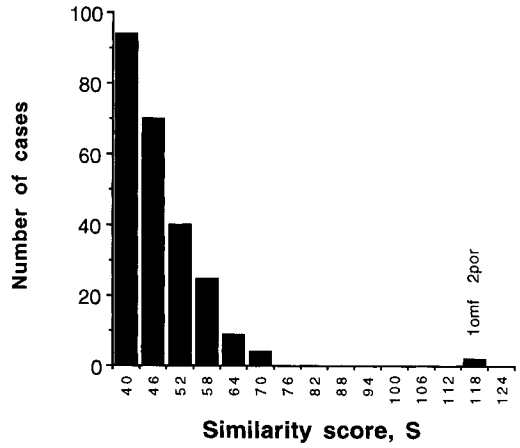

Fig. 7. Distribution of similarity score S for 3bcl-structure, compared with all other proteins from our set.

Extension and Refinement of the Match

After finding large ensembles of mutually compatible SSEs, we need to find the correspondence between C^α-atoms from two proteins. For this purpose we apply an iterative procedure, beginning from the initial superposition of the centers of compatible fragments of SSEs. This preliminary match between protein structures can be improved significantly by small additional rotations and/or translations. To obtain a better superposition we first try to include as many C^α-atoms as possible from the SSEs. We allow deletions in SSEs and use the dynamic programing technique to find the best correspondence. Subsequently, we try to collect other fragments of the C^α-trace, which may not be picked up at previous stages of the algorithm, and may not be in the helical or extended conformations. We unite those fragments into one ensemble unless the rmsd exceeds the threshold value (usually 3.2 Å). Finally, we recalculate the rotation matrix from this new C^α-correspondence and repeat the procedure several (usually five) times.

Significance of Similarities: Similarity Score

We evaluate significance of the similarity from the statistical distribution of the similarity score. Because all of the matches are within a small threshold of the rmsd (3.2 Å), we could use a number of residues N in the match as an approximate measure of the similarity: the more residues in the match, the better the similarity. In fact, we use a similarity score S = 3*N/(1 + rmsd) to increase the score of matches with smaller rmsd. If rmsd = 2.0 Å, then S = N. Comparing a structure of protein A with all the others we compute a distribution of the

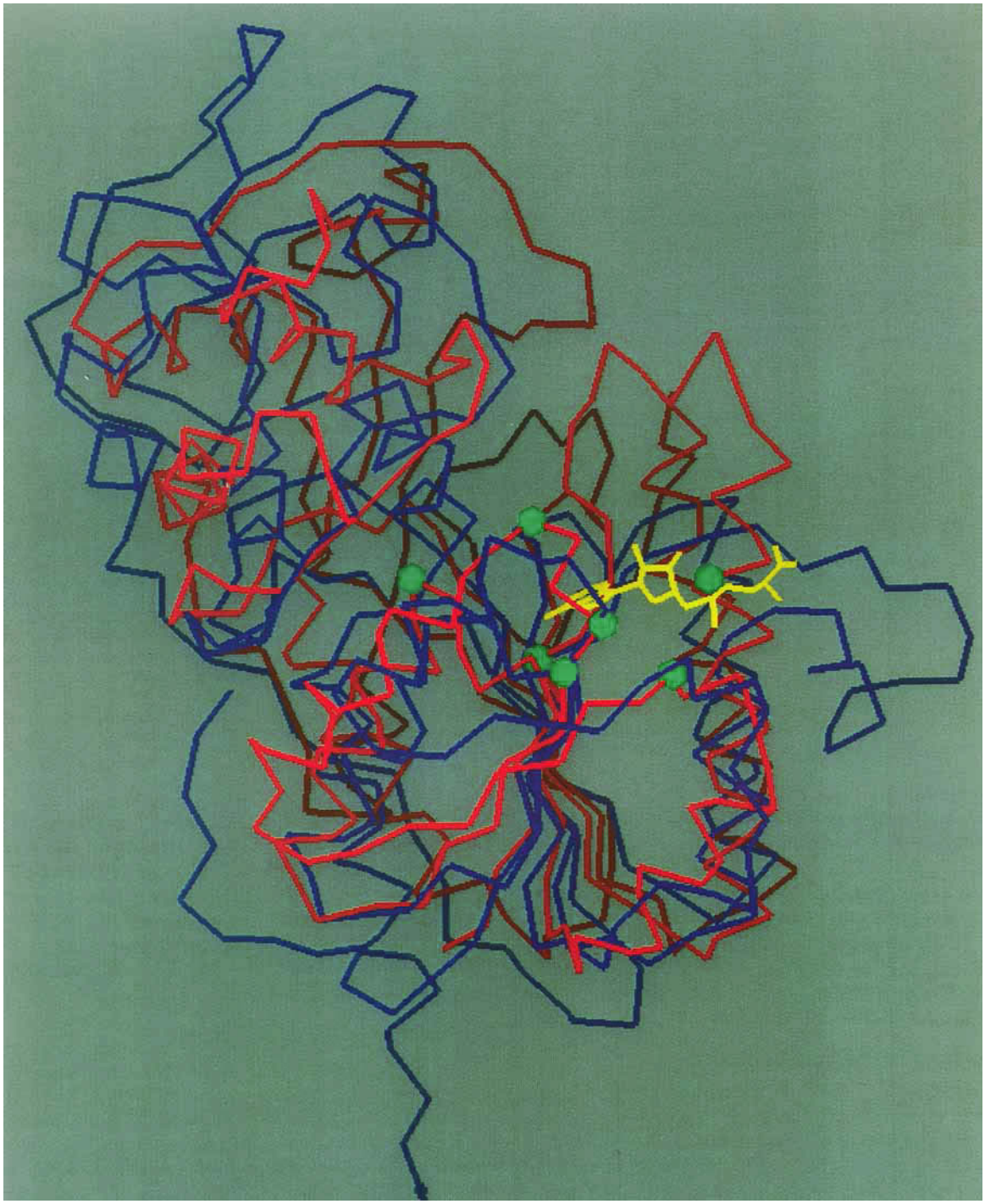

Fig. 8. Superposition of similar domains from UDP-galactose 4-epimerase (1udp) and DNA methyltransferase (1hmy). One hundred forty-eight C^{α}-atoms can be superimposed with an rmsd of only 2.32 Å, and a Z-score of 7.59. Active-site residues of 1udp and a ligand binding to 1hmy are shown in green.

similarity score S. From this distribution, for each protein B we can calculate the statistical significance of the similarity between proteins A and B in units of standard deviations Z(A, B). Although Z(A, B) and Z(B, A) are correlated, Z(A, B) may be large, while Z(B, A) may be a small value. For example, if

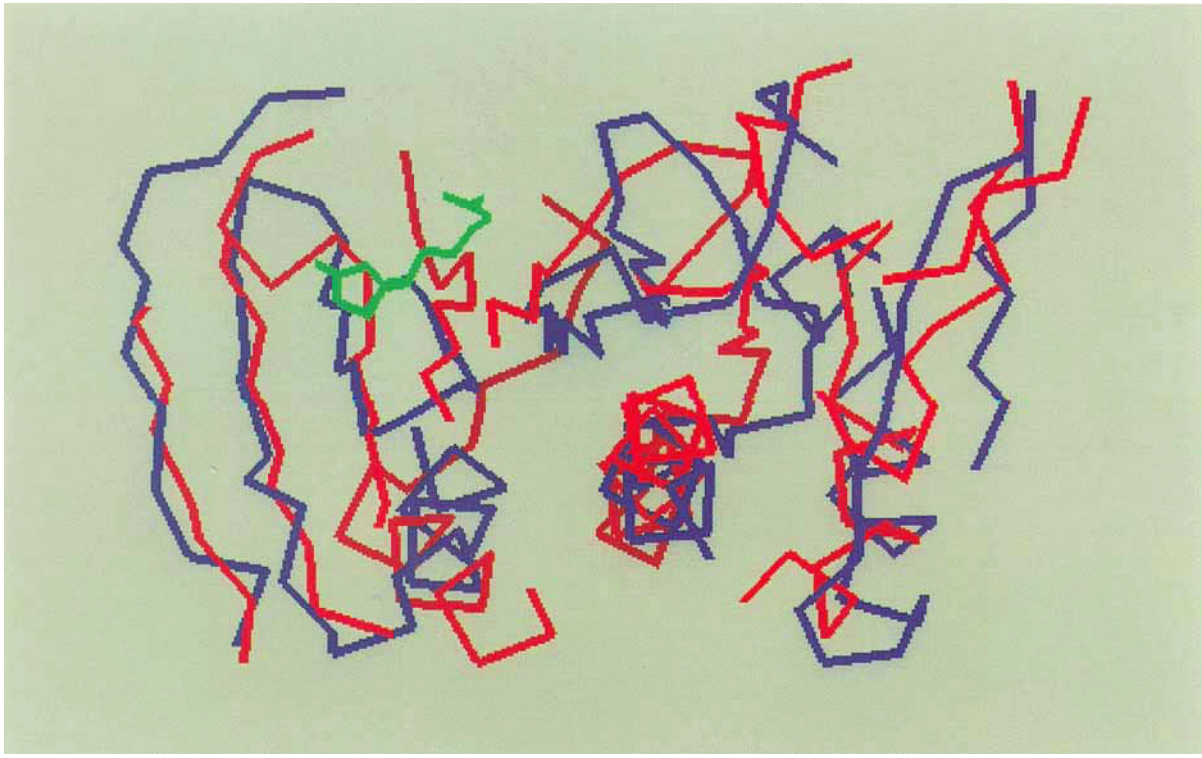

Fig. 9. The common α + β sandwich in BirA protein (1bib) and DNA-polymerase III (2pol). For clarity, only superimposed fragments are shown. In total, 107 C^{α}-atoms can be superimposed with an rmsd of 2.79 Å, resulting in a Z-score of 5.75. The helical part of the sandwich in DNA polymerase binds to DNA, whereas the β-sheet of the sandwich in BirA protein contains a biotine-binding site (biotinylated lysine shown in green).

A is a big protein of 300 residues, it is easy to find a 60-residue match with many proteins. However, for a small protein B, the same 60-residue match would be very rare, resulting in a larger Z-score. To determine the structural similarity between proteins A and B, we used the score $Z_{Ab} = Z(A, B) + Z(B, A)$, which gives a symmetrical value of the significance of similarity between two proteins. This score is the one we use to cluster the structures.

Compilation of Protein Structures

The PDB[20,21] is highly redundant and nonhomogenous. It contains the structures of >3,000 chains, but only <1,000 entries are not (almost) identical to another entry. In this study the list of structures used are obtained from a previously derived representative set.[11]

Fischer et al.[11] derived five structurally nonredundant data sets with five different similarity thresholds. The structural similarity threshold for the fifth data set required that for any two chains in the data set the following condition holds: no more than half the residues of the larger structure can be matched with an rmsd below 3.0 Å. Here we used the data set obtained at stage four, which contains the structures of 320 chains. This set features some structural redundancy, i.e., there are several pairs of homologous structures (for details on the structural threshold used, see Table 1 of Fischer et al.[11]) In Fischer et al.,[11] all chains below 60 residues were excluded as well as structures which the DSSP program would not accept (e.g., C^{α}-only chains). We added these chains to the 320 original data set, obtaining a total of 426 chains.

Hierarchical Cluster Analysis

Having obtained the matrix of similarities between each of the chains in our data set, a cluster analysis was performed using the Phylip package (Phylogeny Inference Package, Joseph Felsenstein and the University of Washington), with the NEIGHBOR program. This program implements the Neighbor-Joining method of Saitou and Nei.[22] The result of this analysis is a tree where similar proteins are placed together in one subbranch. This analysis was carried out completely automatically.

RESULTS

We compared all the structures from the structurally nonredundant representative list of the PDB entries. To test the program, we checked previously reported cases of biologically important similarities.

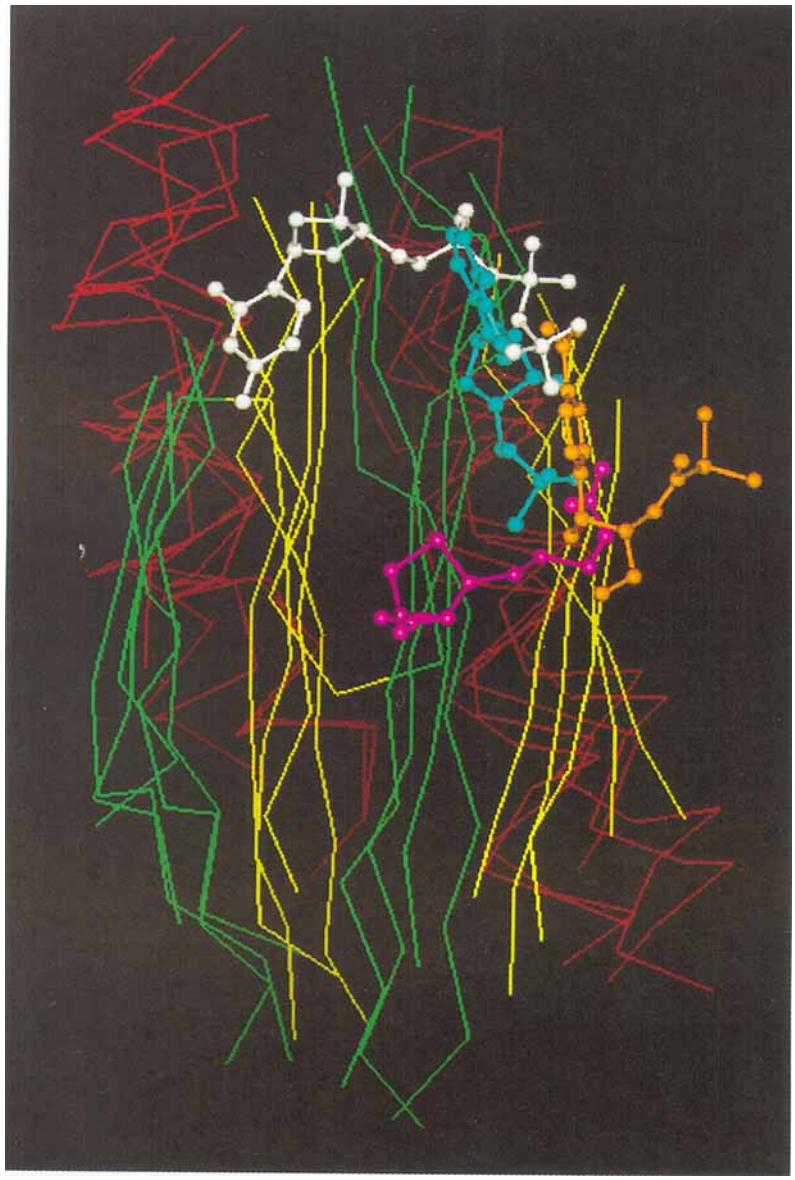

Fig. 10. Common structural motif in five protein structures: BirA protein (1bib), DNA-polymerase III (2polA), aminoacyl tRNA synthetase (1sry), chloramphenicol acetyltransferase (4cla), and DNA polymerase (Klenow fragment) (1kfd). The ligands binding to 1bib, 4cla, and 1kfd are placed in relatively similar positions. The ATP-binding site in DNA-polymerase III is not known.

For example, we have a clearly distinguishable distribution of the similarity score in the comparison of the TIM-barrel structures (Figs. 2, 4).

Having obtained all the pairwise similarities, we constructed a hierarchical tree of protein structures. In this tree, the three main classes of protein folds are clearly separated: (1) mostly α chains, (2) α/β chains, and (3) mostly β chains. In addition, the $\alpha + \beta$ and the small chain classes showed some grouping, but not so tight as the above three. Here we only show some of the major superfamilies that can easily be identified in our tree (Figs. 3–5). The complete tree is available through the WWW at http:www-lmmb.ncifcrf.gov/~nicka/tree.html.

Due to the biologically adequate method of similarity search and evaluation of the significance of similarities, we were able to obtain a reasonable classification completely automatically. The major superfamily clustering, as well as the overall distribution of the major three classes of proteins in our tree, is consistent with previous classifications.[7,15,16]

To find the most interesting cases of similar structures, we sorted all pairs of protein structures ac-

cording to their similarity score Z. As one could expect, the list was topped by large similarities of TIM barrels, viral proteins, and other structures belonging to the same class of folding patterns with conserved topology. However, there were several interesting, significant pairwise similarities. Some of these, to the best of our knowledge, have not been discussed.

Bacteriochlorophylll *a* Protein and Porins

The most significant case among nontopological matches is a similarity between bacteriochlorophyll *a* protein (or Bchl protein) from the green photosynthetic bacterium *Prosthecochloris aestuarii*[23] and bacterial porin.[24] Both structures have a huge, common, similarly-bent fragment of β-sheet (Fig. 6). One hundred thirty-seven C^{α} atoms from these structures can be superimposed with a rmsd = 2.35 Å. This size of similarity is typical for a TIM-barrel motif, or for viral coat proteins. Distribution of the S-score indicates that this similarity is unique, and present only in the structures of the Bchl protein and porins (Fig. 7). This statistically very significant match should have a biologically meaningful explanation.

The Bchl protein is a part of the light-gathering complex of the green photosynthetic bacteria, which also include the chlorophyll bodies and the reaction centers. The Bchl protein is thought to transmit excitation energy from the light-harvesting chlorophyll to the reaction center, incorporated within the cytoplasmic membrane. The Bchl protein is membrane-bound in vivo, and is water-soluble in vitro.

Porins form channels in the outer membrane of bacteria, to allow the passive diffusion of nutrients and waste products.

Besides the remarkably large, similarly-bent fragment of the β-sheet, these two proteins have a similar general shape, and both are functional in the trimer form.

This may imply that the common structural motif in porin and the Bchl protein is well-suited for their functioning within membranes.

Membrane Helices

There are several significant similarities between proteins incorporated into a membrane or a lipid layer. They have long, almost parallel helices, penetrating through the lipid layer(s). On the top of the list are the similarities between bacteriorodopsin (1brd) and other membrane/lipid proteins: colicin (1colA), apolipophorin-III (1aep), and apolipoprotein E2 (1le2). The topology of these folds is different, but all of them have obvious similarly-arranged helices and a similar ability to penetrate through the lipid layers.

Similarities in the α/β Class

The number of different topologies in this class is very large. For example, there are 56 different superfamilies of this class in the classification of protein structures (SCOP[15]). However, many of these proteins have a very similar three-dimensional arrangement of secondary-structure elements. One example is a similarity between UDP-galactose 4-epimerase (1udp[25]) and DNA methyltransferase (1hmy[26]). Both proteins consist of two domains. The larger domain in 1upd is a typical Rossman fold. One hundred forty-eight C^{α}-atoms from the larger domain of 1hmy can be superimposed with an rmsd of only 2.32 Å (Fig. 8). The active sites of both molecules appear to be at the same place, confirming the biological significance of this similarity.

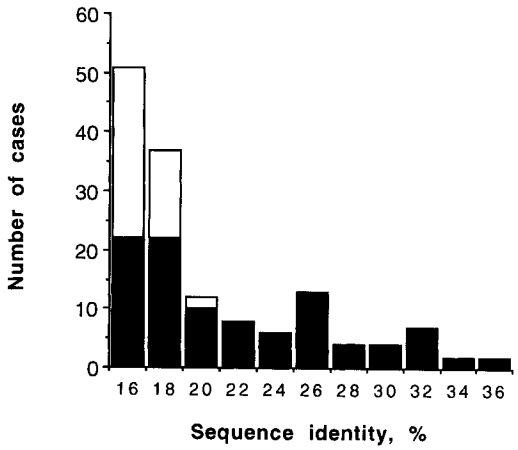

Fig. 11. Dependence of number of significant and nonsignificant structural similarities on percentage of sequence identity. Solid bars show significant similarities (defined as those with Z-score >3.0); open bars show nonsignificant similarities. All similarities with sequence identity >20% are significant.

Citrate Synthase and Cytochrome P450

The common SARF between citrate synthase (1csc) and cytochrome P450 (2cpp) consists of six α-helices. One hundred twenty-two C^{α}-atoms from these structures can be superimposed with an rmsd of 2.61 Å. This similarity is statistically significant, and these proteins are clustered together in our tree, and yet the functional meaning of this similarity is not clear: it may just be a result of the compact arrangement of α-helices, in line with the ideas of Finkelstein and Ptitsyn.[27]

BirA Protein and DNA Polymerase III

The similarity detected between these two proteins may have a functional meaning. BirA is a bifunctional protein, acting as a biotin synthase and a repressor of the biotin operon.[28] The β subunit of *E. coli* DNA polymerase III makes a clamp for DNA, with a specific arrangement of its inner helices providing quick sliding of the DNA through the protein.[29] The second domain of the BirA structure (residues 64–273) and two domains of DNA polymerase III share a common core of three helices and seven

β-strands, arranged as an α + β sandwich (Fig. 9). Although the specific DNA-binding function of the BirA protein was localized in the first domain, the helices from the second domain may provide nonspecific binding to DNA, in a manner similar to that of the corresponding helices from DNA polymerase III.

DISCUSSION

Consistency of our results of the exhaustive comparison of PDB structures with previous (not always automatic) classifications of protein folds indicates that the program SARF2 detects the biologically meaningful similarities between protein structures.

In our analysis of the results we considered three types of similarities.

The first type of similarities occurs between pairs of related proteins, i.e., proteins belonging to the same functional family. These families showed clear subbranches in our automatic hierarchical classification. Examples of such are the immunoglobulins, the globins, the lipocalins, and several families of the α/β fold. These similarities are well-known and have been discussed previously.

The second type of similarities occurs between proteins of different families and functions which share the same topology. Examples of such similarities are the various TIM-barrels, immunoglobulin-like folds, helical bundles, and various double-wound α/β folds. These have been named "superfolds" by Orengo et al.[30] These superfolds are also clearly clustered into the same subbranches in our tree.

The above two types of similarities include the large majority of significant similarities found.

The third type of similarities, which we have discussed in some detail, occurs between evolutionary nonrelated proteins, in which the connectivity of matched backbone fragments is not conserved. This new type of structural similarity has not been studied in detail. Here we have shown that these similarities are statistically significant.

What is the proportion of nontopological similarities? To answer this question, we have introduced a definition of the essentially nontopological similarity. Sometimes, e.g., in the case of UDP-galactose 4-epimerase (1udp) and DNA methyltransferase (1hmy), the similarity is almost topological, just slightly decorated by nontopological fragments. We call the similarity essentially nontopological, if the number of C^{α}-atoms in the topological match between these proteins is <75% of the nontopological match.

It appeared that for a very significant level of the structural similarity ($Z_{AB} > 6.0$), 11% (56 cases) of all the 506 similarities were essentially nontopological. Forty-four of these 56 nontopological similarities (79%) were parallel helix bundles with different connectivities. Another 12 similarities belonged to membrane β-barrels, α/β proteins, α + β sandwiches, and β-sandwiches.

In multidomain structures, it can be the case that one domain is similar to one subbranch and the other domain to another subbranch. Here, the automatic clustering placed each chain in the cluster that best fit it. Thus, it may be interesting to make a domain cut for such chains and redo the automatic classification at the domain level. The disadvantage of cutting into domains is that we may miss similarities within domain interfaces.

Comparison With SCOP, FSSP, and CATH

There are at least three different structural classifications, which are available through the WWW: SCOP (URL http://scop.mrc-lmb.cam.ac.uk/scop/), which is essentially a manual classification; FSSP (http://swift.embl-heidelberg.de/fssp/), constructed with the DALI program;[7] and CATH (http://www.biochem.ucl.ac.uk/bsm/cath/CATHintro.html), constructed with the SSAP program.[31] SCOP and CATH classify structures in terms of their topology, while FSSP can group together nontopological matches as well. The difference between principles of these classifications can be seen in a simple example of four-helix bundle motif. CATH, FSSP, and SARF2 automatically cluster apolipoprotein (1aep) and cytochrome b562 (256b) together, whereas authors of SCOP subjectively decided to separate them. Four-helix bundles with different topologies, e.g., ferritin (1fha) and human growth hormone (1huw), are clustered in different folds in SCOP and CATH, whereas in FSSP and SARF they are placed together.

Although SCOP is a very sensible classification, it is not a purely structural, but rather a functional, classification. The authors avoid placing functionally unrelated proteins together. For example, they do not unite globin and colcin into one cluster, while all automatic methods of structural classification cluster them together.

In general, our automatic classification of the structural universe is quite consistent with that of previous works. However, we have noted several cases of proteins that are clustered in the same class, although they were previously related to different groups. This is because they have a common nontopological arrangement of the SSEs, which matched nicely.

Let us see how these classifications deal with the structural similarities mentioned in this paper.

Representative TIM-barrels are grouped together in all these classifications with the following minor differences: cellobiohydrolase II (3cbh) is clustered into a separate fold of cellulases in SCOP; in CATH, cellobiohydrolase II (3cbh) and adenosine deaminase (1add) are placed into the "unclassified" section, and four structures (1ads, 1btc, 2taa, and 2tmd) into the "multidomain" section; in SARF2 classification, trimethylamine dehydrogenase (2tmd) is clustered

into the FAD/NAD binding motif because of its other domains.

Topological classifications (SCOP and CATH) do not cluster porins and bacteriochlorophyll *a* protein together. The FSSP and SARF2 classifications place them together into a tight cluster. However, the FSSP alignment (128 residues, 5.1 Å rmsd) is less convincing than the SARF2 alignment (137 residues, 2.3 Å rmsd).

None of the previous classifications cluster UDP-galactose 4-epimerase (1udp) and DNA methyltransferase (1hmy) together, in spite of the remarkable structural alignment (148 residues, 2.32 Å rmsd) and the similar location of active sites. Interestingly, even the order of SSEs is almost the same in both proteins.

SARF2 clusters BirA proteins (1bib) with DNA-polymerase III (2pol), which are not related in other classifications. Recently, a similarity between BirA and aminoacyl tRNA synthetase (1sry) has been reported.[32] In our tree, 1sry is also clustered together with 1bib and 2pol, along with two other proteins: chloramphenicol acetyltransferase (4cla), and DNA polymerase (Klenow fragment) (1kfd). Although the cluster is not very tight, as if only a portion of these big proteins are matched, the cluster could have a functional meaning, because after superposition of their common motif, active sites of these proteins appear to be similarly located (Fig. 10).

Sequence-Structure Relationship

We have analyzed the distribution of the matches obtained in the all-vs.-all comparison as a function of the percentage of identities found in the match. In this analysis we only consider those matches having at least 60 residues matched.

Figure 11 shows a fraction of the significant and presumably nonsignificant similarities depending on the percentage of the sequence identity in the match. We define a structural similarity as significant if its Z-score is >3.0. The distribution shows that matches having >20% identical residues imply structurally-related chains. This is considerably lower than the typical 25–30% twilight-zone range of sequence similarity, obtained from the optimum sequence alignment of two proteins.

Some other interesting observations can be derived from our analysis.

1. A Z-score >3.0–3.5 most certainly implies important, biologically meaningful similarity. This may be the cutoff where the twilight zone of structural similarity appears by chance.

2. In several nontopological equivalences in the α/β proteins, the active sites are at the same position, confirming a possible biological role of the nontopological similarities.

3. The α + β class is difficult to classify automatically, because proteins of this class can partially match α/β, mostly β, or mostly α proteins.

4. Small chains are also difficult to classify in the major three classes; they group by themselves and along different subbranches.

CONCLUSIONS

We have developed a method for the structural comparison of proteins which was demonstrated to work well, i.e., it identifies previously detected similarities, and, in addition, allows identification of novel matches. The method detects both topological and nontopological similarities.

To evaluate the significance of a structural match, we defined a statistical score which reflects an intuitive understanding of interesting similarities.

In an all-vs.-all comparison, followed by a full automated clustering, we obtained a classification which is consistent with previous classifications.

The method is robust and able to detect "difficult" similarities with relative large rmsds. The program is available in the WWW at the address: http:llwww-lmmb.ncifcrf.gov/~nicka/info.html.

REFERENCES

1. Farber, G.K., Petsko, G.A. The evolution of α/β barrel enzymes. Trends Biochem Sci 15:228–234, 1990.
2. Branden, C.-I. Relation between structure and function of α/β proteins. Q. Rev. Biophys. 13:317–338, 1980.
3. Holm, L., Sander, C. Globin fold in a bacterial toxin. Nature 361:309, 1993.
4. Holm, L., Sander, C. Searching protein structure database has come of age. Proteins 19:165–173, 1994.
5. Orengo, C. Classification of protein folds. Curr. Opin. Struct. Biol. 4:429–440, 1994.
6. Alexandrov, N.N., Takahashi, K., Go, N. Common spatial arrangements of backbone fragments in homologous and non-homologous proteins. J. Mol. Biol. 225:5–9, 1992.
7. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.
8. Orengo, C.A., Taylor, W.R. A local alignment method for protein structure motif. J. Mol. Biol. 233:488–497, 1993.
9. Maiorov, V.N., Crippen, G.M. Significance of RMSD in comparing three-dimensional structures of globular proteins. J. Mol. Biol. 235:625–634, 1994.
10. Alexandrov, N.N., Go, N. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. Protein Sci. 3:866–875, 1994.
11. Fischer, D., Tsai, C.J., Nussinov, R. A 3-D sequence-independent representation of the protein data bank. Protein Eng., 8:981–997, 1995.
12. Abagyan, R.A., Maiorov, V.N. A simple qualitative representation of polypeptide chain folds: Comparison of protein tertiary structures. J. Biomol. Struct. Dynam. 5:1267–1279, 1988.
13. Mitchell, E.M., Artymiuk, P.J., Rice, D.W., Willet, P. Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J. Mol. Biol. 212:151–166, 1989.
14. Mizuguchi, K., Go, N. Comparison of spatial arrangements of secondary structural elements in proteins. Protein Eng. 8:353–362, 1995.
15. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C. scop: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. 247:536–540, 1995.
16. Orengo, C.A., Flores, T.P., Taylor, W.R., Thornton, J.M. Identification and classification of protein fold families. Protein Eng. 6:485–500, 1993.
17. Holm, L., Sander, C. The FSSP database of structurally

aligned protein families. Nucleic Acids Res. 22:3600–3609, 1994.
18. Unger, R., Harel, D., Wherland, S., Sussman, J.L. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5:355–373, 1989.
19. Kraulis, P.J. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Cryst. 24:946–950, 1991.
20. Abola, E.E., Bernstein, F.C., Bryant, S.H., Koetzle, T.F., Weng, J. Protein data bank. In: "Crystallographic Databases: Information Content, Software Systems, Scientific Applications." Allen, F.H., Bergerhoff, G., Sievers, R. (eds.): Bonn/Cambridge/Chester: Data Commission of the International Union of Crystallography, 1987:107–132.
21. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-based archival file for macromolecule structures. J. Mol. Biol. 112:535–542, 1977.
22. Saitou, N., Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425, 1987.
23. Tronrud, D.E., Schmid, M.F., Matthews, B.W. Structure and X-ray amino acid sequence of a bacteriochlorophyll a protein from *Prosthecochloris aestuarii* refined at 1.9 Å resolution. J. Mol. Biol. 188:443–454, 1986.
24. Cowan, S.W., Schirmer, T., Rummel, G., Steiert, M., Ghosh, R., Pauptit, R.A., Jansonius, J.N., Rosenbusch, J.P. Crystal structures explain functional properties of two *E. coli* porins. Nature 358:727–733, 1992.
25. Bauer, A.J., Rayment, I., Frey, P.A., Holden, H.M. The molecular structure of UDP-galactose 4-epimerase from *Escherichia coli* determined at 2.5 angstroms resolution. Proteins 12:372–381, 1992.
26. Cheng, X., Kumar, S., Posfai, J., Pflugrath, J.W., Roberts, R.J. Crystal structure of the Hhal DNA methyltransferase complexed with S-adenosyl-L-methionine. Cell 74:299–307, 1993.
27. Finkelstein, A.V., Ptitsyn, O.B. Why do globular proteins fit the limited set of folding patterns? Prog. Biophys. Mol. Biol. 50:171–190, 1987.
28. Wilson, K.P., Shewchuk, L.M., Brennan, R.G., Otsuka, A.J. *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA-binding domains. Proc. Natl. Acad. Sci. U.S.A. 89:9257–9261, 1992.
29. Kong, X.-P., Onrust, R., O'Donnell, M., Kuriyan, J. Three-dimensional structure of the beta subunit of *E. coli* DNA polymerase III holoenzyme: A sliding DNA clamp. Cell 69:425–437, 1992.
30. Orengo, C.A., Jones, D.T., Thornton, J.M. Protein superfamilies and domain superfolds. Nature 372:631–634, 1994.
31. Taylor, W.R., Orengo, C.A. Protein structure alignment. J. Mol. Biol. 208:1–22, 1989.
32. Artymiuk, P.J., Rice, D.W., Poirrette, A.R., Willet, P. A tale of two synthetases. Struct. Biol. 1:758–760, 1994.