PROTEINS: Structure, Function, and Genetics 46:405–415 (2002)
DOI 10.1002/prot.1176

Automated Assignment of SCOP and CATH Protein Structure Classifications From FSSP Scores

Gad Getz,[1] Michele Vendruscolo,[2] David Sachs,[3] and Eytan Domany[1]*
[1]*Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel*
[2]*Oxford Centre for Molecular Sciences, New Chemistry Laboratory, Oxford, United Kingdom*
[3]*Department of Physics, Princeton University, Princeton, New Jersey*

ABSTRACT **We present an automated procedure to assign CATH and SCOP classifications to proteins whose FSSP score is available. CATH classification is assigned down to the topology level, and SCOP classification is assigned to the fold level. Because the FSSP database is updated weekly, this method makes it possible to update also CATH and SCOP with the same frequency. Our predictions have a nearly perfect success rate when ambiguous cases are discarded. These ambiguous cases are intrinsic in any protein structure classification that relies on structural information alone. Hence, we introduce the "twilight zone for structure classification." We further suggest that to resolve these ambiguous cases, other criteria of classification, based also on information about sequence and function, must be used. Proteins 2002;46:405–415.**

Key words: protein structure; protein databases; CATH; FSSP; SCOP; classification; clustering

INTRODUCTION

The first step to analyze the vast amount of information provided by genome sequencing projects is to organize proteins (the gene products) into classes with similar properties. Because during evolution protein structures are much more conserved than sequences and functions,[1] proteins are usually classified first by their structural similarity (phenetic classification) and then by the similarity of their sequences or by the similarity of their functions (phylogenetic classification).[2]

A reliable structural classification scheme is useful for several reasons. Perhaps the most exciting perspective is the possibility to routinely assign a function to newly identified genes.[3] This goal may be achievable because a classified database provides a library of representative structures to perform prediction of protein structure by homology[4,5] or by threading,[6–8] and it allows for the identification of distant evolutionary relationships.[9] In addition, given a particular protein, it provides a tool to identify other proteins of similar structure and function.[10] The knowledge of the structure helps to reveal the mechanism of molecular recognition involved in catalysis, signaling, and binding[2] and may lead to the rational design of new drugs.[11] At a more abstract level, the physical principles dictating structural stability of proteins are revealed by their folded state. Therefore, most of the recently proposed methods to derive energy functions to perform protein fold predictions rely in different ways on structural data.[12,13]

The most comprehensive repository of three-dimensional structures of proteins is the Protein Data Bank (PDB).[14] The number of released structures is increasing at the pace of about 50 per week, and >12,000 complete sets of coordinates were available at the time of writing. Many research groups maintain web-accessible hierarchical classifications of PDB entries. The most widely used are FSSP,[15] CATH,[16] SCOP,[17] HOMSTRAD,[18] MMDB,[19] and 3Dee[20] (see Table I for a list of abbreviations). Here we consider three of these: the FSSP, the CATH, and the SCOP databases. Each group has its own way to compare and classify proteins; these three classification schemes are, however, consistent with each other to a large extent.[21,22]

FSSP Database

The FSSP (Fold classification based on Structure-Structure alignment of Proteins) uses a fully automated structure comparison algorithm, DALI (Distances ALIgnment algorithm),[23,24] to calculate a pairwise structural similarity measure (the S-score) between protein chains.

The algorithm searches for that amino acid alignment between the two protein chains that yields the most similar pair of C_α distance maps. In general, the more geometrically similar two chain structures are, the higher their S-score is. The mean and standard deviations of the S-scores obtained for all the pairs of proteins are evaluated. Shifting the S-scores by their mean and rescaling by the standard deviation yield the statistically meaningful Z-scores.

For classification of structures, the FSSP uses the Z-scores for all pairs in a representative subset of the PDB. A fold tree is generated by applying an average-linkage hierarchical clustering algorithm[25] to this all-against-all

Grant sponsor: Minerva Foundation; Grant sponsor: Germany-Israel Science Foundation; Grant sponsor: US-Israel Science Foundation (BSF).

*Correspondence to: Eytan Domany, Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot 76100, Israel. E-mail: fedomany@wicc.weizmann.ac.il.

Received 23 February 2001; Accepted 13 July 2001

TABLE I. Abbreviations and Definitions

Abbreviation	Definition
3Dee	Database of protein domain definitions
ASTRAL	The ASTRAL compendium for sequence and structure analysis
CATH	Protein structure classification
CO	Classification by optimization
DALI	Protein structure comparison by alignment of distance matrices
DHS	Dictionary of homologous superfamilies
FSSP	Fold classification based on structure-structure alignment of proteins
HOMSTRAD	Homologous structure alignment database
MMDB	Molecular modeling database
PDB	Protein data bank
SCOP	Structural classification of proteins
SSAP	Structure comparison algorithm

Z-score matrix. An alternate classification based on a more common four-level hierarchy is also available.[24]

CATH Database

Orengo and coworkers use a combination of automatic and manual procedures to create a hierarchical classification of domains (CATH).[16] They arrange domains in a four-level hierarchy of families according to the protein class (C), architecture (A), topology (T), and homologous superfamily (H). The class level describes the secondary structures found in the domain[26] and is created automatically. There are four class types: mainly-α, mainly-β, α-β, and proteins with few secondary structures (FSS). The architecture level, on the other hand, is assigned manually (using human judgment) and describes the shape created by the relative orientation of the secondary structure units. The shape families are chosen according to a commonly used structure classification (e.g., barrel, sandwich, roll, etc.). The topology level groups together all structures with similar sequential connectivity between their secondary structure elements. Structures with high structural and functional similarity are put in the same fourth-level family, called *homologous superfamily*. Both the topology and homologous superfamily levels are assigned by thresholding a calculated structural similarity measure (SSAP) at two different levels, respectively.[27,28] The CATH database has been recently linked to the Dictionary of Homologous Superfamilies (DHS) database,[29] which allows further analysis of structural and functional features of evolutionary related proteins. There is a growing need for annotating proteins classified in structural databases because structural genomic initiatives are providing a large number of new proteins whose function might be gathered by distant homology informations.

SCOP Database

The Structural Classification of Proteins (SCOP)[17] database is organized hierarchically. The lower two levels (family and superfamily) describe near and distant evolutionary relationship, the third (fold) describes structural similarity, and the top level (class) describes the secondary structure content.[26] SCOP is linked to the ASTRAL compendium,[30] which provides a series of tools for further analysis of the classified structures, mainly through the use of their sequence. At variance with FSSP and CATH, SCOP is constructed manually, by visual inspection and comparison of not only structures but also sequences and functions.

Automated Assignment of SCOP and CATH Classifications

In this work we present a method, Classification by Optimization (CO), to predict without human intervention the SCOP fold level and the CATH topology level from the FSSP pairwise structure similarity score. A protein for which the Z-score is available is classified into a SCOP fold and into a CATH topology by the CO method, an optimization procedure that finds the assignment of minimal cost, where the cost is defined in terms of Z-scores (see Materials and Methods). The query for the classification of any such protein can be submitted to the web site.[31]

RESULTS

Consistency of the FSSP, CATH, and SCOP Classifications

We found that the FSSP and CATH databases are consistent.[21] In this section we show that SCOP is also consistent with these to a large extent (see also Ref. 22). In the rest of this work we use this fact to derive an automated procedure to assign the CATH and SCOP classifications starting from the FSSP Z-scores (which are updated weekly) in a fully automated fashion to include new releases in the PDB.[1] Here we further discuss the consistency of the three classification schemes by introducing concepts and quantities that are later used in the prediction of the CATH and SCOP classifications.

We first illustrate the correlation between the FSSP similarity score and the CATH classification. A simple and visually appealing way to study this problem is shown in Figure 1. The element Z_{ij} of the Z-score matrix [Fig. 1(a)] represents the score for superimposing structure i with structure j of the set PFrCs (a subset of the proteins in FSSP and CATH, see Table III and Materials and Methods) using the DALI algorithm.[23,24] In Figure 1(a) only the pairs with $Z > 2$ are shown; therefore, the matrix is sparse and the proteins are ordered in a random fashion. Figure 1(b) is produced by reordering the rows and columns of the original Z-score matrix [Fig. 1(a)]. The reordering is performed according to the CATH classification in the following way: for each of the proteins in this set we have the CATH classifications at all levels. First, we order the proteins by their class; within the class, by the architecture; within it by the topology, and so on. This reordering generates a permutation of the columns and rows of the Z matrix. The solid black grid in Figure 1(b) separates the proteins according to their CATH class, and a thin grid is placed at the boundaries between architectures.

Figure 1(b) shows the underlying order behind the apparent randomness of Figure 1(a) and reveals the extent

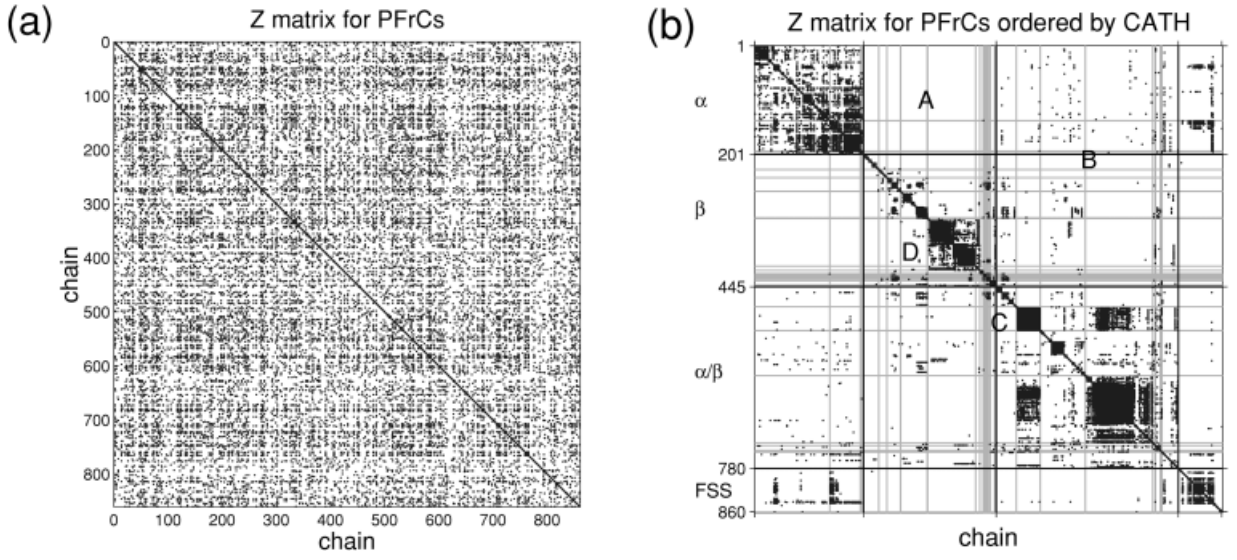

Fig. 1. **a:** Z-score matrix between all pairs of proteins in the PFrCs set. A black dot represents $Z > 2.0$. **b:** Same Z-score matrix with rows and columns rearranged by using the CATH classification (see text). Part (b) shows the underlying order behind the apparent randomness of part (a) and illustrates the extent to which the FSSP Z-scores reflect the CATH classification. The regions A, B, C, and D are discussed in the text.

to which the FSSP Z-scores reflect the CATH classification.

Several interesting observations can be made. First, consider the Class level of CATH. As can be seen in Figure 1(b), there are no matrix elements with $Z > 2.0$ in region A that connect proteins of the mainly-α class to the mainly-β class. At variance with this, some proteins from both of these classes have large Z-scores with proteins from the α-β class (region B). This is reasonable, because of the way similarity is defined by FSSP; a mainly-α protein can have a high Z-score with an α-β protein because of high similarity with the α part. Second, in the Architecture level, we observe that there are architecture families that are highly connected within themselves, e.g., α-β barrels (482–525: region C), whereas for others the intrafamily connections are more sparse. The similarities within the mainly-β sandwich family (318–406: region D) have two relatively distinct subgroups, which suggest an inner structure corresponding to the lower levels in the CATH hierarchy. Checking the topology level (the third CATH level) for this architecture, one indeed finds two large topology subfamilies, the immunoglobulin-like proteins (324–366: upper left part of region D) and the Jelly-Rolls (373–402: lower right part of region D), which correspond precisely to the two strongly connected subgroups that appear in Figure 1(b).

We found that the CATH classification at the level of topology is reflected in the Z-matrix. This is to be expected because the Z-score measures the structural similarity of two aligned proteins while preserving their connectivity. Overall, this analysis shows that the Z-matrix is correlated with the CATH classification. In a similar way it is possible to show that the Z-score is correlated with the SCOP classification. The results are available at the web site.[31]

These findings suggest that Z-scores can be used to predict the CATH and SCOP classifications of yet unclassified proteins. In what follows, we demonstrate that this indeed can be done. We also estimate the success rate of our predictions and provide a web site[31] that can be used to retrieve our predictions for the CATH topology and the SCOP fold for new entries in FSSP.

We also verified that the CATH and SCOP classifications are to a large extent mutually compatible. An immediate consequence of this is that it is possible to construct a "translation table," $\hat{T}$, from the proteins that have already both a CATH and a SCOP classification. In this way, given a CATH entry, one can obtain the corresponding SCOP classification (see Fig. 2). Row i of the table refers to a particular CATH topology and column j to a particular SCOP fold. The element $\hat{T}_{ij}$ of the table is the measured fraction of times that a protein has a CATH topology i and a SCOP fold j. This number is calculated by enumerating all the 10,197 single-domain proteins with known CATH and SCOP classifications (PCsSs), and it is an estimate of T_{ij}, the joint probability distribution for a protein to have CATH topology i and SCOP fold j. If the CATH and SCOP classifications had been independent, every element T_{ij} could have been expressed as a product of C_i, the fraction of proteins that belong to CATH topology i, and S_j, the fraction that belongs to SCOP fold j, that is, $T_{ij} = C_i * S_j$. Randomly placing 10,197 proteins using such a probability distribution yields 4780 ± 40 nonzero elements in the matrix. In the other extreme case, if there had been a full correspondence between the SCOP and CATH classifications, the table would have had a single nonzero element in each row and column (in each CATH topology row the nonzero element would have been in that SCOP fold column that corresponds to it). In this case, the proteins in PCsSs would have been distributed among 284

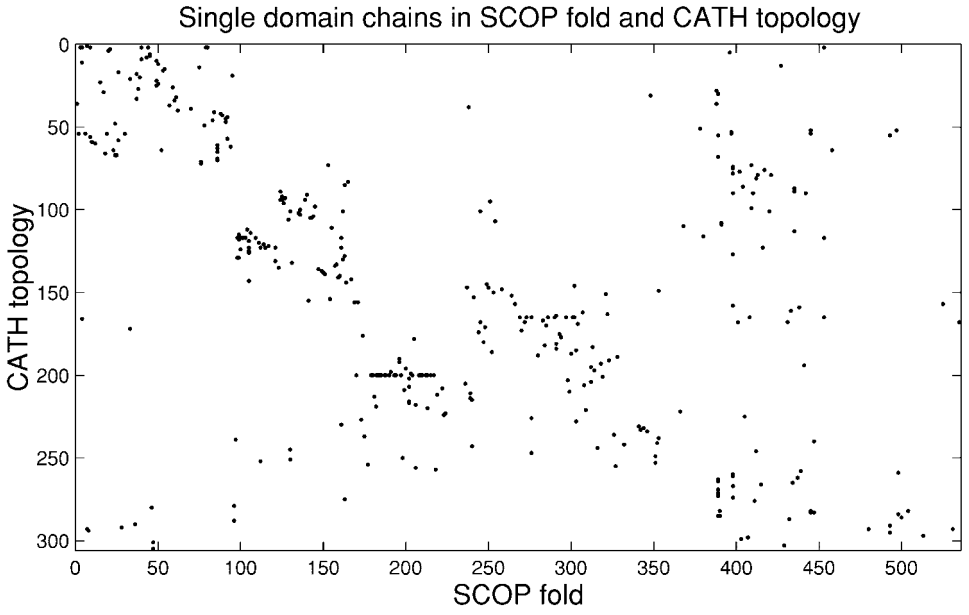

Fig. 2. Translation table from the CATH topology to the SCOP fold and vice versa. Nonzero entries of $\hat{T}_{ij}$ appear as black dots. $\hat{T}_{ij}$ is proportional to the number of proteins of CATH topology *i* that have a SCOP fold *j* in PCsSs.

nonzero elements (the number of distinct CATH topologies in PCsSs).

We found 369 nonzero elements in $\hat{T}$, meaning that the CATH and SCOP classifications are highly dependent. Still, the correspondence is not entirely one-to-one; in general, more than one SCOP fold corresponds to a given CATH topology. The number of such folds is, however, typically small. Such a translation table may be used to predict the SCOP classification of a structure already classified in CATH or at least to significantly restrict the number of possibilities and vice versa. For example, the assignment of the CATH topology to a protein with known SCOP fold can be done by selecting the CATH topology with the largest value in the translation table for that particular SCOP fold. Such an assignment is correct in 93% of the cases. The corresponding assignment of the SCOP fold from the CATH topology is correct in 82% of the cases. Although this is possibly useful information, in this work we do not assign classifications in this way.

SUMMARY OF THE COCLASSIFICATION PERFORMANCE

Every time the FSSP Z-scores are updated (once a week) the CO classification can be applied to all the proteins that appear in the new FSSP release but are not yet classified in CATH or in SCOP. The possible outcomes of the classification procedure are as follows:

1. Correct classification: the predicted classification will agree with the future release of the databases.
2. Rejection: the program is unable to classify the structure.
3. Ambiguous classification: a classification is returned (both for CATH and SCOP), but a later release provides a different classification.

The frequencies of these outcomes greatly depends on the statistics of the set of proteins to be classified. More specifically, rejected proteins are of two types: proteins that do not have high Z-scores with any other proteins ("islands"; see Materials and Methods) and clusters of proteins that are similar among themselves but do not have high Z-scores with other proteins outside their cluster ("superislands"). The fraction of islands and superislands is a feature of the particular set of proteins to be classified. The occurrence of a superisland suggests that a new classification type (a new topology for CATH and and new fold for SCOP) might be needed. The work of maintaining CATH and SCOP can be thus focused on the classification of a representative from each of these superislands.

For the set PFCs, the fraction of islands and superislands is 5%. We used this set to provide an upper bound for the performance of the CO method (see below); however, for the set PF$\bar{C}$ the fraction of rejections goes up to 22%. If rejections are not counted, we classify correctly 98% of the PFCs proteins. On the other hand, we could test our predictions also against the new CATH release v2.0. Of 1582 proteins that were assigned to previously existing CATH topologies, CO has classified correctly 80%. The difference in success rates between PFCs and PF$\bar{C}$ is due to the different way in which the test set is nested in the larger set of structures with known classification. In the first case, the test set consisted of 20% of the members of PFCs, selected at random; the remaining 80% were used to "predict" the classification of the test set. In the second case, the members of CATH v1.7 were used to predict the classification of the new proteins that were added when

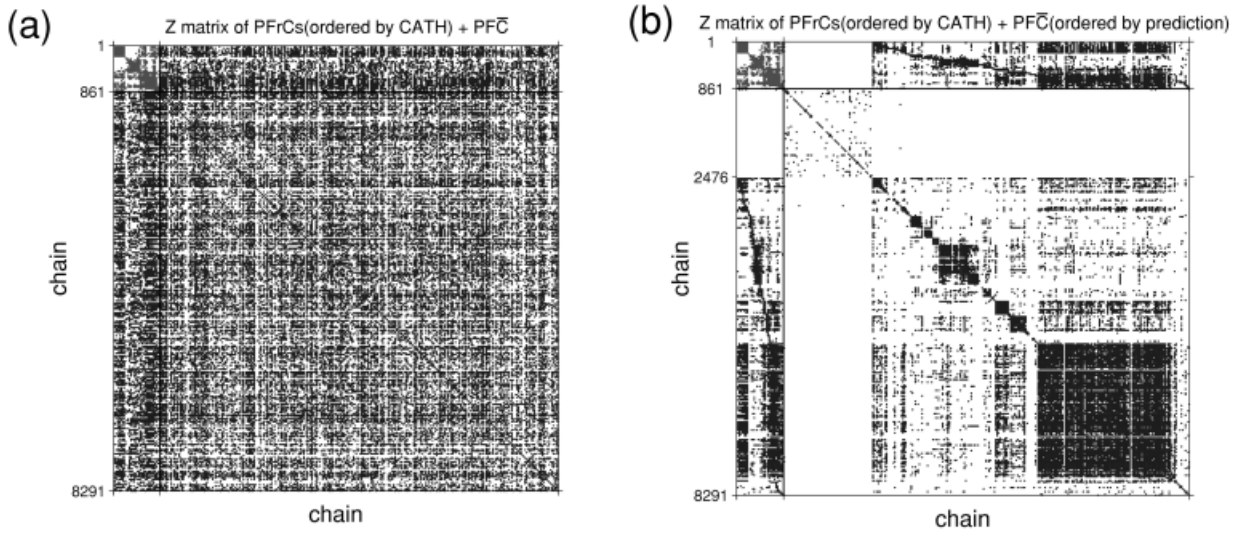

Fig. 3. **a:** Z-score matrix between all pairs of proteins in the combined PFrCs + PFC̄ sets. The submatrix in the upper left corner is the reordered Z-score matrix of the set PFrCs, which was already shown in Figure 1(b). The rest of the matrix presents the Z-scores for the proteins in the set PFC̄. **b:** The same matrix as in (a) with the rows and columns relative to the proteins in PFC̄ reordered according to our assignment of their CATH topology. With the CO method, the original order in the submatrix PFrCs is propagated to the entire matrix.

CATH v2.0 was released. These new structures are not distributed uniformly at random among the members of CATH v1.7.

Ambiguous classifications are due to two different mechanisms. The first stems from a well-known problem with the way the FSSP similarity index is calculated (the "Russian doll effect"; see below). The second kind of "mistake" is actually not a wrong classification; rather, it happens when the newly classified structure lies within the ambiguous "twilight zone" between two closely related topologies (for CATH) or folds (for SCOP), as demonstrated in detail below.

Automated Assignment of CATH Classification From FSSP

In this section we describe the procedure that we used to predict CATH topology level from the FSSP scores. We identified a set of 7431 proteins (PFC̄; see Materials and Methods) that appear in FSSP but were not yet processed by CATH 1.7. Our goal is to predict the CATH topology of these 7431 proteins by using (a) the Z-scores between all proteins in PF (see Materials and Methods) and (b) the known classifications of the set PFrCs (see Materials and Methods).

Predicting topologies is a classification problem that we treated with pattern recognition tools. We tested several prediction algorithms using cross-validation to estimate their performance.[21] Every one of the algorithms that were tested can be viewed as a two-stage process. In the first stage, a new similarity measure is produced from the original Z-scores. This is done either by a direct rescaling of the original Z-scores or by using the results of various hierarchical clustering methods to produce new similarity measures. The second stage consists of using these similarities as the input to some classification method, yielding predictions for the classes and architectures. In this work we present only results obtained by one particular method (CO), which uses the original Z-score as a similarity measure (see Materials and Methods). A complete list of the results obtained by using other methods can be found in Ref. 21, which is available on the web site.[31]

Our final assignments for the set PFC̄ using the CO method are listed in the web site. A more illustrative way to present these results is shown in Figure 3. In Figure 3(a) we present the Z-score matrix for the combined set PFrCs + PFC̄. The submatrix in the upper left corner is the reordered Z-score matrix of the set PFrCs, which was already shown in Figure 1(b). The rest of the matrix in Figure 3(a) presents the Z-scores of PFrCs with the set PFC̄ (randomly ordered) and the Z-scores of PFC̄ among themselves. In Figure 3(b) we reordered the rows and columns whose index was >860, corresponding to proteins in PFC̄. Although in the matrix of Figure 3(a) these proteins appear in a random order, in Figure 3(b) they appear in the order imposed by our prediction of their CATH topology. One can see that the original order in the submatrix PFrCs is propagated by our assignment procedure to the set PFC̄. For example, focus on the small black square at the upper left corner of the matrix. This small black square represents the high Z-scores among the mainly-α class of proteins in PFrCs. In the corresponding top rows of the full matrix we see high Z-scores between these structures and some proteins from PFC̄. In particular, the small group with indices near 2476 are "close" to these mainly-α structures and hence are also classified as such. On the other hand, there is a large group of structures from PFC̄ (between 861 and 2476), which do not have high Z-scores with any of the proteins in PFrCs or with any of the other structures in PFC̄ with index >2476. Hence,

we are unable to classify this group of structures on the basis of their FSSP scores.

Figure 3(b) illustrates the central idea of this work. We perform a task that is intermediate between clustering and classification. We take proteins of known classification and we use them as fixed a priori values in a clustering procedure.

The overall success rate of our prediction estimated by cross-validation was 93%. To understand the significance of these success rates, we derived a statistical (see Materials and Methods) upper bound for this kind of prediction. This upper bound is 95% (see Materials and Methods), hence the figure of 93/95 = 98% given above.*

We estimated the accuracy of the prediction by using the following procedure. First, the set PFCs was randomly "diluted"; that is, we randomly chose a certain fraction of the proteins in PFCs and placed them in a test set, pretending that we did not know their classification. The FSSP scores of the entire set were then used to classify the test set. For each protein from the test set, we either return a predicted classification or reject the protein (i.e., we declare that we are unable to classify it). The quality of any classification algorithm (see Materials and Methods) is measured by its success rate (fraction of correctly classified proteins, out of the test set) and by the purity (success rate out of the nonrejected proteins). For the CO method, the results were 93% for the success rate and 98% for the purity (using a dilution of 20%). More extensive tests at other dilutions and for other methods are of classification are discussed in Ref. 21 and available at the web site.[31]

We also tested directly the reliability of the CO assignments by using the CATH version 2.0 (PC2). In PC2, 1640 single-domain proteins that are present in $PF\bar{C}$ were assigned to one of the topologies that existed in v1.7. Fifty-eight of these we "rejected." In 1266 cases of the remaining 1582 (80%), our prediction agrees with the one given in CATH v2.0. Almost all the cases in which we misassigned a domain can be explained in a simple way. These cases are discussed in detail in a following section.

The CO method can also be used to predict directly the C level and the A level of CATH. We found that when the C and A levels were predicted as a byproduct of predicting the T level, the resulting C and A were consistent with those predicted directly.

Automated Assignment of SCOP Classification From FSSP

We used the CO method to predict the SCOP fold for a set of 3451 proteins ($PF\bar{S}$) that belong in PF but not yet in PS. The results are available on the web site.[31] The estimated success rate (by cross-validation) was 93%. As in the case of CATH, this number increased when we discarded proteins in the "twilight zone" (see the next section).

Twilight Zone for Protein Classification

The attempt to assign a new protein to a known fold might lead to frustration because at times one is undecided about two or more possibilities. To assess that two proteins have similar structures, a similarity score is needed. FSSP uses the Z-score, CATH uses the SSAP score, and SCOP uses a subjective evaluation, which is also a kind of score. The problem arises when the protein to be classified has high scores with two proteins already classified, but to different topologies. In this article, these proteins are called *borders* (see Materials and Methods). Being a border protein depends on the similarity score. We showed, however, that FSSP, CATH, and SCOP are to a large extent consistent classifications. Therefore, we suggest that there are "intrinsically" ambiguous cases—cases that are unavoidable in structure comparison. We refer to these ambiguous regions in structure space as the "twilight zone" in analogy with the case of protein sequence comparison where proteins with sequence similarity below 30% cannot be reliably assigned to the same fold. We illustrate this concept by a typical case, shown in Figure 4. This is a border protein. Protein 1dhn (the central one) is the one to be classified (in fact, it is a three-layer sandwich according to CATH). It has a Z-score of 9.3 with protein 1a8rA (on the left), which is a three-layer sandwich topology and a Z-score of 8.7 with protein 1b66A (on the right), which is a two-layer sandwich topology. This example illustrates how structural information alone might not provide a clear-cut criterion for classification of this protein. The incidence of the twilight zone is shown in Figure 5. In Figure 5(a) we present the histogram of the number of protein pairs that have different CATH topologies as a function of their Z-score. This number is a rapidly decaying function of Z. On the contrary, the number of pairs with the same CATH topology is a slowly decaying function of Z. For $Z > 3$, the probability of having the same CATH topology becomes greater than that of having different topologies. For $Z > 7.5$, the probability to have the same topology is 97.5%. In Figure 5(b) we show the corresponding figure for SCOP. The number of folds in SCOP is larger than the number of topologies in CATH; therefore, there is more ambiguity. However, also in this case for $Z > 7.5$, the probability to have the same topology is 93.5%. Taken together, these results indicate that the twilight zone for structure comparison can be bound by $Z \leq 7$.

There are other cases in which the classification of a particular protein is inconsistent with that of all its neighbors. For example, proteins that we called *colonies* (see Materials and Methods) are such that none of their neighbors are of their own kind. This means that the FSSP scores imply that these proteins are similar only to proteins of different classes and architectures. Identifying these proteins can also focus the attention to possible misclassification or to drawbacks of the Z-score. For example, 1 of the 49 colonies (at the architecture level) that

* One must keep in mind that the estimated success rate is calculated for all proteins; both FSSP representatives (≈10% of the proteins) and nonrepresentatives. Because the presence of homologous proteins can create a bias in these estimates, we also tested the success rate of predicting the CATH topology only for the FSSP representatives, which yielded 63%, to be compared with the corresponding upper bound of 74%.

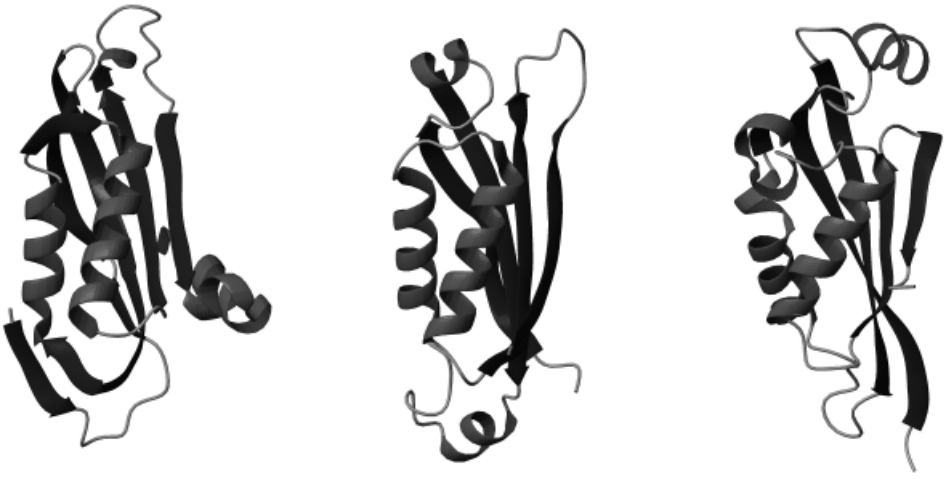

Fig. 4. **Center:** Protein 1dhn, which has a CATH α β three-layer (ββα) sandwich Aspartyl-glucosaminidase chain B (3.50.11) topology. **Left:** Protein 1a8rA, which has also a CATH α β three-layer (ββα) sandwich Aspartylglucosaminidase chain B (3.50.11) topology and has Z-score of 9.3 with protein 1dhn. **Right:** Protein 1b66A, which has a CATH α β two-layer sandwich Tetrahydropterin Synthase, subunit A (3.30.479) topology and has Z-score of 8.7 with protein 1dhn. This example illustrates how structural information alone might be insufficient to provide a clear-cut criterion for the classification of this protein.

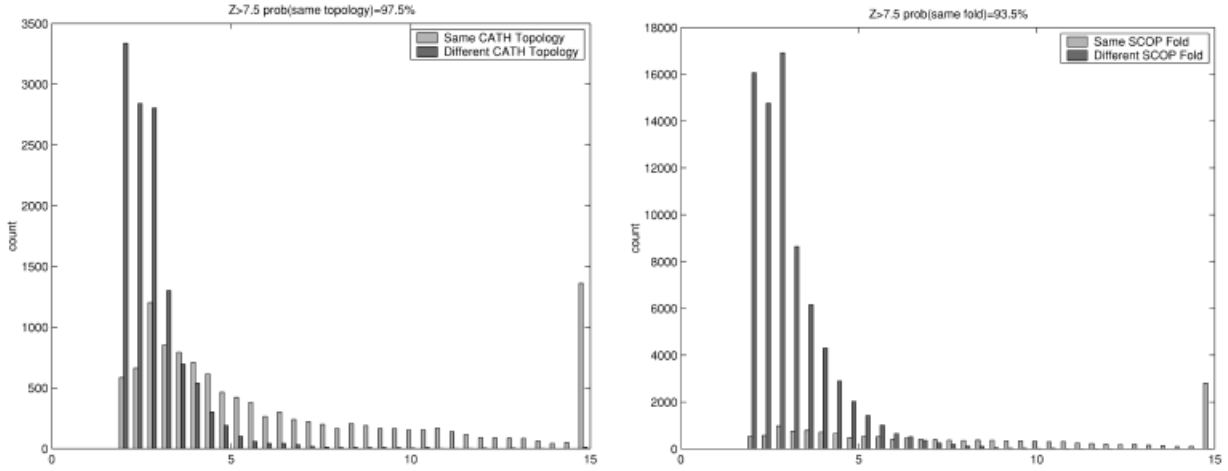

Fig. 5. Twilight zone for protein structure classification. **a:** The number of protein pairs of with a given FSSP Z-score that have different CATH folds is a rapidly decaying function of *Z*. On the contrary, the number of proteins pairs with the same CATH fold is decaying slowly. For $Z < 5$ there is a non-negligible probability to have different folds. We call this threshold the "twilight zone for structure classification." **b:** The corresponding histogram for SCOP folds. The number of SCOP folds is larger than the number of CATH topologies; hence the twilight zone is $Z \simeq 7$.

we found in CATH is the PDB entry 1rboC, which is classified as a α-β two-layer sandwich. It has 15 neighbors in PC, 14 of which are classified as mainly-β sandwiches.

We summarize the results about the assignments of the CATH architecture for proteins that already have a CATH classification (PFCs) in a "confusion table" (see Table II). The first column lists the "correct" classification (as given in CATH v1.7 for the test set); the second column gives the assignments by CO (correct, incorrect, or reject), and the third column lists the corresponding percentages. A full list of the inconsistent proteins is available on the web site.[31]

Another problem is that there are some large Z-scores between proteins of different architectures. Such large Z-scores arise when a protein of one particular architecture has a similar structure to a part of a protein of a different architecture. Swindells et al.[32] call the phenomenon of structures within structures, the "*Russian doll" effect.* Such cases are common between architectures of long proteins that contain substructures corresponding to architectures of shorter proteins; for example, there are many two-layer sandwich proteins that resemble a part of three-layer sandwich proteins. Such relationships can occur at the class level [e.g., α-β proteins that contain mainly-α or mainly-β proteins (1rboC, 1hgeA)]. They can also occur at the architecture level within the same class [e.g., α-β complex architecture contains α-β two-layer sandwich (1regX)]. Other inconsistencies occur when proteins fit two architecture definitions.

TABLE II. Summary of a "Confusion Table"

Original classification	Assigned classification	Cases (%)
Mainly alpha		
1.10 Orthogonal bundle	1.10 Orthogonal bundle	96.3
	reject	3.3
1.20 Up-down bundle	1.20 Up-down bundle	97.7
	4.10 Irregular	1.2
	1.10 Orthogonal bundle	0.7
1.25 Horseshoe	1.25 Horseshoe	100.0
1.50 Alpha-alpha barrel	1.50 Alpha-alpha barrel	100.0
Mainly beta		
2.10 Ribbon	2.10 Ribbon	93.9
	reject	5.7
2.20 Single sheet	2.20 Single sheet	97.2
	reject	2.3
2.30 Roll	2.30 Roll	97.2
	reject	2.1
	3.10 Roll	0.7
2.40 Barrel	2.40 Barrel	91.0
	reject	8.8
2.50 Clam	2.50 Clam	94.4
	2.40 Barrel	5.6
2.60 Sandwich	2.60 Sandwich	86.1
	reject	13.9
2.70 Distorted sandwich	2.70 Distorted sandwich	96.1
	2.60 Sandwich	3.9
2.80 Trefoil	2.80 Trefoil	100.0
2.90 Orthogonal prism	2.90 Orthogonal prism	100.0
2.100 Aligned prism	2.100 Aligned prism	100.0
2.102 3-layer sandwich	2.102 3-layer sandwich	78.6
	2.30 Roll	21.4
2.110 4 Propellor	2.110 4 Propellor	100.0
2.120 6 Propellor	2.120 6 Propellor	96.1
	reject	3.9
2.130 7 Propellor	2.130 7 Propellor	100.0
2.140 8 Propellor	2.140 8 Propellor	85.3
	reject	14.7
2.160 3 Solenoid	2.160 3 Solenoid	100.0
2.170 Complex	2.170 Complex	83.3
	2.60 Sandwich	8.6
	reject	8.0
Mixed alpha-beta		
3.10 Roll	3.10 Roll	99.9
3.20 Barrel	3.20 Barrel	100.0
3.30 2-layer sandwich	3.30 2-layer sandwich	93.5
	reject	6.0
3.40 3-layer(aba) sandwich	3.40 3-layer(aba) sandwich	96.1
	reject	3.8
3.50 3-layer(bba) sandwich	3.50 3-layer(bba) sandwich	72.1
	reject	27.2
	3.30 2-layer sandwich	0.7
3.60 4-layer sandwich	3.60 4-layer sandwich	99.7
3.70 Box	3.70 Box	100.0
3.75 5-stranded propeller	3.75 5-stranded propeller	100.0
3.80 Horseshoe	3.80 Horseshoe	100.0
3.90 Complex	3.90 Complex	97.9
	reject	0.7
Few secondary structures		
4.10 Irregular	4.10 Irregular	90.8
	reject	8.3
	1.20 Up-down bundle	0.8

This table summarizes the results about the assignments of the CATH architecture for proteins that have already a CATH classification. Only cases that occur >0.5% are listed. These figures were calculated by using 100 cross-validation runs at 20% dilution.

TABLE III. The Search Result When Submitting "1cuoA" to the Web Site http://www.weizmann.ac.il/physics/complex/compphys/f2cs/

	CATH v1.7				CATH v2.0				CATH prediction			SCOP 1.53			SCOP prediction	
Chain id	#	C	A	T	#	C	A	T	C	A	T	#	C	F	C	F
1cuoa	−1				1	2	60	40	2	60	40	−1			2	5

This protein was classified by neither CATH v1.7 nor SCOP 1.53, which are the basis of our predictions. We predicted it to belong to CATH topology 2.60.40 and SCOP fold 2.5. Later it was indeed classified by CATH v2.0 as 2.60.40. The −1 in both CATH v1.7 and SCOP 1.53 represents that it was not classified by them.

Class Prediction Using the Web Site

To retrieve our prediction for the CATH topology or SCOP fold of a protein, one can use the web site[31] by entering the protein chain identifier in the search box and submitting the query. If the protein appears in our database, then a table will be returned containing both the known and the predicted SCOP and CATH classifications. For example, the submission of the chain identifier "1cuoA" returns Table III. This protein was classified by neither CATH v1.7 nor SCOP 1.53, which are the basis of our predictions. We predicted it to belong to CATH topology 2.60.40 and SCOP fold 2.5. Later, the release CATH v2.0 identified 1cuoA as 2.60.40.

CONCLUSIONS

The rapidly increasing number of experimentally derived protein structures requires a continuous updating of the existing structure classification databases. Each group adopts different classification criteria at the level of sequence, of structure, and of function similarities. A comparison between different classification schemes can help to understand the optimal interplay between different levels, it can reveal possible misclassification, and it can ultimately offer a fully automated updating procedure. Manual steps can be automated in an ever-increasing way by using the tools made available by other databases.

In this work we showed that it is possible to automatically predict the CATH topology and the SCOP fold from the FSSP Z-scores. It is possible to submit a protein of unknown CATH or SCOP classifications but known FSSP Z-scores to the web site[31] to obtain its CATH and SCOP classifications. Because the FSSP database is updated weekly, our procedure offers the possibility to update also CATH and SCOP with the same frequency (at least down to the topology and fold level, respectively). We introduced a classification method that clusters together structures of known and unknown classification according to their Z-scores. When proteins outside the twilight zone for structure comparison are considered, our method is highly reliable. We suggest that, to classify proteins within the twilight zone, other classification criteria, based on sequence and function similarity, must be adopted.

The advent of genome projects is multiplying the efforts in the field of protein classification. In the past, the aim was to find the structure of the particular protein that was interesting at a given time. Now the hope is to find a large representative set of structures that can encompass most of the existing folds, possibly all of them.[3] In such a large-scale project, human intervention, which is precious in setting the principles of classification, should be gradually replaced by automated procedures.

ACKNOWLEDGMENTS

We thank Liisa Holm for making the raw FSSP data available to us and for useful discussions during the initial stages of this project. This work is based on a thesis for the M.Sc. degree submitted by G.G. to Tel-Aviv University (1998). We also thank Noam Shental for discussions. M.V. is supported by an European Molecular Biology Organization (EMBO) long-term fellowship; he also thanks the Einstein Center for Theoretical Physics for partial support of his stay at the Weizmann Institute. D.S. thanks the Weizmann Institute of Science for hospitality while part of this work was carried out.

MATERIALS AND METHODS

Databases and Protein Sets

Because the CATH and SCOP databases classify domains and FSSP deals with chains, we considered only chains that form a single domain; therefore, these proteins appear as a single entry in the three databases. Several groups have developed methods to identify protein domains.[20,23,33–35] In this work, we used the Dali Domain Dictionary[24] to identify single-domain proteins.

We used the following databases. The CATH release 1.7, which contains 15,802 protein chains, among which 10,906 are classified as single domain. This latter set is called PCs. We also used the CATH release 2.0, which contains 20,780 protein chains, among which 14,389 are single domain (PC2s). The SCOP release 1.53, which contains 20,021 protein chains, among which 15,375 are single domain (PSs). The FSSP release from 14 January 2001, which contains 22,660 protein chains (PF). The FSSP proteins are grouped into 2,494 homology classes so that within a class the sequence similarity is >25%. One protein per class is selected as representative, and we call PFr the set of all representatives. All the protein sets and their sizes are listed in Table IV.

Classification by Optimization (CO) Method

The classification scheme that we used is based on the minimization of a particular cost function, defined as follows (for the case of the prediction of CATH topology; a similar definition holds for SCOP folds). Each protein is

TABLE IV. Protein Sets and Their Sizes

Name	Description	Size of set
PF	All chains in FSSP (14 Jan, 2001)	22,660
PFr	Representative chains in FSSP (14 Jan, 2001)	2,494
PC	Chains in CATH v1.7	15,802
PCs	Single-domain chains in CATH v1.7	10,906
PC2	Chains in CATH v2.0	20,780
PC2s	Single-domain chains in CATH v2.0	14,389
PS	Chains in SCOP 1.53	20,021
PSs	Single-domain chains in SCOP 1.53	15,375
PCsSs	Single-domain chains in SCOP 1.53 and CATH v1.7	10,197
PFrCs	Single-domain chains in CATH that are representatives FSSP (PFr ∩ PCs)	860
PFrSs	Single-domain chains in SCOP that are representatives FSSP (PFr ∩ PSs)	1,626
PFCs	Chains in FSSP and single domain in CATH v1.7	10,541
PFSs	Chains in FSSP and single domain in SCOP 1.53	14,716
PF$\bar{C}$	Chains in FSSP and not in CATH v1.7	7,431
PF$\bar{S}$	Chains in FSSP and not in SCOP 1.53	3,451

assigned an integer number c_i, describing its topology (1–305). We assign to proteins with known classification the value of $c(i)$ determined by their CATH classification. To the yet unclassified proteins we assign initially random values from 1 to 305. A cost is calculated for each configuration $C = \{c_i\}$ of topologies, which penalizes the assignment of different topologies to any pair of proteins. The value of this penalty is chosen to be the similarity measure Z_{ij} between proteins i and j; the higher the similarity Z_{ij}, the more costly it is to place proteins i and j in different topologies. The cost function is defined as the sum of penalties for all protein pairs $\langle i, j \rangle$,

$$E(C) = \sum_{\langle i,j \rangle} Z_{ij}[1 - \delta(c_i, c_j)]. \quad (1)$$

The classification problem is stated as finding the minimal cost configuration of the unclassified proteins, while keeping the topologies (i.e., the c_i values) of the classified proteins fixed. This problem corresponds to finding the ground state of a random field Potts ferromagnet.

We search for a classification C of minimal cost by an iterative greedy algorithm described in detail elsewhere.[21] The algorithm identifies at which iteration, if any, it performed a heuristic decision. For low fractions of unknown topologies, the algorithm usually reaches the global minimum of the cost function.

Bounds on the Success Rate of the Prediction

In this section we establish a statistical upper bound for the prediction success rate relevant to a family of prediction algorithms.

The Z-matrix can be reinterpreted as a weighted graph; each vertex in the graph represents a protein and the weights on the edges connecting two vertices are the corresponding Z-scores. Edges with $Z < 2.0$ are absent from the graph. Following this representation, we define two proteins as neighbors if they are connected by an edge. By analyzing the connectivity properties of set PC we make inferences about our predictive power.

One can characterize the FSSP-based neighborhood of a protein according to the CATH classification of itself and its neighbors. Every protein must belong to one of four categories:

"Island": The protein has no neighbors.

"Colony": It has no neighbors of its own kind.

"Border": It has neighbors of its own kind as well as of other kinds.

"Interior": The protein has only neighbors of its own kind.

Using these definitions we can arrange the proteins of PC in groups according to their neighborhood category at the class, architecture, and topology levels. The distribution of the proteins among these groups can be used to calculate an upper bound for the CO method, if we assume that the set of unclassified proteins has the same distribution as the classified ones. For example, islands cannot be classified and are therefore rejected. Colonies are bound to be misclassified because none of their neighbors give a clue on their type. Because the fraction of proteins in each category was estimated on the basis of a sample, it can be interpreted only as a statistical upper bound.

We consider the set PFCs to obtain a first type of upper bound for the success rate of the CO method. This set (see Table IV) is formed by 10,541 proteins, among which 5% are islands, a negligible fraction (0.2%) are colonies, 6% are borders, and 88% are interiors. Therefore, the upper bound that we found is about 95% for predicting the topology level in CATH.

The actual prediction performed in this work is done on the set PF$\bar{C}$, which is formed by the 7431 proteins that are in FSSP (14 January 2001) but not in CATH1.7 (see Table IV). Within PF$\bar{C}$ there is a subset of 1617 (about 22%) proteins that are either islands or superislands, that is, they are connected only with other proteins in the subset and therefore they have no connection to proteins with known classification. Thus, the upper bound for this second type of prediction is about 78%.

Evaluating a Classification Prediction Algorithm

Because an algorithm can output either a predicted classification or a "rejection," if it does not have any prediction, one has to estimate two probabilities: P_{success} and P_{reject}. Robust estimation of these parameters is produced by cross-validation, a procedure that consists in averaging over many (T) randomly sampled test trials. In each trial, the set is divided into two subsets; one is used for training the algorithm and the other set, of N_{test} proteins, is used to test the algorithm by comparing its prediction to the true classification. The probability estimates are given by

$$\hat{P}_{\text{success}} = 1/T \sum_{t=1}^{T} \frac{N_{\text{success}}}{N_{\text{test}}} \tag{2}$$

$$\hat{P}_{\text{non-reject}} = 1 - \hat{P}_{\text{reject}} = 1/T \sum_{t=1}^{T} \frac{N_{\text{test}} - N_{\text{reject}}}{N_{\text{test}}} \tag{3}$$

Another figure of merit, the purity P_{pure}, is the probability of correctly classifying nonrejected proteins. It is estimated by

$$\hat{P}_{\text{pure}} = \frac{\hat{P}_{\text{success}}}{1 - \hat{P}_{\text{reject}}} \tag{4}$$

REFERENCES

1. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
2. Thornton JM, Orengo CA, Todd AE, Pearl FMG. Protein folds, functions and evolution. J Mol Biol 1999;293:333–342.
3. Šali A. 100,000 protein structures for the biologist. Nat Struct Biol 1998;5:1029–1032.
4. Martí-Renom MA, Ashley AC, Fiser A, Sanchez R, Melo F, Šali A. Comparative protein structure modeling og genes and genomes. Annu Rev Biophys Biomol Struct 2000;29:291–325.
5. Heger A, Holm L. Towards a covering set of protein family profiles. Prog Biophys Mol Biol 2000;73:321–337.
6. Bowie JU, Lüthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
7. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. Nature 1992;358:86–89.
8. Fisher D, Rice D, Bowie JU, Eisenberg D. Assigning amino acid sequences to 3-dimensional protein folds. FASEB J 1996;10:126–136.
9. Gerstein M, Levitt M. A structural census of the current population of protein sequences. Proc Natl Acad USA 1997;94:11911–11916.
10. Murzin AG. Structural classification of proteins: new superfamilies. Curr Opin Struct Biol 1996;6:386–394.
11. Blundell TL, Mizuguchi K. Structural genomics: an overview. Prog Biophys Mol Biol 2000;73:289–295.
12. Finkelstein AV. Protein structure: What is possible to predict now? Curr Opin Struct Biol 1997;7:60–71.
13. Simons KT, Bonneau R, Ruczinski I, Baker D. Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins 1999;37(Suppl 3):171–176.
14. Bernstein F, Koetzle T, Williams G, Meyer EJ, Brice M, Rodgers J, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
15. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 1997;25:231–234.
16. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
17. Conte LL, Ailey B, Hubbard TJP, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257–259.
18. Mizuguchi K, Deane CM, Blundell TL, Overington JP. HOMSTRAD: a database for protein structure alignments for homologous families. Protein Sci 1998;7:2469–2471.
19. Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opin Struct Biol 1996;6:377–385.
20. Siddiqui AS, Barton GJ. Continuous and discontinuous domains: an algorithm for the automatic generation of reliable protein domain definitions. Protein Sci 1995;4:872–884.
21. Getz G. Clustering and classification of protein structures. M.Sc. Thesis, Tel-Aviv University, 1998.
22. Hadley C, Jones DT. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 1999;7:1099–1112.
23. Holm L, Sander C. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 1994;22:3600–3609.
24. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L. A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res 2001;29:55–57.
25. Jain AK, Dubes RC. Algorithms for clustering data. Englewood Cliffs, NJ: Prentice-Hall; 1988.
26. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.
27. Taylor WR, Orengo CA. Protein structure alignment. J Mol Biol 1989;208:1–22.
28. Orengo CA, Brown NP, Taylor WR. Fast structure alignment for protein databank searching. Proteins 1992;14:139–167.
29. Bray JE, Todd AE, Pearl FMG, Thornton JM, Orengo CA. The CATH Dictionary of Homologous Superfamilies: a consensus approach to analyze distant structural homologues. Protein Eng 2000;13:153–165.
30. Brenner SE, Koehl P, Levitt M. The ASTRAL compendium for protein structure and sequence analysis. Nucleic Acids Res 2000; 28:254–256.
31. http://www.weizmann.ac.il/physics/complex/compphys/f2cs/index.html.
32. Swindells MB, Orengo CA, Jones DT, Hutchinson EG, Thornton JM. Contemporary approaches to protein structure classification. Bioessays 1998;20:884–891.
33. Islam SA, Luo J, Sternberg MJE. Identification and analysis of domains in proteins. Protein Eng 1995;8:513–525.
34. Swindells MB. A procedure for detecting structural domains in proteins. Protein Sci 1995;4:103–112.
35. Sowdahamini R, Rufino SD, Blundell TL. Nuclear dynamics and electronic transition in a photosynthetic reaction center. J Am Chem Soc 1997;119:3948–3958.