PROTEINS: Structure, Function, and Genetics 33:88–96 (1998)

Dictionary of Recurrent Domains in Protein Structures

Liisa Holm* and Chris Sander
EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, United Kingdom

ABSTRACT **The rapid growth in the number of experimentally determined three-dimensional protein structures has sharpened the need for comprehensive and up-to-date surveys of known structures. Classic work on protein structure classification has made it clear that a structural survey is best carried out at the level of domains, i.e., substructures that recur in evolution as functional units in different protein contexts. We present a method for automated domain identification from protein structure atomic coordinates based on quantitative measures of compactness and, as the new element, recurrence. Compactness criteria are used to recursively divide a protein into a series of successively smaller and smaller substructures. Recurrence criteria are used to select an optimal size level of these substructures, so that many of the chosen substructures are common to different proteins at a high level of statistical significance. The joint application of these criteria automatically yields consistent domain definitions between remote homologs, a result difficult to achieve using compactness criteria alone. The method is applied to a representative set of 1,137 sequence-unique protein families covering 6,500 known structures. Clustering of the resulting set of domains (substructures) yields 594 distinct fold classes (types of substructures). The Dali Domain Dictionary (http://www.embl-ebi.ac.uk/dali/) not only provides a global structural classification, but also a comprehensive description of families of protein sequences grouped around representative proteins of known structure. The classification will be continuously updated and can serve as a basis for improving our understanding of protein evolution and function and for evolving optimal strategies to complete the map of all natural protein structures. Proteins 33:88–96, 1998.** © 1998 Wiley-Liss, Inc.

Key words: fold classification; substructures; Dali; protein families; structural similarity

Introduction

It has long been recognized that proteins exhibit a modular architecture consisting of globular and compact substructures (structural domains) that may be autonomous folding units.[1] Sequence analysis has revealed numerous evolutionarily mobile domains, also called modules, that often carry specific biological functions.[2] In a number of cases, the structural autonomy of such modules has been verified by structure determination (e.g., SH2 and SH3 "src-homology" domains[3,4]). The notions of substructures, domains, modules, or folding units all relate to attempts to define protein architecture at an intermediate size level, i.e., intermediate between secondary structure elements and entire proteins or polypeptide chains. As experimental structure determination accelerates, the need for automatic processing of protein substructure organization using objective and quantitative criteria will increase.

Many algorithms for delineating domains from the three-dimensional coordinates of protein structures are based on measures of compactness, surface area, globularity, or chain flexibility.[5–9] In contrast, human intuition on domain assignment is largely guided by the visual recognition of recurrent folding patterns.[10] In our opinion, ignoring the information about recurrence is an important shortcoming of current domain decomposition algorithms. For example, if domains are computed independently for any given coordinate set, then there is no guarantee for consistent parsing between structures that are evolutionarily related members of a protein family. In this work, we address the problem of consistency by introducing the concept of recurrence into automated domain recognition. Recurrent domains are substructures with similar folds that are found in different combinations in distantly related or functionally unrelated proteins (Fig. 1).

We quantify recurrence in terms of the statistical significance of structural similarity for many pairs of domains. The statistical significance is highest for structural similarities that involve large units and that completely cover a substructure unit. Exploiting these effects, we define a sum-of-pairs objective

*Correspondence to: Liisa Holm, EMBL-EBI, Wellcome Trust Genome Campus, CB10 1SD Cambridge, United Kingdom.
Received 27 January 1998; Accepted 7 May 1998

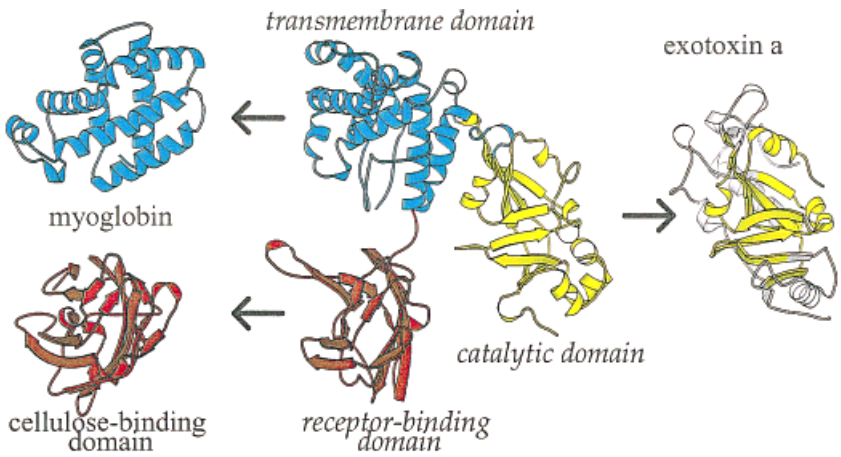


Fig. 1. Recurrent domains in diphtheria toxin (1ddt).[24] Structural domains are the basic units of protein architecture, function, and evolution. A beautiful example is diphtheria toxin, made up of three domains, each of which is involved in a different stage of infection (receptor binding, membrane penetration, and catalysis of ADP-ribosylation of elongation factor 2). A structural neighbor is depicted next to each domain of diphtheria toxin (middle). Drawn with Molscript v.2.[25] The crystallographers delineated three domains spanning residues 1–187, 200–380, and 381–535 (188–199 not traced). The automatic algorithm delineates recurrent domains as (1–66,77–170), (67–76, 171–187, 200–376), and 377–535. The difference between the two definitions is in the border between the catalytic and transmembrane domains, where the automatic algorithm trims the catalytic domain to the part that is common between diphtheria toxin and its relative exotoxin, enterotoxin, and pertussis toxin and poly(ADP-ribose) polymerase.

function that (1) favors recurrences of large substructures with distinct topological arrangement and packing of secondary structure elements, and (2) disfavors small substructures consisting of one or two secondary structure elements despite their higher frequency of recurrence. Although other formulations of the optimization problem are possible, the empirically chosen objective function combined with a heuristic algorithm for optimization yields a useful set of substructures (domains). Although we do not foresee that automatically delineated domains will be accepted as the gold standard of the trade, the domain definitions resulting from the present method are consistent within protein families and often coincide with biologically functional units. The method recovers the well-known folding topologies with many members and yields clusters with good coverage of common secondary structure elements, providing a useful basis for large-scale structure analysis and classification.

METHODS

Structure Dataset

In structure comparison and classification, it is practical to work in a representative subset of structures that is purged of sequence redundancy. We use a representative set that contains no sequence similar protein pair *(A, B)* with sequence identity above the threshold T (adapted from Sander and Schneider[11]):

$$T(L_A, L_B) = 290.15 * \left(\frac{L_A + L_B}{2}\right)^{-0.562} \quad (1)$$

where L is the length of the chain, and sequence identity is computed from the structure alignment. The threshold T levels off at 25% identity for chains longer than 80 residues and is higher for shorter chains. The threshold T was originally derived to identify structurally conserved homologs based on sequence similarity to a known structure.[11] Analogous safe thresholds have been recently restated using more sophisticated measures of sequence similarity.[12–14]

The representative set is constructed using algorithm 1 of Hobohm et al.[15] Briefly, each structure is assigned a quality index as a function of crystallographic resolution, secondary structure content, number of undefined atomic coordinates, etc. The list of structures in the Protein Data Bank (PDB)[16] is processed in order of decreasing quality, adding a new structure to the representative set unless it is sequence similar to a previous representative. These definitions yielded a mapping of all known structures to a representative set of 1,137 proteins of known structure (November 1997).

Based on equation 1, proteins in the representative set may be more than 25% identical over some part of the structure but are both retained in the representative set if unique segments reduce the global average identity to below 25%; the goal is to guarantee retention of every structurally unique motif in the database. In some cases, a biologically functional unit is divided over several chains in the PDB entry, but we currently treat each chain individually.

Structure Alignment

Structurally equivalent substructures between proteins are defined using the Dali method.[17–19] Struc-

ture similarity S is quantified as a weighted sum of similarities of intramolecular distances:

$$S(A, B) = \sum_{i \in \text{core}} \sum_{j \in \text{core}} \left(0.2 - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^{AB}} \right) \cdot e^{-(d_{ij}^{AB}/20\text{Å})^2} \quad (2)$$

where d_{ij}^A and d_{ij}^B are intramolecular Cα-Cα distances in proteins A and B, respectively; $d_{ij}^A = (d_{ij}^A + d_{ij}^B)/2$, a relative deviation of 0.2 (20 %) is the threshold of similarity; and the exponential factor downweights contributions from pairs at longer distances. The *core* is a set of equivalences between residues in A and B (i_A equivalent to i_B) analogous to sequence alignment. Null correspondences are allowed (e.g., loop segments in different conformation or domains of dissimilar architecture are not part of the core). In the present application, we additionally require that the aligned segments are collinear, i.e., $i_A < j_A$ implies $i_B < j_B$.

The algorithmic problem is to find the common core (structure alignment of A,B) that maximizes S (Eq. 2). The Dali system uses a collection of algorithms (see Holm and Sander[18] and references therein) ranging from fast to sensitive to efficiently identify significant structural similarities in all-on-all comparison of the database. The present analysis includes suboptimal alignments (internal duplications, repeats, etc.) obtained after disallowing reuse of equivalenced residue pairs (i_A , i_B).

Length Normalization of Structure Similarity Scores

The Dali-score (Eq. 2) is an open scale of structural similarity. The maximum score for comparing a globular structure to an identical copy increases with the size of the protein. Similarly, the expected score for random pairwise comparison increases with the number of amino acids L_A and L_B in proteins A and B. We have empirically determined the background distribution of similarity scores as a function of protein size from a large number of pairwise alignments of random protein structure pairs (allowing permutations; see Holm and Sander[17]). The relationship between mean score m and size $L = \sqrt{L_A L_B}$ is closely approximated by the polynomial

$$m(L) \approx 7.95 + 0.71L - 2.59 \cdot 10^{-4} L^2 - 1.92 \cdot 10^{-6} L^3, \quad L \leq 400 \quad (3)$$

As there were few data points above $L = 400$, for $L > 400$ we use an extrapolation that is a linear increment over $m(400)$. As a measure of the statistical significance of a pairwise comparison score S (Eq. 2), we use the Z-score defined as

$$Z(A, B) = \frac{S(A, B) - m(L)}{0.50 \cdot m(L)} \quad (4)$$

where the denominator is an estimation of the average standard deviation.

The above calibration of statistical significance is based on score statistics for uncut protein chains. We also apply the calibration to substructures resulting from cut chains (domains), which is justified by the notion of domains as independent folding units. In structure database searches with multidomain proteins, ranking structural neighbors by Z-scores (rather than S of Eq. 2) brings biologically interesting hits (with complete coverage of a compact substructure) to the top, even though the domains may be of very different sizes, e.g., the DNA binding, catalytic, and SH3-like domains of biotin repressor.[19]

Alternative Substructure Partitions

The Z-score implies a penalty for incomplete coverage of an aligned substructure, and its application requires predefined substructure units. To generate a set of candidate substructures, we use the PUU algorithm,[9] which yields a hierarchy (unfolding tree) of compact substructure units at all size levels. Here we modify the algorithm to use recurrence information to guide the division of the polypeptide chain into unfolding units.

The domain decomposition algorithm is based on the intuitive idea that folding units are compact and the interactions between them are weak. This intuition is made quantitative in a simple harmonic approximation of protein dynamics. The relative motion between compact domains is governed by the strength of nonbonded interactions across the interface and the distribution of masses. The most likely domain separation involves units for which the time constant of relative motion is largest.

Nonbonded interactions can be represented by a N*N contact matrix, where N is the number of residues in a structure, and interresidue contact strength is computed from interatomic distances using a simple potential function. The PUU algorithm[9] uses a linear-time heuristic that yields a binary division of a structure into two substructures while allowing multiple traversals of the chain across the domain boundary. This is based on reordering the polypeptide sequence by band diagonalization of the contact matrix and scanning for a single cut point in the reordered sequence that corresponds to the largest time constant of relative motion.

In this work, recurrence information is used to enhance the contrast of the contact map, so that residues belonging to the same recurrent motif are more likely to stay together in the unfolding process than residues that are not part of the same recurrent motif. For the reordering step, we add on to the

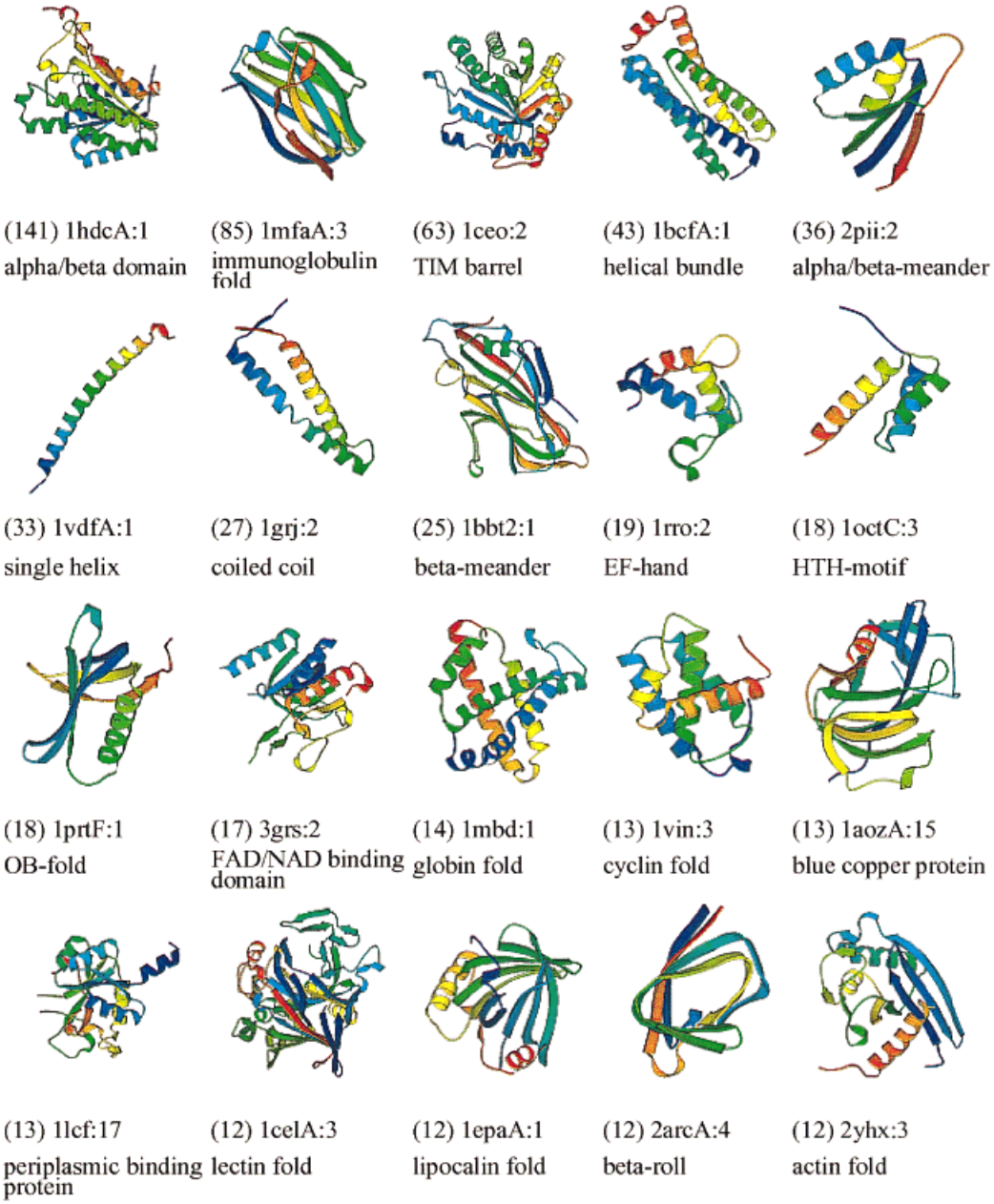


Fig. 2. Dominant domain fold types. The automatic identification of recurrent domains can be used to define a classification of fold types in the currently known protein structures. Here we show the 20 most populated fold types in the representative set (number of member domains in parentheses). The structure shown is a central member of its cluster, i.e., it has the largest sum of similarities to other members of the cluster. Structures are identified by the PDB identifier[chain]:node number in unfolding tree (node 1 is the uncut chain). A common name is listed with the fold types. Drawn using Molscript v.2[25] with rainbow coloring (blue-green-yellow-red) to indicate the course of the polypeptide chain.

contact map of the query protein, the contact maps of structurally similar proteins (same class in FSSP database; see Holm and Sander[20]) at structurally equivalent positions. The external structures are weighted according to their Z-score for pairwise structure comparison with the query structure. After some experimentation, we chose to scale the total relative external:query contributions to 1:1. The step of calculating time constants of motion uses only the original contacts and masses. With this procedure, the sequence of unfolding events is more concordant between structural homologs than with the original PUU method.[9] In particular, with TIM (triose isomerase) barrels, recurrence successfully balances the sometimes uneven distribution of masses. However, Achromobacter protease (1arb), with its very tight

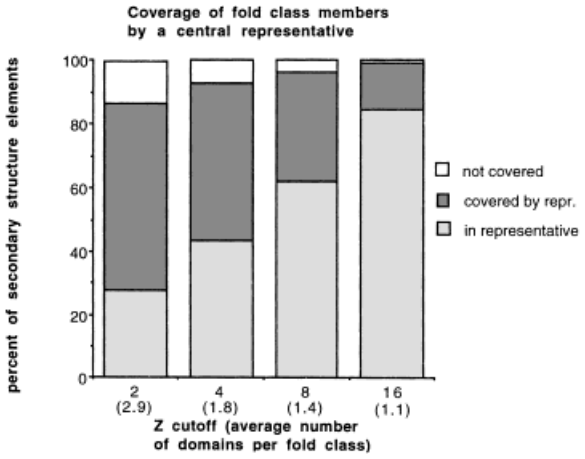


Fig. 3. How well are the diverse members of a fold class described by a single representative? To remove structural redundancy, domains are clustered into fold types by average linkage clustering (terminating at average similarity Z = 2 and subclusters defined at Z = 4, 8, and 16). Each cluster is represented by a central member, selected based on strongest average similarity to all other members of the cluster. The coverage of all members of the fold cluster by the single representative is here quantified in terms of secondary structure elements that are (1) part of the representative, (2) covered ("reused," structurally equivalent according to the Dali alignment) in other member domains, and (3) unique (additional elements, not structurally equivalent to the representative) in other member domains. Here, 100% corresponds to the total number of secondary structure elements in the representative set (generated by removing sequence redundancy) of 1,137 protein chains or 1,724 domains. To further remove structural redundancy, we select one central member domain to describe all members of a fold class. The compression is largest for the broadest definition of fold classes. For example, at a Z-cutoff of 2, the representatives (light gray bars) make up one-third-to one-fourth of secondary structure elements yielding a compression factor of 3–4. The severity of the approximation is indicated by the ratio of the white and dark gray bars. For example, at a Z-cutoff of 2, the structurally redundant domains contain about one in five elements in redundant domains that are not covered by the fold class representative. For typical size domains this means one or two not covered elements. Inspection shows that these are typically protuberances (short helices, beta-hairpins) outside a compact common core.

interdomain interface, remains pathological as the first division plane going through the middle of one of the internally duplicated beta-barrels.

Alternative partitions of the protein into substructure units are generated by traversing the branches of the unfolding tree to different depths. For example, a protein A may be taken as a whole (A), divided in two halves yielding the partition (A.1, A.2), and either or both halves may be further divided yielding the partitions (A.1, A.2.1, A.2.2), (A.1.1, A.1.2, A.2), and (A.1.1, A.1.2, A.2.1, A.2.2). We systematically generate all combinations of units in this fashion down to a unit size of 40 residues. The unfolding trees are usually well balanced, yielding in the order of 10 alternative partitions. Only for a few structures was the number of alternative partitions more than 1,000.

Size Selection

The Z-score is used as a quantitative criterion to identify optimal coverage of structural motifs chosen from among the alternative candidate substructures. The Z-score (statistical significance) increases with increasing size of the equivalent substructures, until the substructures grow too large and include increasingly nonequivalent portions (length increases while Dali-score remains the same) and the Z-score starts to decrease. Our goal is to multiply this effect by selecting a consensus set of substructures (domains) based on the all-on-all structural alignments within a representative set of proteins, so that many of the chosen substructures are common to different proteins at a high level of statistical significance.

In the present method, recurrence is quantified as the sum of Z-scores between the selected substructure units. Formally, we have to solve a global optimization problem of finding partitions p for each of N structures in the representative set that maximize recurrence R :

$$R = \sum_{i=1}^{N} \rho(p_i) \tag{5}$$

where the recurrence ρ of a partition of protein i that is made up of substructures a, is defined as

$$\rho(p_i) = \sum_{a \in p_i} \sum_{b \in c_a} Z_{a,b} \tag{6}$$

where $c_a = \{b_1, b_2, \cdots, b_n\}$ denotes a clique of substructures b (in the current selection) that are structural neighbors of substructure a. The clique is constructed by processing the neighbors of a in decreasing order of similarity, i.e., $Z_{a,b_1} \geq Z_{a,b_2} \geq \cdots Z_{a,b_n} \geq 2$, and requiring that $Z_{b_k,b_l} \geq 2.0$, $k, l = 1, \ldots, n$, i.e., all members of the clique have structural

TABLE I. Comparison of Domain Decompositions by Recurrence and Crystallographers†

Number of cases

		Crystallographers			
DDD	NA‡	1 domain	2 domains	3 domains	4 domains
1 domain	558	178	17		
2 domains	204	6	36	6	
3 domains	62	2	4	7	
4 domains	26	1	1	7	1
5 domains	11				
6 domains	5		1		
7 domains	3				
8 domains	1				

†The number of domains assigned to the representative set (1,137 chains) by our recurrence algorithm (DDD, rows) is cross-tabulated with the number of domains assigned by crystallographers (columns). For example, 178 chains are assigned as single-domain structures by both methods. A total of 267 chains are common between the representative set and the reference set of crystallographers' domain assignments compiled by Islam et al.[5]
‡PDB entry not in reference set.

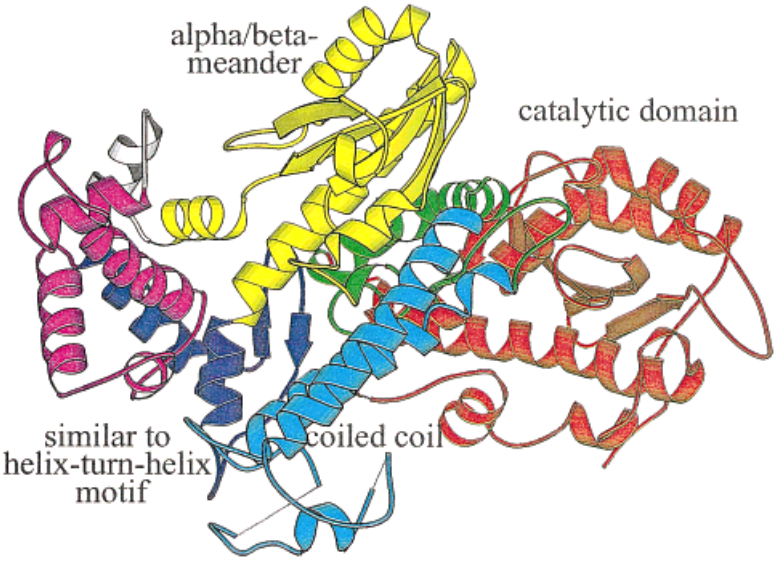


Fig. 4. Domain decomposition of the Klenow fragment of DNA polymerase (1klnA).[26] Depending on the scientific question, there can be alternative, equally reasonable ways to cut a large protein into domains. Crystallographers identified two domains in the Klenow fragment of DNA polymerase, the catalytic domain (red) and a second domain (the rest). Our algorithm identifies the catalytic domain as recurring in Taq and T4 polymerases; the second domain is cut into several structural units. The domain with the $\alpha\beta$-meander topology recurs in Taq polymerase, adenylyl cyclase, and reverse transcriptase; the latter two have only this domain in common with the Klenow fragment. Two other motifs that recur in a large number of proteins are two helices in coiled-coil arrangement, and three helices in similar arrangement as in the helix-turn-helix motif of several DNA-binding domains.

similarity scores above the threshold Z = 2 for each pairwise comparison. Counting only structural neighbors that form a clique ensures that one captures a consistently recurring motif, rather than bits and pieces from different parts of substructure *a*, and helps to reduce noise due to marginally similar structures. The final selection of units results from the size dependence of the Z-score and from the recurrence of the selected units in the representative set. Internal duplications of domains are included in the sum, but pairs of substructures that correspond to parent and child nodes in the unfolding tree are excluded.

In principle, equations 5 and 6 pose a complicated combinatorial problem, because the product of the number of alternative partitions for structures in the representative set is astronomically large. Given the heuristic nature of the objective function expressed in terms of Z-scores (the sum of statistical significance values has no physical interpretation), we used a simple greedy optimization procedure rather than more rigorous stochastic optimization algorithms. Initially, the partitions p are set to complete chains (no cutting). The selection is then refined by iterative cycles, and the iteration typically converges to a local optimum in fewer than 10 cycles. During an iteration cycle, each structure is considered in turn, and the recurrence ρ is evaluated for all the alternative partitions p_i of protein i, given the current field of selected partitions for the other proteins. The partition p_i with the highest recurrence ρ is selected, and R (Eq. 5) is updated. The combination of partitions that yields the maximal value of total recurrence R is retained.

Postprocessing

Fold types are defined by average linkage clustering of the domains based on their structural similarities, starting at high Z values and terminating at an average Z-score of 2. Based on the resulting fold dendrogram, singlets (single members of a cluster) are demoted to the category "decorations" if they are smaller than 80 residues, have only 2 or fewer secondary structure elements, and are not a complete chain from an experimental structure dataset. Decorations are not counted as domains.

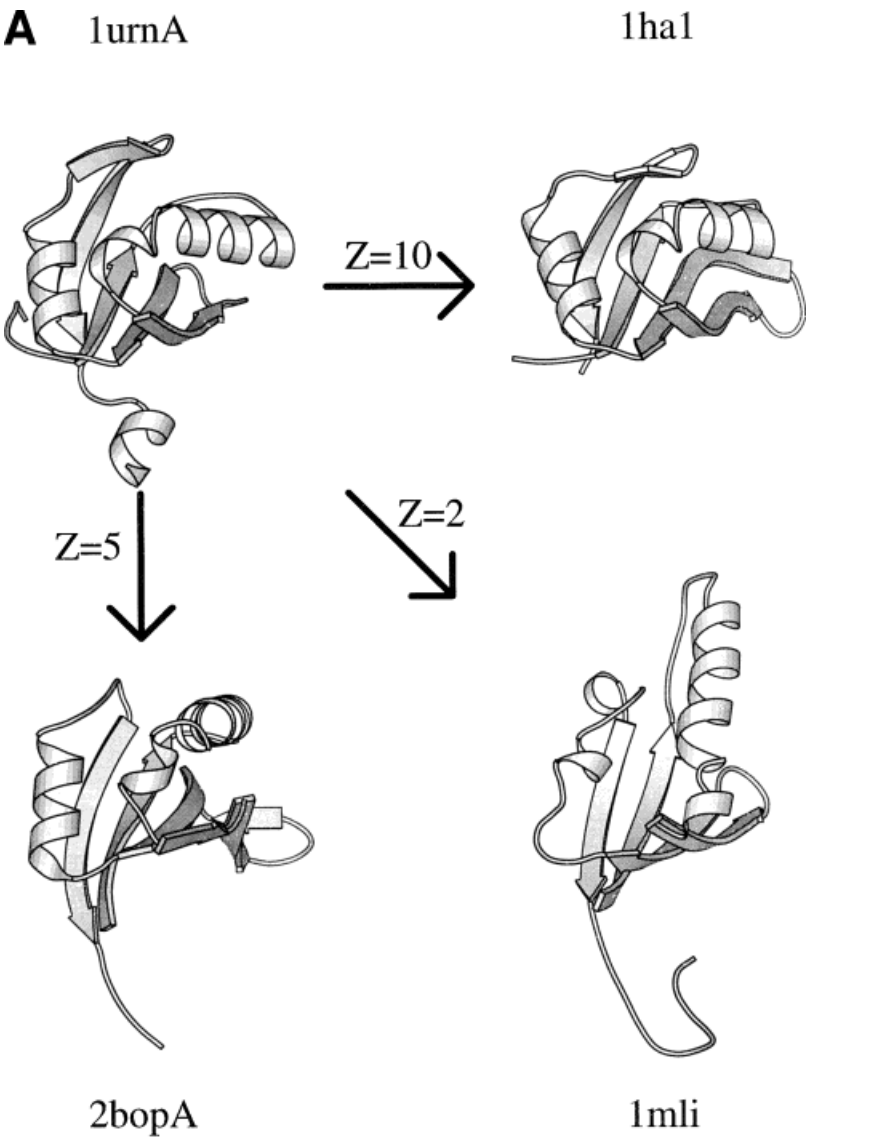
RESULTS AND DISCUSSION

The method was applied to a representative subset of the PDB (November 1997; 1,137 protein chains with less than 25% pairwise sequence identity). The main results are a description of the domain architecture of individual protein structures and a global classification of fold types at the domain level.

Domain Definitions

The algorithmic innovation of the present method is to couple the domain definition for a query protein to the domain definition in its structural neighbors. This coupling through recurrence yields consistent domain definitions between remote homologs and recovers the classic fold types. Many classic folds are among the 20 most frequently recurring fold types that are shown in Figure 2. For example, an alpha/beta domain (resembling the "Rossmann fold") recurs 141 times, an immunoglobulin-like domain recurs 85 times, and TIM barrels recur 63 times in the representative set of structures. Average linkage clustering yields groups of substructures that are homogeneous in size, an indication of overall consistency of the domain selection. Typically, there is a compact structural core common to all members of a fold cluster, and individual members have relatively few secondary structure elements dangling outside the common core (Figure 3). In particular, protein superfamilies made up of functionally and evolutionarily related proteins get consistent domain definitions (e.g., trypsin-like serine proteases).

The assessment of the "correctness" of domain definitions is, generally speaking, an academic ques-



B

```
1urnA --RPNHTIYINNNLNEKI----KKDELKKSLHAIFSRFG---QILDILV-SRS---LKM---
Z=10         *    *              *   *         *  *          *
1ha1  ahLTVKKIFVGGIKEDT--------EEHHLRDYFEQYG---KIEVIEI-MTDrgsGKK---
Z=5           *
2bopA ----sCFALIS-GTANQ-----vKCYRFRVKKNHRHR-----YENCTTtWFT---Vadnga
Z=2                          *
1mli  ---mlFHVKMTVKLpvdmdpakatglkadeKELAQRlgreqTWRHLWR-IAG---------


1urnA ----RGQAFVIFKEV--SSATNALRSMQGFPFYDKPMRIQYAKTDSDIIAKM---------
Z=10      ** *** *       *                      *
1ha1  ----RGFAFVTFDDH--DSVDKIVIQ-kYHTVNGHNCEVRKAL------------------
Z=5         *    *     *     *      *     * *
2bopA ergqQAQILITFGSP--SQRQDFLKHVPLPP----GMNISGF-----tASLDf--------
Z=2           *           *     **     * *
1mli  ----HYANYSVFDVpsvEALHDTLMQLpLFPY----MDIEVD-----gLCRHpssihsddr
```

Fig. 5. Different levels of structural similarity. Dali detects overall topological similarity of structures, but differences of architectural details yield lower similarity scores. **(A)** Four members of the alpha/beta-meander cluster (1urnA, u1a spliceosomal protein; 1ha1, heterogeneous nuclear ribonucleoprotein a1; 2bopA, bovine papillomavirus DNA-binding domain; 1mli, muconolactone isomerase). Each structure is superimposed onto the frame of u1a spliceosomal protein (top left). The Z-score decreases with increasing structural changes, including differences in the length, relative orientation, and number of secondary structure elements. 1urnA compared with itself yields a Z-score of 24. Drawn with Molscript v.2.[25] **(B)** Useful components of the Dali Domain Dictionary are explicit structure-based alignments that allow one to see how the different similarity levels are reflected in sequence. Amino acid identities with respect to 1urnA are marked by asterisks above the aligned sequence (uppercase letters indicate structural equivalence with 1urnA, and lowercase letters indicate structural nonequivalence with 1urnA). The first pair (1urnA-1ha1) are clearly evolutionarily related and have many conserved hydrophobic contacts. The common hydrophobic contacts between 1urnA and the bottom two proteins map to opposite ends of the beta-sheet than between 1urnA and 1ha1. Helices are underlined, and strands are double-underlined.

tion. In most cases, direct experimental data are lacking to verify the structural, functional, or evolutionary autonomy that might be implied by the concept of domain (Fig. 1). Comparison of the objective domain assignments based on recurrence to subjective assignments made by crystallographers (Table I) yields the same number of domains in 83% of the cases (an increase of 7 percentage points over the original PUU algorithm on the same set; data not shown). There is not necessarily a conflict between alternative domain decompositions. For example, the Klenow fragment structure can be equally well described in terms of two functional domains or six recurrent domains (Fig. 4).

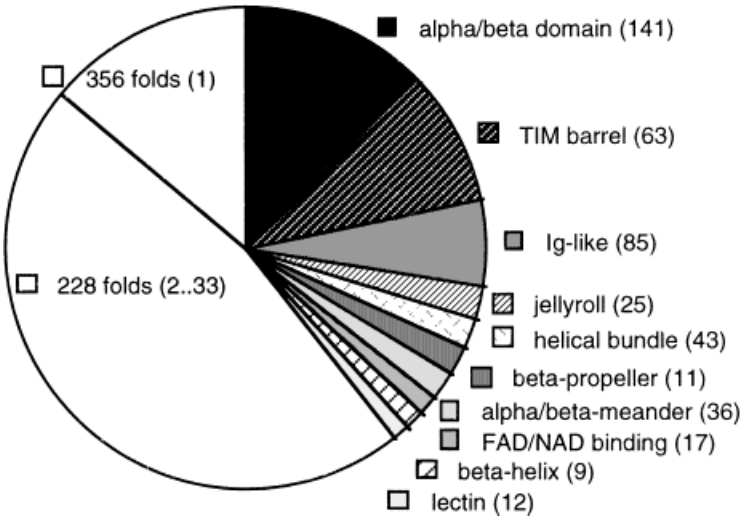


Fig. 6. Population of fold types. The distribution of domains into fold types is weighted by the number of secondary structure elements.[27] Nearly 40% of secondary structure elements in the representative set are covered by only 10 fold types. Each of the remaining fold types all cover less than 1% of the total number of secondary structure elements and are shown as merged slices. Singlet folds are typically smaller than the average structure (14% of secondary structure elements, 20% of domains). Numbers in parentheses are the number of member domains per fold type.

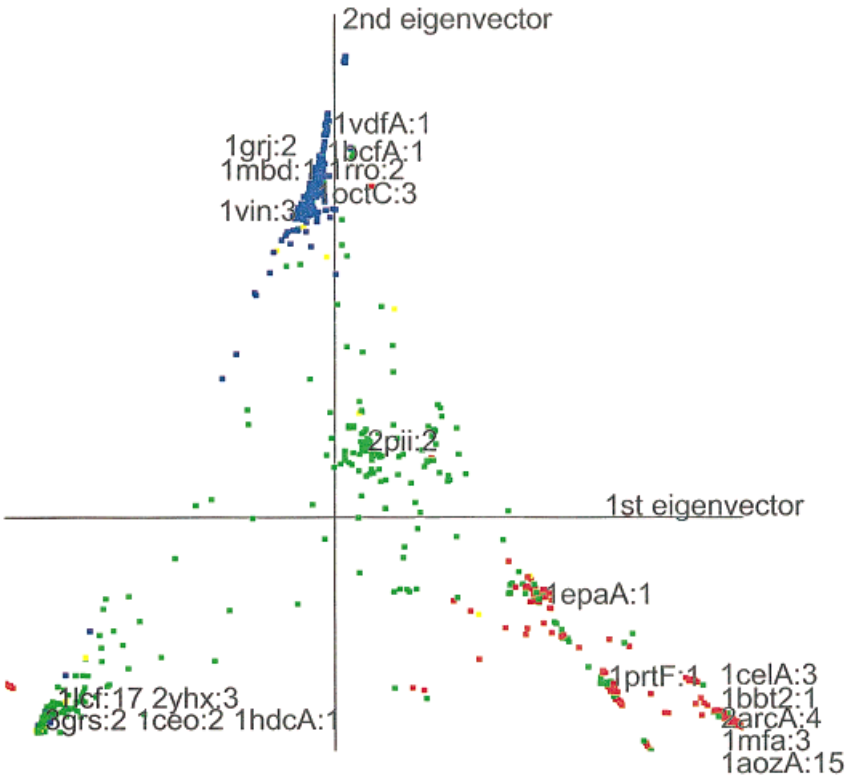


Fig. 7. Map of fold space. Quantification of pairwise structural similarities in an all-on-all comparison allows one to position each domain structure relative to the others in an abstract, high-dimensional fold space. The long-range distribution of different architectures is revealed in a projection down onto a disclosing plane based on multivariate scaling. Mathematically, an eigenvalue problem is solved; the input is the matrix of all-on-all structural similarities, and the plane is defined by the two largest eigenvectors.[18] Each point in the plot is a domain from the representative set, and neighborhoods in the map correspond to clusters of similar folds (extreme outliers have been removed, leaving 1,250 domains that form a connected set where each domain has at least 4 structural neighbors within this set). Labels indicate the position of domains shown in Figure 2 (some labels have been moved to reduce cluttering). Coloring is by secondary structure composition: blue, dominantly alpha-helical (>40 % of residues in helix, <15 % of residues in strand); red, dominantly beta-strands (<15 % helix, >30 % strand); green, mixed helix and strand (>15 % helix, >15 % strand); yellow, otherwise. The first eigenvector primarily discriminates between parallel (left) and antiparallel (right) beta-sheets. The red speck at the left corresponds to beta-helix structures (pectate lyase etc.). The second eigenvector mostly discriminates between all-helical structures and structures with beta-sheets.

Domain Fold Classification

The Dali Domain Dictionary is a compendium of currently known domain folds. The dictionary is organized into structural neighborhoods based on the quantitative all-on-all structure comparison by Dali. The dictionary contains 1,137 representative protein structures decomposed into 1,724 representative domains that are grouped into 594 fold types. Fold types are defined using an empirical cutoff for the radius of a structural neighborhood that groups together domains with similar overall topological arrangement of secondary structure elements (Fig. 5). It is a common observation that the distribution of domains into fold types is very uneven.[18,21,22] Indeed, 40% of the secondary structure elements in the representative set are covered by only 10 fold types (Fig. 6). Although the definition of discrete fold types is useful for counting purposes, the Dali Domain Dictionary also makes explicit the graded similarities that exist between members of the same fold type and that may extend beyond the borders of fold categories. The major architectural trends in the

whole representative set are revealed by multivariate analysis yielding a map of fold space (Fig. 7). The gross distribution of domains in fold space correlates with secondary structure composition, i.e., parallel versus antiparallel beta-sheets versus all-helical architectures. The map further reveals a small number of densely populated regions where the common features are topological motifs at the core of the domains. These dense clusters may represent physical attractor regions in fold space with implications for protein folding.[18]

The World Wide Web is a popular medium for presenting derived databases with added information value. Several groups maintain servers for domain libraries[5–7] and fold classifications based on either hierarchical principles[21,22] or structural neighborhoods.[20,23] The Dali Domain Dictionary is available over the World-Wide Web at http://www.embl-ebi.ac.uk/dali. Browsing the dictionary, the user can search PDB text records, inspect a random series of PDB structures, sort folds by population or novelty, or select groups of similar fold types mapping to the same neighborhood in fold space. From any selected structure, the user can follow links to structural neighbors and can view the structural superimposition with selected neighbors in three dimensions or as explicit multiple alignments of the one-dimensional sequences. Around the known structures, there is typically a whole family of homologous sequences that can be unambiguously aligned to the known structure.[11] In many cases, the Dali Domain Dictionary constructs "multiple alignments of multiple alignments" of several remotely related protein families where the register between the evolutionarily distant families is determined by structural alignment. The unified alignments can reveal conserved sequence patterns that are required for biochemical function and that may lead to evolutionary discoveries.[18]

We look forward to linking this survey of structural domains with an equally comprehensive survey of protein sequence modules, using sensitive methods for the detection of distant sequence similarities. One exciting use of such a joint sequence and structural classification of protein families is the prospect of judiciously choosing novel targets for crystallization and thereby maximizing the information return from experimental structure determination.

REFERENCES

1. Wetlaufer, D.B. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc. Natl. Acad. Sci. USA 70:697–701, 1973.
2. Bork, P. Mobile modules and motifs. Curr. Opin. Struct. Biol. 2:413–421, 1992.
3. Musacchio, A., Noble, M., Pauptit, R., Wierenga, R., Saraste, M. Crystal structure of a Src-homology 3(SH3) domain. Nature 359:851–855, 1992.
4. Waksman, G., Kominos, D., Robertson, S.C., et al. Crystal structure of the phosphotyrosine recognition domain SH2 of v-src complexed with tyrosine-phosphorylated peptides. Nature 358:646–653, 1992.
5. Islam, S.A., Luo, J., Sternberg, M.J. Identification and analysis of domains in proteins. Protein Eng. 8:513–525, 1995.
6. Siddiqui, A.S., Barton, G.J. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. Protein Sci. 4:872–884, 1995.
7. Sowdhamini, R., Rufino, S.D., Blundell, T.L. A database of globular protein structural domains: Clustering of representative family members into similar folds. Fold. Des. 1:209–220, 1996.
8. Swindells, M.B. A procedure for detecting structural domains in proteins. Protein Sci. 4:103–112, 1995.
9. Holm, L., Sander, C. Parser for protein folding units. Proteins 19:256–268, 1994.
10. Richardson, J.S. The anatomy and taxonomy of protein structure. Adv. Protein Chem. 34:167–339, 1981.
11. Sander, C., Schneider, R. Homology-derived secondary structure of proteins and the structural meaning of sequence homology. Proteins 9:56–68, 1991.
12. Abagyan, R.A., Batalov, S. Do aligned sequences share the same fold? J. Mol. Biol. 273:355–368, 1997.
13. Brenner, S.E., Chothia, C., Hubbard, TJ (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 95, 6073–6078.
14. Fischer, D., Eisenberg, D. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. Proc. Natl. Acad. Sci. USA 94:11929–11934, 1997.
15. Hobohm, U., Scharf, M., Schneider, R., Sander, C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci. 1:409–417, 1992.
16. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., et al. The protein data bank: A computer-based archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
17. Holm, L., Sander, C. Protein structure comparison by alignment of distance matrices. J. Mol. Biol. 233:123–138, 1993.
18. Holm, L., Sander, C. Mapping the protein universe. Science 273:595–602, 1996.
19. Holm, L., Sander, C. Alignment of three-dimensional protein structures. Methods Enzymol. 266:653–662, 1996.
20. Holm, L., Sander, C. Touring protein fold space with Dali/FSSP. Nucleic Acids Res. 26:316–319, 1998.
21. 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.
22. Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M. CATH—A hierarchic classification of protein domain structures. Structure 5:1093–1108, 1997.
23. Gibrat, J.-F., Madej, T., Bryant, S.H. Surprising structural similarities. Curr. Opin. Struct. Biol. 6:377–385, 1996.
24. Bennett, M.J., Choe, S., Eisenberg, D. The crystal structure of diphtheria toxin. Nature 357:216–222, 1992.
25. Kraulis, P. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.
26. Beese, L.S., Derbyshire, V., Steitz, T.A. Structure of DNA polymerase I Klenow fragment bound to duplex DNA. Science 260:352–355, 1993.
27. Kabsch, W., Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22:2577–2637, 1983.