Protein Structural Domains: Analysis of the 3Dee Domains Database

Uwe Dengler,[1] Asim S. Siddiqui,[2] and Geoffrey J. Barton[1,2*]
[1]*EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom*
[2]*Laboratory of Molecular Biophysics, University of Oxford, Oxford, United Kingdom*

ABSTRACT **The 3Dee database of domain definitions was developed as a comprehensive collection of domain definitions for all three-dimensional structures in the Protein Data Bank (PDB). The database includes definitions for complex, multiple-segment and multiple-chain domains as well as simple sequential domains, organized in a structural hierarchy. Two different snapshots of the 3Dee database were analyzed at September 1996 and November 1999. For the November 1999 release, 7,995 PDB entries contained 13,767 protein chains and gave rise to 18,896 domains. The domain sequences clustered into 1,715 *domain sequence families,* which were further clustered into a conservative 1,199 *domain structure families* (families with similar folds). The proportion of different domain structure families per domain sequence family increases from 84% for domains 1–100 residues long to 100% for domains greater than 600 residues. This is in keeping with the idea that longer chains will have more alternative folds available to them. Of the representative domains from the domain sequence families, 49% are in the range of 51–150 residues, whereas 64% of the representative chains over 200 residues have more than 1 domain. Of the representative chains, 8.5% are part of multichain domains. The largest multichain domain in the database has 14 chains and 1,400 residues, whereas the largest single-chain domain has 907 residues. The largest number of domains found in a protein is 13. The analysis shows that over the history of the PDB, new domain folds have been discovered at a slower rate than by random selection of all known folds. Between 1992 and 1997, a constant 1 in 11 new domains deposited in the PDB has shown no sequence similarity to a previously known domain sequence family, and only 1 in 15 new domain structures has had a fold that has not been seen previously. A comparison of the September 1996 release of 3Dee to the Structural Classification of Proteins (SCOP) showed that the domain definitions agreed for 80% of the representative protein chains. However, 3Dee provided explicit domain boundaries for more proteins. 3Dee is accessible on the World Wide Web at http://barton.ebi.ac.uk/servers/3Dee.html. Proteins 2001; 42:332–344.** © 2000 Wiley-Liss, Inc.

Key words: protein domain; classification; database; protein structure; protein fold

INTRODUCTION

The protein structural hierarchy runs from primary (the amino acid sequence) through secondary (α-helices and β-strands) to tertiary (single chain, all atoms) and quaternary (multiple chains). A commonly cited intermediate unit of structure in this hierarchy is the domain.[1–3] Although there is no universally agreed definition for a domain, domains are normally compact units of protein structure that can comprise the entire protein chain. If a protein has more than one domain, domains are often thought to be able to exist in isolation from the rest of the protein. Domains may be functional units or modules[4–7] or simply distinct units of protein structure that make up part of the fully functional protein.[2,3] It is thought that such units may be able to fold into a native structure if cleaved from the rest of the protein.

Classification of protein structure at the level of the domain is important in studies of protein structure and function because there are many examples of domains in multidomain proteins that show similarity to single domain proteins. Accordingly, techniques to search for structural similarities[8–17] or to predict protein structure by fold recognition[18–23] must normally consider the protein at the domain level, rather than the complete chain or complex.

The Protein Data Bank (PDB)[24,25] is the repository for protein three-dimensional (3D) structures, but it does not systematically store domain definitions for those structures. As a result, several groups have developed techniques that attempt automatically to determine the domain organization from protein coordinates.[26–37] Subsequently, domain definitions for representative subsets of the PDB have been reported,[30–32,34,35,38] but because the PDB is growing rapidly, these are quickly out of date. To overcome this limitation, comprehensive databases for protein structure classification and protein structural domain definitions have been developed.[39–42] Recently, the protein structure classifications in SCOP,[39] Class, Architecture Topology, and Homologous Superfamily (CATH),[40] and fold classification based on Structure–Structure alignment of Proteins (FSSP)[41] have been compared with the goal of developing reliable template libraries

Grant sponsors: BBSRC; Royal Society; and EMBL.

*Correspondence to: Geoffrey J. Barton, EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD Cambridge, United Kingdom. E-mail: geoff@ebi.ac.uk

Received 7 April 2000; Accepted 19 October 2000

Published online 21 December 2000

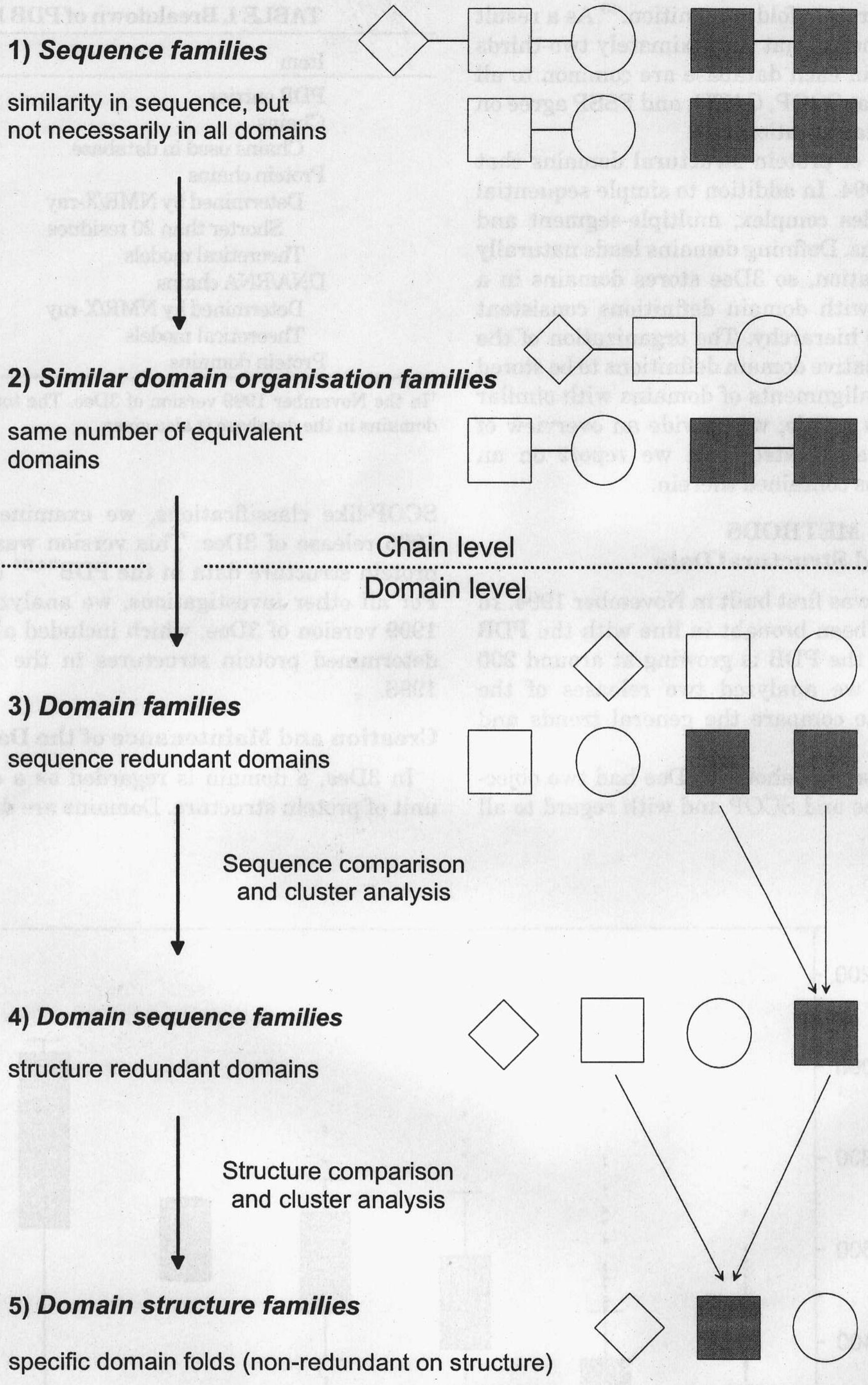
1) *Sequence families*
similarity in sequence, but not necessarily in all domains
2) *Similar domain organisation families*
same number of equivalent domains
Chain level
Domain level
3) *Domain families*
sequence redundant domains
Sequence comparison and cluster analysis
4) *Domain sequence families*
structure redundant domains
Structure comparison and cluster analysis
5) *Domain structure families*
specific domain folds (non-redundant on structure)

Fig. 1. Although different domain types (single-segment, multisegment, and multichain) are stored in the database, for simplicity domains are represented in this figure as differently shaped, filled or unfilled beads on a string. The different shapes represent different domains. Domains with the same shape have the same structure. Domains with the same shape and fill share structural and sequence similarity. The flow chart shows how chains grouped in (1) sequence families are split into (2) similar domain organization families such that chains in each similar domain organization family have the same number of equivalent domains. The step from 2 to 3 leads from the level of chains to the level of domains as similar domain organization families are separated into domains to give (3) domain families. The domains in the domain families are clustered by sequence to produce (4) domain sequence families and then are clustered by structure to form (5) domain structure families.

and benchmarks for protein fold recognition.[43] As a result of this work, it was shown that approximately two-thirds of the protein chains in each database are common to all three database and that SCOP, CATH, and FSSP agree on the majority of their classifications.

3Dee is a database of protein structural domains that was first started in 1994. In addition to simple sequential domains, 3Dee includes complex, multiple-segment and multiple-chain domains. Defining domains leads naturally to structural classification, so 3Dee stores domains in a structural hierarchy with domain definitions consistent within families in the hierarchy. The organization of the database allows alternative domain definitions to be stored as well as structural alignments of domains with similar 3D structures. In this article, we provide an overview of how the database was created, and we report on an analysis of the domains contained therein.

METHODS

Versions of 3Dee and Structural Data

The 3Dee database was first built in November 1994. In the meantime, it has been brought in line with the PDB seven times. Because the PDB is growing at around 200 structures a month, we analyzed two releases of the database to be able to compare the general trends and perspectives observed.

The analysis of these snapshots of 3Dee had two objectives. To compare 3Dee and SCOP and with regard to all SCOP-like classifications, we examined the September 1996 release of 3Dee. This version was in line with the protein structure data in the PDB[24,25] on April 22, 1996. For all other investigations, we analyzed the November 1999 version of 3Dee, which included all X-ray- or NMR-determined protein structures in the PDB on July 21, 1998.

TABLE I. Breakdown of PDB Entries†

Item	Number
PDB entries	7,995
Chains	15,463
Chains used in database	12,458
Protein chains	13,767
Determined by NMR/X-ray	13,423
Shorter than 20 residues	965
Theoretical models	344
DNA/RNA chains	1,696
Determined by NMR/X-ray	1,652
Theoretical models	44
Protein domains	18,896

†In the November 1999 version of 3Dee. The total number of protein domains in the database is also given.

Creation and Maintenance of the Database

In 3Dee, a domain is regarded as a compact, globular unit of protein structure. Domains are defined on a purely

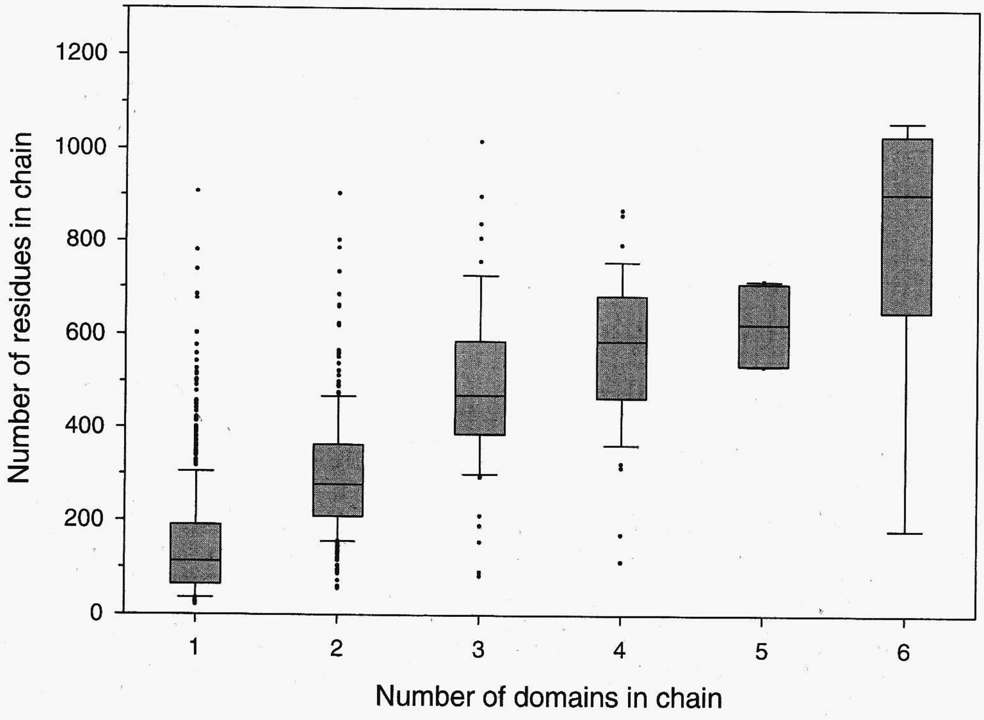

Fig. 2. This box plot diagram[67] shows the number of residues in a chain plotted against the number of domains in the chain for the representative set of chains. Chains that are part of multichain domains have been excluded from the analysis. The horizontal line through the box is the median of the data; the upper and lower ends of the box are at the upper and lower quartiles (75th and 25th percentiles), respectively. The 10th and 90th percentiles are shown as error bars. Very extreme points are plotted by themselves.

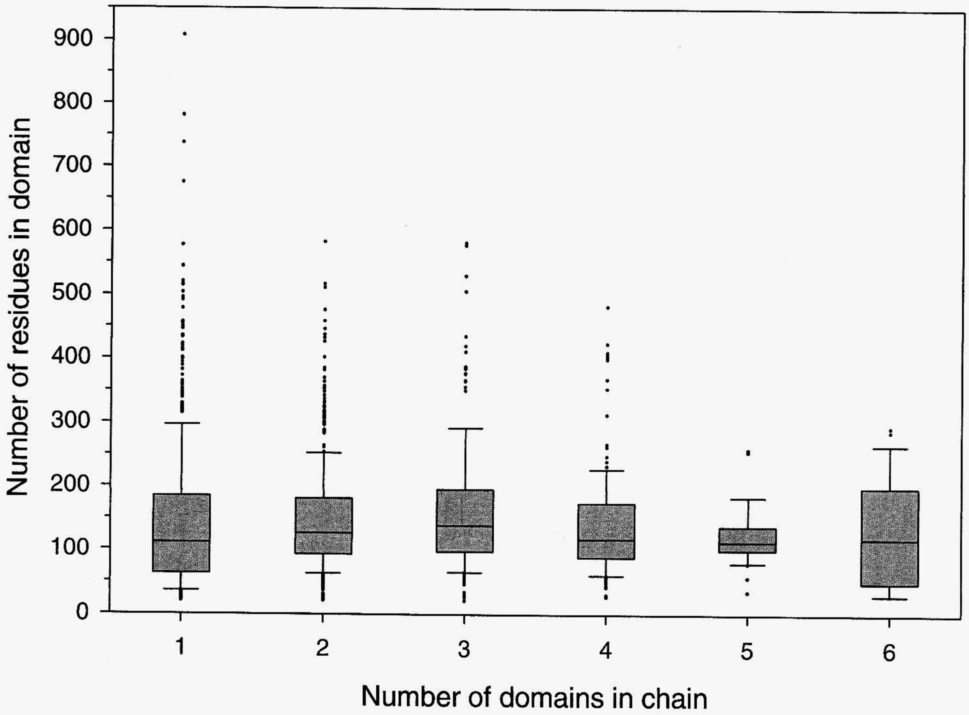

Fig. 3. Box plot showing the number of residues in a domain plotted against the number of domains in the chain for the representative set of chains, excluding chains that are part of multichain domains. See the legend of Figure 2 for an explanation of the box plot notation.

structural basis, so functional features are not taken into account. When the database was first built, all domains were defined by the DOMAK program.[32] DOMAK locates globular domains by maximizing the ratio of the number of internal contacts to the number of external contacts for a given set of coordinates. Compared to domain definitions from the literature, DOMAK gives an accuracy of 70%. After three reliability screens were applied, the DOMAK accuracy rose to 97% for 75% of the proteins in the reference set.[32] However, because the goal was to obtain a comprehensive set of accurate domain definitions, all DOMAK definitions were then checked by eye. The basic rule in assigning domains by eye is not to break β-sheets unless there is an obvious case of twofold or higher symmetry.

In subsequent updates to the database, the domains have been defined automatically by sequence alignment to existing domain definitions or by eye. The update process was made tractable by the development of http-based client-server software that allows domain definitions to be assessed quickly. The update tools perform error checking to prevent invalid or inconsistent definitions from entering the database. Checks are performed to ensure that the start and end residues of a newly entered definition exist and that a new multichain domain definition is consistent with all other multichain domain definitions of the corresponding chains.

A PDB file often contains several distinct chains, each of which may comprise several domains. Two multidomain proteins may share only one domain in common. These complexities require that the database be built in several stages.

Sequence comparison and clustering

To remove redundancy, the chains are first clustered on the basis of local sequence similarity to form *sequence families.* Two proteins may show similarity in only a few of their domains. For example, in Figure 1(1), the two filled chains are correctly clustered together. However, only one chain contains the diamond-shaped domain. So that similar domains may be identified, the sequence families are then divided into those that have *similar domain organization,* that is, the same number of domains, with the same number of segments and a similar number of residues [Fig. 1(2)]. The chains, chosen from the similar domain organization families, without those having identical domain definitions, form a representative set of chains. The formation of similar domain organization families allows the chains to be divided into their constituent domains to give *domain families* [Fig. 1(3)]. When the sequence families are divided into similar domain organization families, domains with significant sequence similarity may be separated. As a result, the domain families do not make up a nonredundant set of sequences. Accordingly, the domain families [Fig. 1(3)] that show sequence similarity were clustered to produce *domain sequence families* [Fig. 1(4)].

The alignment scores used for the sequence clustering came from the program SCANPS,[44] which implements a variant of the Smith–Waterman local alignment algorithm.[45] Alignments are scored by length-dependent statis-

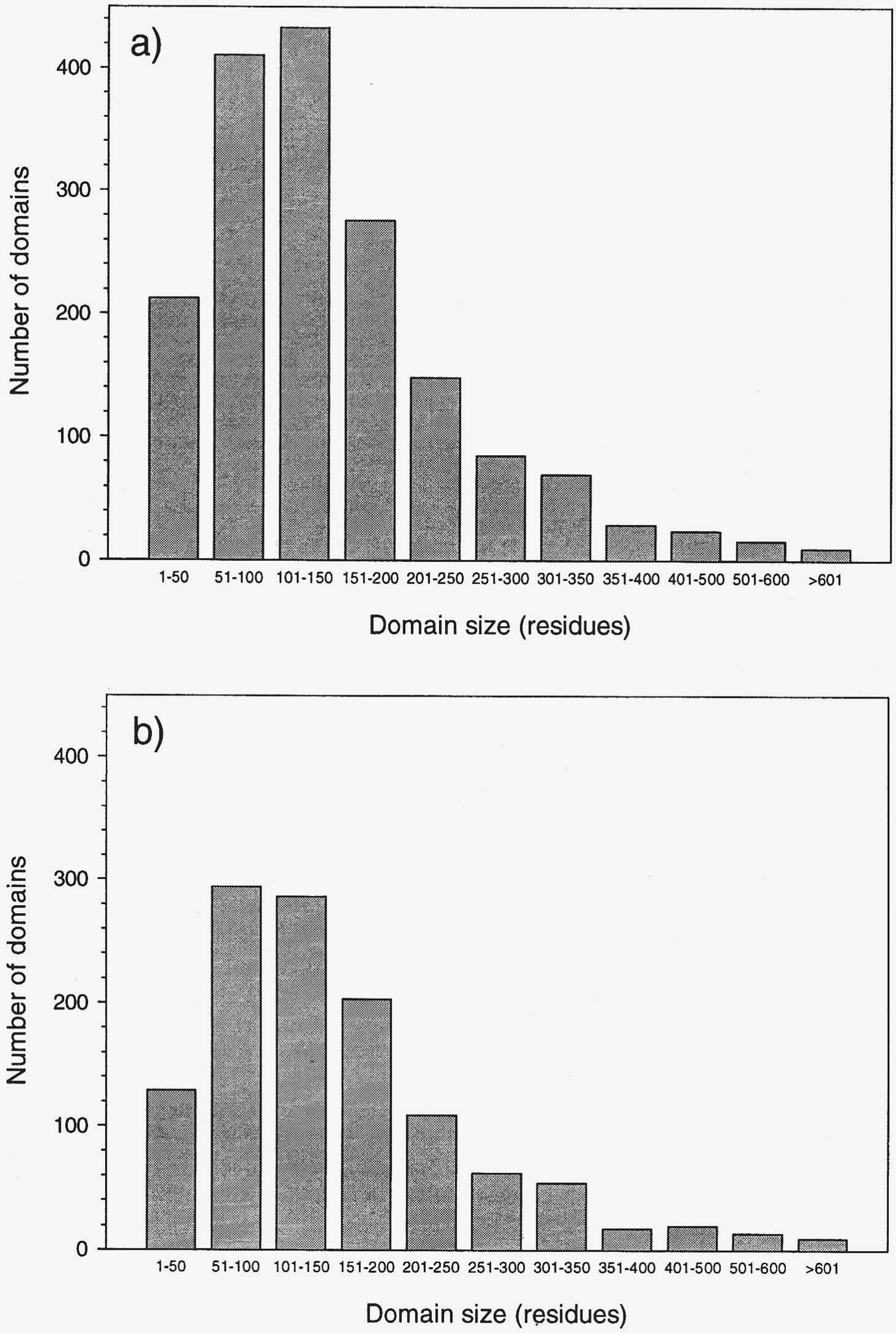

Fig. 4. Distribution of the size of domains for (a) the representatives of the domain sequence families, that is, the nonredundant set of domain sequences; (b) the 5.0 structural families, (c) the 4.0 structural families, and (d) the fold name structural families (September 27, 1996).

tics to estimate the probability of an alignment with a specific score and length being produced by chance. The SCANPS statistics give similar probabilities to BLAST[46] when used to compare the same sequences. The domain sequences were clustered by single linkage clustering to a threshold probability of 10^{-9} and then complete linkage clustering to a threshold of 10^{-7} with the OC program.[47] A size cutoff for comparison of domains, such that the larger domain had to be no more than 1.6 times the size of the smaller domain, was used to prevent fragments of domains from being matched to the complete protein. The effectiveness of this scoring and clustering regime was verified by

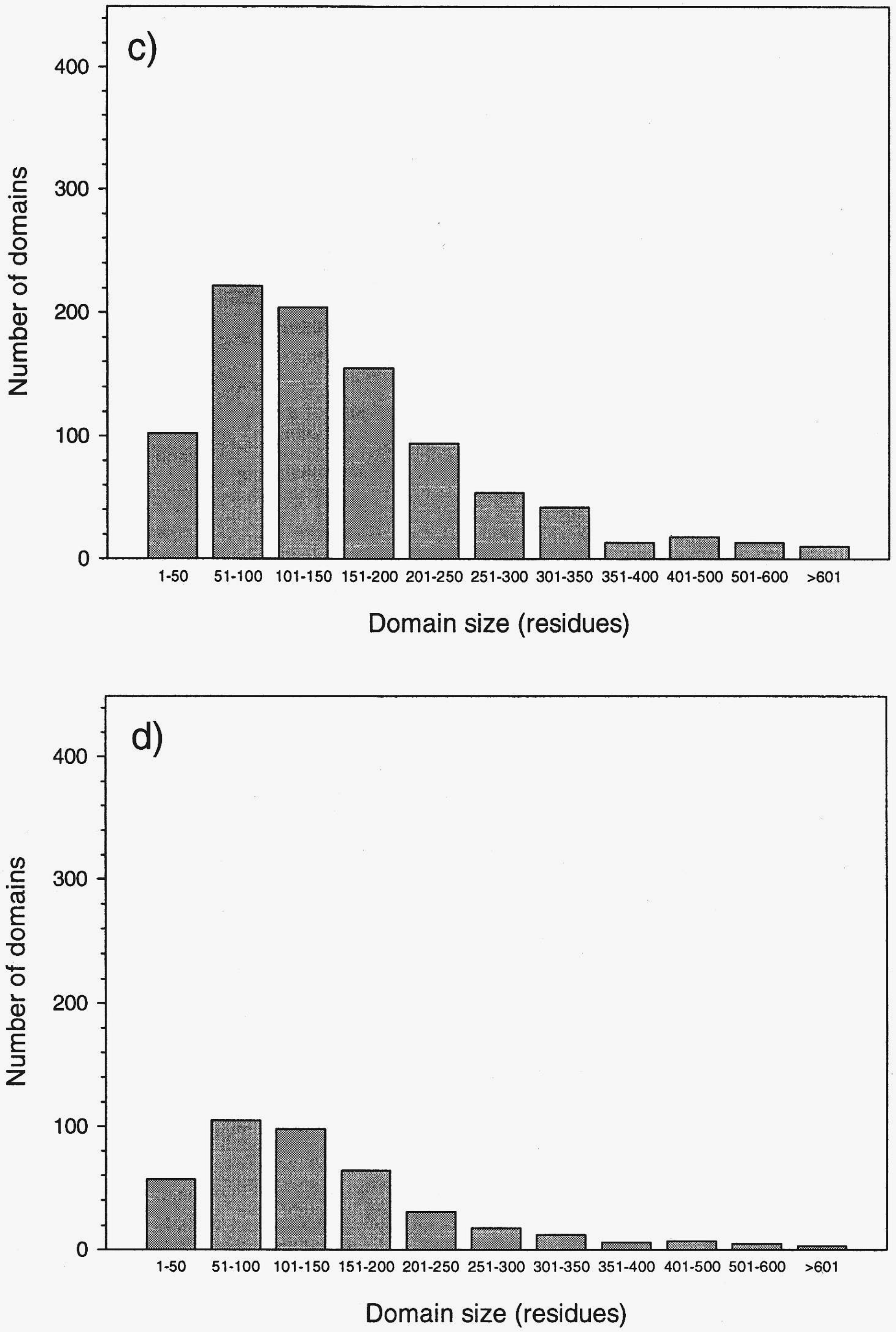

Figure 4. (Continued.)

the plotting of the percentage identity against the alignment length and the inspection of outliers (data not shown).

Structural classification

With the domains grouped into domain sequence families, domains with similar 3D structures but little or no sequence similarity were identified by structural comparison with the widely used STAMP program.[14] STAMP finds multiple structure alignments by an iterative dynamic programming procedure. The formation of *domain structure families* is illustrated in Figure 1(4,5) by the clustering of domains with the same shape (structure) but different fills (no significant sequence similarity). STAMP structural similarity scores[14] were modified to account for the proportion of the domains that are similar and the length of the alignment. Means linkage hierarchical cluster analysis with OC[47] was used to classify the domains.

The resulting clusters are called *5.0* and *4.0 structural families,* depending on the threshold of structural similarity used during the classification.

After the initial database was created, subsequent updates only required comparisons to be made with and between new domains.

Comparison with SCOP

As an independent consistency check of 3Dee's domain definitions and automatic structural classification, it was compared with SCOP,[39] which at the level of similar folds is derived principally by inspection. For this comparison, SCOP version 1.32, released in May 1996, and the version of 3Dee from September 27, 1996, were used.

To compare 3Dee and SCOP, each domain in the representative set of chains was labeled with a SCOP-like fold and class name. A comparison of domain definitions must be performed for chains because domain boundaries in SCOP are defined as positions in protein chains. Accordingly, the comparison was performed on each of the representative chains in 3Dee. Where the domain definitions agreed with SCOP domain definitions, the fold name from the SCOP database was used. If the domain definitions did not agree, a new name for the fold was chosen. With representative members from each domain family having been labeled with a fold and class, fold and class names were transferred to all domains in the 3Dee database. Domains with the same fold name are called *fold name structural families,* which allow a direct comparison between the classification of domains in 3Dee and SCOP.

RESULTS AND DISCUSSION

Breakdown of PDB Entries

The database analyzed comprised 7,995 PDB entries containing 15,463 chains. After the removal of chains that refer to DNA, RNA, theoretical models of proteins, or those containing less than twenty residues, the number of chains was reduced to 12,458, 80.5% of the original set (Table I). Eleven percent of the chains in the PDB represent DNA or RNA; 587 PDB entries (7.3%) contain only DNA or RNA and no protein.

From Protein Chains to Families of Domains

The 12,458 protein chains group into 1,324 sequence families, 1,613 similar domain organization families are generated from the sequence families (Fig. 1), 224 of the 1,324 sequence families are split into two or more similar domain organization families, and 1,613 chains are selected as representatives from the similar domain organization families. Removing chains with identical domain definitions leaves 1,535 of the 1,613 chains. All further classification of structures in the database is carried out only at the level of domains. However, because analysis is often performed on protein chains, not domains, we first discuss statistics from these 1,535 chains, called the *representative set of chains.*

In the representative set of chains, 1,405 chains (91.5%) give rise to single-chain domains, and 923 of these chains comprise a single domain. Of the representative chains, 130 (8.5%) are part of multichain domains formed by two or more chains coming together.

Among the PDB entries containing multichain domains, there are two with nine domains and one with thirteen domains. The nine-domain proteins are phaseolin (PDB ID 2phl)[48] and the lactose operon repressor (1lbi),[49] consisting of six and four chains, respectively. The protein with thirteen domains is a different crystal structure of the lactose operon repressor (1lbg).[49] This structure has been solved in complex with four DNA chains and includes the N-terminal DNA-binding domains with a helix-turn-helix motif and a hinge helix. As the lactose operon repressor is a homotetramer, including one tetramerization domain, this gives an additional four domains. With multichain domains excluded, the largest number of domains seen in a protein is six, as in *Thermus aquaticus* DNA polymerase (Taq polymerase, 1taq)[50]; hexon, an adenovirus type 2 coat protein (1dhx)[51]; carbamoyl phosphate synthetase (1jdb)[52]; and β-galactosidase (1bgm).[53]

Figure 2 shows that there is a steady increase in the chain length with the number of domains, but there is considerable overlap in the sizes of chains with different numbers of domains. This makes it difficult to predict the number of domains in a protein chain given the number of residues. However, 64.3% of the representative chains having over 200 residues consist of more than one domain, 42.9% of the representative chains greater than 350 residues have more than two domains, and 38.5% of the representative chains over 500 residues in length have more than three domains. One of the most prominent outliers in Figure 2 is the 907-residue protein aldehyde oxidoreductase (1alo).[54] In 3Dee, it has been defined as single-domain as it integrates a four-helix bundle and several segregated α- and β-regions in a very compact manner. A 737-residue fragment of the head of myosin (1mmn)[55] is also defined as single-domain in 3Dee. This structure consists of a central α/β-unit surrounded by three small β-sheets and several α-helices. There are two chains with less than 100 residues and two chains with less than 200 residues that consist of three and four domains, respectively. Each of the three domain chains contains three classical zinc fingers, which represent the minimal requirement for a duplex oligonucleotide binding site.[56] Of the two four domain chains, one contains four classical zinc fingers (1ubd),[57] and the other is the lectin wheat germ agglutinin (1wgt),[58] consisting of four forty-residue domains.

Dividing the 1,535 representative chains from the similar domain organization families into separate domains produces 2,558 domain families. Sequence comparison between representatives from each of the 2,558 four domain families reveals 1,715 domain sequence families. A representative from each domain sequence family is used to construct a nonredundant set of domain sequences.

Structural comparisons between these representatives identified similarities not apparent from sequence comparison and so grouped the domain sequence families into domain structure families. This led to a conservative 1,199 5.0 structural families, or 927 4.0 structural families if the more lenient threshold for structural similarity was

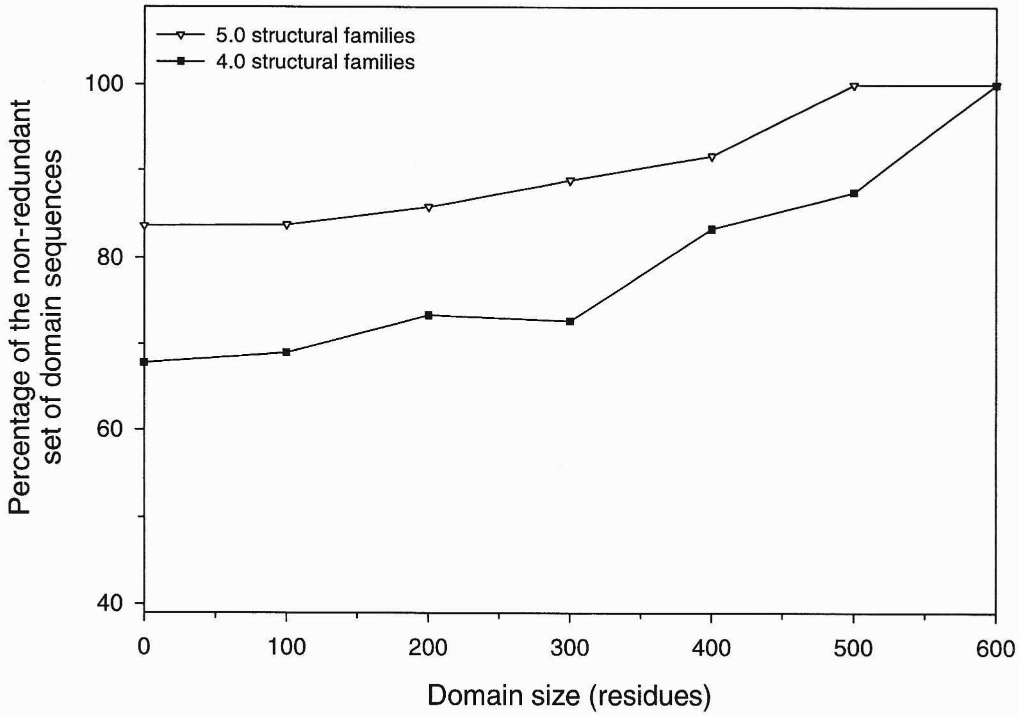

Fig. 5. Number of representative members of the 5.0 and 4.0 structural families in each size interval expressed as a percentage of the number of domains in the nonredundant set of domain sequences in the same size interval. This ratio increases from 84 and 68% to 100%, supporting the hypothesis that larger domains containing more secondary structures have access to a greater number of possible folds.

adopted. The database as a whole stores information about 18,896 domains.

Size of Domains

As shown in Figure 3, most of the domains contain between 50 and 150 residues, regardless of the number of domains in the original chain. There are twenty-six very large single-chain domains greater than 500 residues. Aldehyde oxidoreductase and the fragment of the head of myosin were mentioned in the previous section. Other examples are the tailspike viral adhesion protein, which consists of a β-helix (1tyx),[59] the vanadium-containing chloroperoxidase from *Curvularia inaequalis* (1vnc),[60] and the central domain of lipoxygenase-1 (2sbl).[61] Many of the domains smaller than 50 residues are peptides, fragments from bigger macromolecules, classical zinc fingers, and lectins.

The largest domains in the database are multichain domains, including macromolecular assemblies. An example is the proteasome activator REG-α structure (1avo),[62] which has 1,400 residues and is the largest domain in the database. It is not a globular domain but a barrel-shaped assembly of fourteen helices from fourteen different chains that form one of the subunits of the 11S regulator of the human 20S proteasome.[62]

As shown in Figure 4(a–d), almost half of the representatives from the domain sequence families (49.2%) as well as the 5.0 (48.4%), 4.0 (46.0%), and fold name structural families (49.6%) have 51–150 residues. It is not surprising that the distributions in Figure 4 are similar because all the structural families are subsets of the domain sequence families.

TABLE II. Number of *n* Segment Domains[†]

Number of segments in domain	Occurrences	Percentage
1	1,389	81.0
2	285	16.6
3	29	1.7
4	6	0.3
5–9	3	0.2
10–14	3	0.2

[†]Among the representatives of the domain sequence families.

The number of representative domains from the structural families in each size interval expressed as a percentage of the number of representatives from the domain sequence families in the same size interval highlights the reduction in the number of representative domains dependent on domain size. As shown in Figure 5, this ratio increases for the 5.0 and 4.0 structural families from 84 and 68%, respectively, in the size interval 1–100 residues to 100% for domains greater than 600 residues. Larger domains contain more secondary structures, so a greater number of topologies of secondary structures are possible. Accordingly, larger domains should have access to a greater number of possible folds. If the members of the nonredundant set of domain sequences are randomly selected from the pool of all possible domain sequence families, the ratio in Figure 5 should increase with domain size. The results

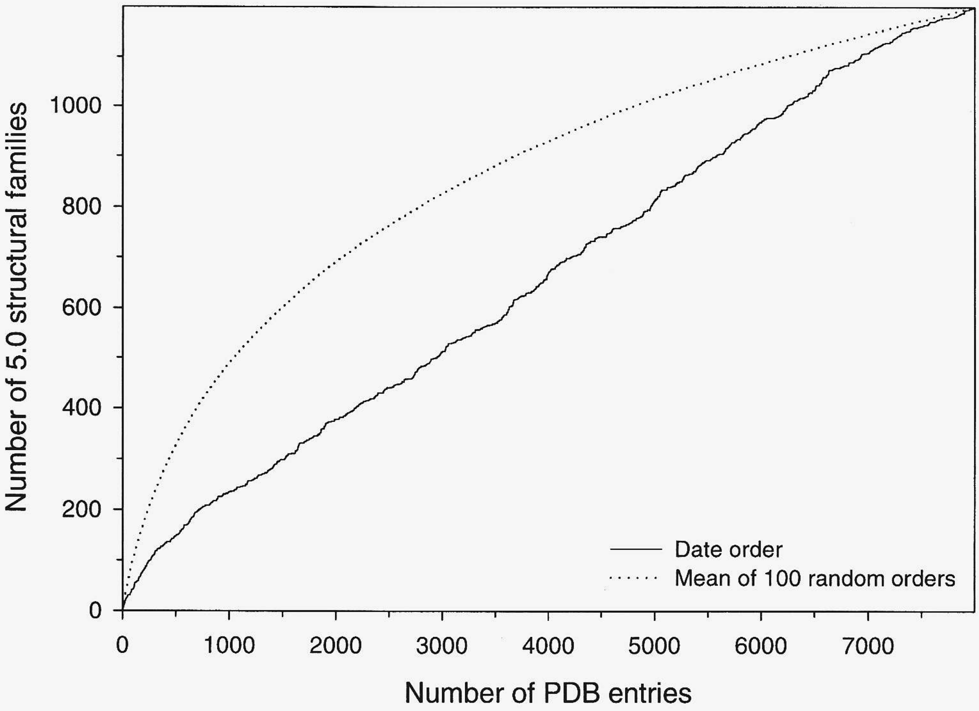

Fig. 6. Mean number of 5.0 structural families for 100 random orders of PDB entries (dotted line). In addition, the actual rate of growth of 5.0 structural families is plotted in date order (solid line). A comparison of the two plots demonstrates that the discovery of new fold families has been nonrandom.

for the 5.0 and 4.0 structural families are in accordance with these considerations.

Number of Segments per Domain

Among the 1,715 representatives of the domain sequence families, 81% of the domains are single-segment. Table II shows that the number of domains drops quickly with an increasing number of segments. Thymidine phosphorylase (1tpt)[63] and phytase (1ihp)[64] are the only examples of proteins with single-chain four-segment domains. There are six domains consisting of more than four segments. All of them are multichain domains. The highest number of segments in a domain is found in the proteasome activator REG-α structure (1avo),[62] which comprises fourteen segments from fourteen distinct chains.

Rate of Growth of the Database

An examination of the 3Dee database shows how domain families have increased since the creation of the PDB. To do this, the PDB entries were ordered by submission date, and the number of domain sequence and structural families present was derived by the counting of the number of new families that were discovered as each entry was submitted.

As of November 1999, the PDB was growing at a rate of approximately 200 new structures per month. The domain sequence families, 5.0, 4.0, and fold name structural families grew exponentially until the end of 1997. Fewer PDB entries were released between January and July 1998 as all curves drop off at the end (data not shown).

Rate of Discovery of New Domain Folds

To probe the rate at which new domain folds have actually been discovered compared to the rate that would be expected by chance, 100 different random orders of PDB entries were generated. For each of the 100 orders, entries were selected in turn, and at each step the number of 5.0, 4.0, and fold name structural families was counted. The mean and the mean error of the mean of the number of families for a given number of PDB entries were then calculated. The result of this calculation for the 5.0 structural families is shown in Figure 6, which also shows the growth in these families given the actual order in which protein structures were deposited with the PDB. The results for the 4.0 and fold name structural families are similar (data not shown).

From these data, it can be concluded that the rate of discovery of new fold families has been slower than would be expected by random selection. This suggests that for many families of domains, the structures of family members have been solved over in close succession; that is, if a certain domain fold is discovered, many of the structures solved subsequently belong to the same structural family, a tendency for structural biologists to focus on certain families of domains once one member is solved or ongoing research in mutants or obvious sequence homologues.

Between January 1992, when the PDB contained 1,200 entries, and March 1997, when there were 6,500 entries, the rate at which structures have been solved for new domain sequence families has been remarkably constant (data not shown). Approximately one new domain se-

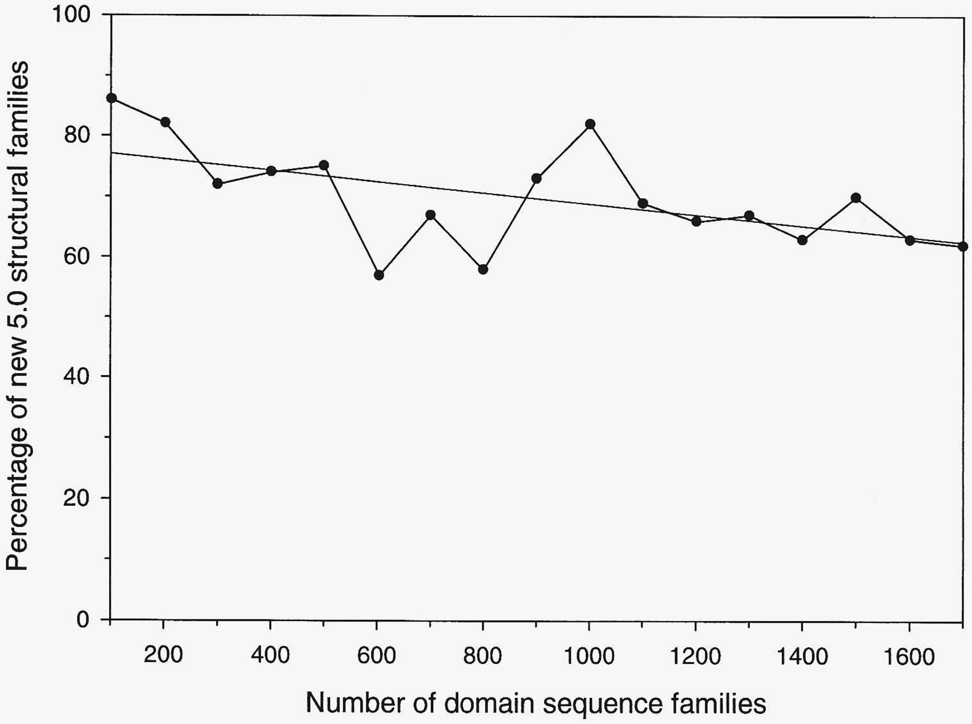

Fig. 7. Percentage of new 5.0 structural families discovered for every 100 new domain sequence families solved. The chance of discovering a domain having a novel fold decreased from over 80% before 1990 to between 60 and 70% in July 1998.

quence family was solved for every five new PDB entries. This corresponds to one in eleven new domain structures.

Every new domain sequence family has the potential to have a new fold as it has no detectable sequence similarity with another structure already present in the database. Figure 7 shows how for the 5.0 structural families, the chance of a domain with the potential to have a novel fold actually having a novel fold has decreased from over 80% before the year 1990 to between 60 and 70% today. This finding suggests that the PDB contains folds corresponding to a significant proportion of all domain sequence families in nature. Taking this into account and with respect to the fold definition of the 5.0 structural families, it is possible to estimate that the total number of folds in nature is less than 2,700.[65]

The last five points in Figure 7 correspond to the 500 most recently solved structures of novel domain sequence family members. The standard deviation and mean of these five data points provide an estimate of the chance that a domain in a new domain sequence family has a novel fold. For 5.0 and 4.0 structural families, the percentage chance is 65 ± 3% and 48 ± 3%, respectively.

Comparison with SCOP

Domain definitions

Of the 853 representative chains in 3Dee on September 27, 1996, 543 (64%) were defined as single-domain in both SCOP and 3Dee. Six (<1%) were defined as single-domain in 3Dee and multiple-domain in SCOP; 116 (14%) were defined as multidomain in 3Dee but only single-domain in SCOP. For 82 of these (71% of 116), SCOP stated that the protein is multidomain, but the domain definitions for these proteins were not given explicitly.

One hundred eighty-one chains (21% of 853) were defined as multidomain in both SCOP and 3Dee. Of these, 139 (77% of 181) had the same definition or showed a minor difference, such as where a protein has short segments at the N or C terminus that pass into other domains. The assignment of these short segments to domains may vary between the databases. For 28 chains (15%), SCOP described the protein as multidomain, but the protein was only partly divided into domains. There were 14 proteins (8%) where the multidomain definitions in SCOP and 3Dee disagreed.

In summary, of the 853 chains, 682 (80.0%) had the same or similar domain definitions in both SCOP and 3Dee; 110 (12.9%) were defined as multidomain in 3Dee, whereas SCOP mentioned that these proteins were multidomain without giving explicit domain definitions. There were 54 examples (6.3%) where the domain definitions disagreed. The remaining 7 examples (0.8%) corresponded to whole or parts of proteins not described in the SCOP database.

The domain definitions of SCOP and 3Dee agreed for the majority of protein chains. 3Dee gave explicit domain boundaries for more proteins. Of the small number of chains with major differences in the domain definitions, most were due to differences between the authors' concepts of domains. For

TABLE III. SCOP-like Classification of Domains (September 27, 1996)

Class	*A*[a]	*B*[b]	*C*[c]
α and β (α/β)	2,016 (22.27%)	262 (26.15%)	88
All β	2,793 (30.86%)	208 (20.76%)	58
All α	1,355 (14.97%)	193 (19.26%)	92
α and β (α + β)	2,091 (23.10%)	191 (19.06%)	90
Small proteins	576 (6.36%)	94 (9.38%)	40
Peptides	101 (1.12%)	32 (3.19%)	26
Membrane and cell surface proteins and peptides	62 (0.69%)	11 (1.10%)	6
Multidomain (α and β)	54 (0.60%)	8 (0.80%)	6
Nonprotein	2 (0.02%)	2 (0.20%)	2
Designed proteins	1 (0.01%)	1 (0.10%)	1
Total	9,051 (100.00%)	1,002 (100.00%)	409

[a]Number and percentage of domains of this class in the entire database.
[b]Number and percentage of domains of this class in the nonredundant set of domain sequences.
[c]Number of folds of this class.

example, SCOP does not split serine proteases because both domains are required for function. There were a small number of errors in both databases (<1%).

Structural classification

Assigning folds and classes to all domains in 3Dee gave ten structural classes and 409 folds. Table III shows that folds in the all β class were the most common in the 3Dee database as a whole. This was primarily because of the immunoglobulin-like β-sandwich domains, which made up 8.1% of the total number of domains in the database. In addition, there were large numbers of trypsin-like serine proteases, Domains I and II (together, both domains made up 4.9% of the database). α/β domains comprised the largest class in the nonredundant set of domain sequences. Domains containing both helices and strands (α/β and α + β domains) contributed to 45.2% of the representative set. The four main structural classes (α, β, α/β and α + β) made up 85.2% of the nonredundant set of domain sequences.

There are fifty-eight differences between the 5.0 structural families and the SCOP derived fold name structural families. All but one of these differences have the same secondary structures in the same topology. The only exception is the spectrin repeat unit, whose three helices match with three of the four helices in a four-helix bundle.

The fold name structural families in some cases classify the data to a finer level, whereas in others they bring together domains of the same fold that the structural scoring scheme does not easily find. The SCOP database is built on functional and evolutionary relationships. Although these relationships are related to structural similarity, they are not exclusively defined by structural measures. Hence, the SCOP database does not group together some proteins that show obvious structural similarity, for example, the globins and colicin A. In the SCOP classification, there is no such thing as degree of structural similarity. Domains either have the same fold or do not. The advantage of this is that all domains with the same fold are placed in the same group. The disadvantage is that the degree of similarity between many folds, for example, many of the α/β domains, is not readily apparent. Hence, the structural classification in 3Dee is a good complement to the SCOP database.

SUMMARY AND FUTURE DEVELOPMENTS

In this study, two different snapshots of the 3Dee database, a comprehensive collection of domain definitions for 3D structures in the PDB, were analyzed.

In contrast to other comprehensive protein structure classification and protein structural domain databases that mainly describe simple sequential domains, 3Dee includes definitions for complex, multiple-segment and multiple-chain domains organized in a structural hierarchy.

The domains in 3Dee are defined for the coordinates stored in a PDB file. For structures solved by X-ray crystallography, these coordinates consist of only the contents of the asymmetric unit. As a result, some domains that span multiple chains will not be identified. Recently, PQS, a database of probable quaternary molecules, was developed by Henrick and Thornton.[66] A future development of 3Dee may make use of this resource to provide a more complete description of domains in proteins.

The PDB is growing by about 200 structures a month, so any analysis such as this will be out of date by the time it is completed and published. Despite this, the trends and perspectives seen in the two snapshots of 3Dee, separated by 2 years and a 77.4% growth in the number of PDB entries, are very similar. This provides confidence in the generality of the results presented in this article. The statistics presented will be updated automatically in future releases of 3Dee and will be available from the 3Dee web site (http://barton.ebi.ac.uk/servers/3Dee.html).

ACKNOWLEDGMENTS

The authors thank Professor L.N. Johnson for her support and encouragement; Dr. R.B. Russell, Dr. R.R. Copley, and Dr. S.M.J. Searle for helpful discussions; and Dr. J. Cuff for contributing some of the domain definitions. This work was supported by a Biotechnology and Biological Sciences Research Council (BBSRC) studentship to A.S. Siddiqui. G.J. Barton thanks the Royal Society and, the European Molecular Biology Laboratory (EMBL) for their support.

REFERENCES

1. Wetlaufer DB. Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci U S A 1973;70:697–701.
2. Rossmann MG, Liljas A. Recognition of structural domains in globular proteins. J Mol Biol 1974;85:177–181.
3. Richardson JS. The anatomy and taxonomy of protein structure. Adv Protein Chem 1981;34:246.
4. Patthy L. Introns and exons. Curr Opin Struct Biol 1994;4:383–392.
5. Baron M, Norman DG, Campbell ID. Protein modules. Trends Biochem Sci 1991;16:13–17.
6. Campbell ID, Baron M. The structure and function of protein modules. Philos Trans R Soc Lond [Biol] 1991;332:165–170.
7. Hegyi H, Bork P. On the classification and evolution of protein modules. J Protein Chem 1997;16:545–551.
8. Argos P, Rossmann M. Exploring. J Mol Biol 1976;105:75–95.
9. Taylor W, Orengo C. Protein. J Mol Biol 1989;208:1–21.
10. Sali A, Blundell TL. Definition of general topological equivalence in protein structures: A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. J Mol Biol 1990;212:403–428.
11. Subbarao N, Haneef I. Defining topological equivalences in macromolecules. Protein Eng 1991;4:877–884.
12. Vriend G, Sander C. Detection of common three-dimensional substructures in proteins. Proteins 1991;11:52–58.
13. Alexandrov NN, Katsutoshi T, Go N. Common spatial arrangement of backbone fragments in homologous and non-homologous proteins. J Mol Biol 1992;225:5–9.
14. Russell RB, Barton GJ. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. Proteins 1992;14:309–323.
15. Grindley HM, Artymiuk PJ, Rice DW, Willett P. Identification of tertiary structure resemblance in proteins using a maximal common subgraph isomorphism algorithm. J Mol Biol 1993;229:707–721.
16. Holm L, Sander C. Protein structure comparison by alignment of distance matrices. J Mol Biol 1993;233:123–138.
17. Holm L, Sander C. Searching protein structure databases has come of age. Proteins 1994;19:165–173.
18. Russell RB, Copley RR, Barton GJ. Protein fold recognition by mapping predicted secondary structures. J Mol Biol 1996;259:349–365.
19. Jones DT, Thornton JM. Potential energy functions for threading. Curr Opin Struct Biol 1996;6:210–216.
20. Fischer D, Rice D, Bowie JU, Eisenberg D. Assigning amino acid sequences to 3-dimensional protein folds. FASEB J 1996;10:126–136.
21. Sippl MJ, Flöckner H. Threading thrills and threats. Structure 1996;4:15–19.
22. Smith TF, Lo Conte L, Bienkowska J, Gaitatzes C, Rogers RGJ, Lathrop R. Current limitations to protein threading approaches. J Comp Biol 1997;4:217–225.
23. Torda AE. Perspectives in protein-fold recognition. Curr Opin Struct Biol 1997;7:200–205.
24. Bernstein FC, Koetzle TF, Williams GJ, Meyer EFJ, Brice MD, Rodgers JR, 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.
25. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucl Acid Res 2000;28:235–242.
26. Crippen GM. The tree structural organisation of proteins. J Mol Biol 1978;126:315–332.
27. Rose GD. Hierarchic organisation of domains in globular proteins. J Mol Biol 1979;134:447–470.
28. Wodak SJ, Janin J. Location of structural domains in proteins. Biochemistry 1981;20:6544–6552.
29. Zehfus MH, Rose GD. Compact units in proteins. Biochemistry 1986;25:5759–5765.
30. Zehfus MH. Binary discontinuous compact protein domains. Protein Eng 1994;7:335–340.
31. Holm L, Sander C. Parser for protein folding units. Proteins 1994;19:256–268.
32. Siddiqui A, Barton GJ. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. Protein Sci 1995;4:872–884.
33. Sowdhamini R, Blundell TL. An automatic method involving cluster analysis of secondary structures for the identification of domains in proteins. Protein Sci 1995;4:506–520.
34. Islam SA, Luo J, Sternberg MJE. Identification and analysis of domains in proteins. Protein Eng 1995;8:513–525.
35. Swindells M. A procedure for the automatic determination of hydrophobic cores in protein structures. Protein Sci 1995;4:103–112.
36. Jones S, Stewart M, Michie A, Swindells MB, Orengo C, Thornton JM. Domain assignment for protein structures using a consensus approach: Characterization and analysis. Protein Sci 1998;7:233–242.
37. Wernisch L, Hunting M, Wodak SJ. Identification of structural domains in proteins by a graph heuristic. Proteins 1999;35:338–352.
38. Sowdhamini R, Rufino SD, Blundell TL. A database of globular protein structural domains: Clustering of representative family members into similar folds. Fold Des 1996;1:209–220.
39. Murzin A, Brenner SE, Hubbard T, Chothia C. SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540.
40. Orengo CA, Michie AD, Jones S, T. JD, Swindells MB, Thornton JM. CATH—A hierarchic classification of protein domain structures. Structure 1997;5:1093–1108.
41. Holm L, Sander C. The FSSP database: Fold classification based on structure–structure alignment of proteins. Nucl Acid Res 1996;24:206–209.
42. Holm L, Sander C. Dictionary of recurrent domains in protein structures. Proteins 1998;33:88–96.
43. Hadley C, Jones D. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. Structure 1999;7:1099–1112.
44. Barton GJ. An efficient algorithm to locate all locally optimal alignments between two sequences allowing for gaps. Comput Appl Biosci 1993;9:729–734.
45. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol 1981;147:195–197.
46. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
47. Barton GJ. OC—A cluster analysis program: Usage notes. http://barton.ebi.ac.uk/manuals/oc/AAREADME. 1997.
48. Lawrence MC, Izard T, Beuchat M, Blagrove RJ, Colman PM. Structure of phaseolin at 2.2 Å resolution. Implications for a common vicilin/legumin structure and the genetic engineering of seed storage proteins. J Mol Biol 1994;238:748–767.
49. Lewis M, Chang G, Horton NC, Kercher MA, Pace HC, Schumacher MA, Brennan RG, Lu P. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science 1996;271:1247–1254.
50. Youngsoo K, Eom SH, Wang J, Lee DS, Suh SW, Steitz TA. Crystal structure of *Thermus aquaticus* DNA polymerase. Nature 1995;376:612–616.
51. Athappilly FK, Murali R, Rux JJ, Cai Z, Burnett RM. The refined crystal structure of hexon, the major coat protein of adenovirus type 2, at 2.9 Å resolution. J Mol Biol 1994;242:430–455.
52. Thoden JB, Raushel FM, Benning MM, Rayment I, Holden HM. The structure of carbamoyl phosphate synthetase determined to 2.1 Å resolution. Acta Crystallogr D Biol Crystallogr 1999;55:8–24.
53. Jacobson RH, Zhang XJ, DuBose RF, Matthews BW. Three-dimensional structure of β-galactosidase from *E. coli.* Nature 1994;369:761–766.
54. Romao MJ, Archer M, Moura I, Moura JJ, LeGall J, Engh R, Schneider M, Hof P, Huber R. Crystal structure of the xanthine oxidase-related aldehyde oxido-reductase from *D. gigas.* Science 1995;270:117–1176.
55. Gulick AM, Bauer CB, Thoden JB, Rayment I. X-ray structures of the MgADP, MgATPgammaS, and MgAMPPNP complexes of the *Dictyostelium discoideum myosin* motor domain. Biochemistry 1997;36:11619–11628.
56. Foster MP, Wuttke DS, Radhakrishnan I, Case DA, Gottesfeld JM, Wright PE. Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA. Nat Struct Biol 1997;4:605–608.
57. Houbaviy HB, Usheva A, Shenk T, Burley SK. Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. Proc Natl Acad Sci U S A 1996;93:13577–13582.
58. Harata K, Nagahora H, Jigami Y. X-ray structure of wheat germ

agglutinin isolectin 3. Acta Crystallogr D Biol Crystallogr 1995;51: 1013–1019.

59. Steinbacher S, Baxa U, Miller S, Weintraub A, Seckler R, Huber R. Crystal structure of phage P22 tailspike protein complexed with *Salmonella sp.* O-antigen receptors. Proc Natl Acad Sci U S A 1996;93:10584–10588.

60. Messerschmidt A, Wever R. X-ray structure of a vanadium-containing enzyme: Chloroperoxidase from the fungus *Curvularia inaequalis.* Proc Natl Acad Sci U S A 1996;93:392–396.

61. Boyington JC, Gaffney BJ, Amzel LM. The three-dimensional structure of an arachidonic acid 15-lipoxygenase. Science 1993;260: 1482–1486.

62. Knowlton JR, Johnston SC, Whitby FG, Realini C, Zhang Z, Rechsteiner M, Hill CP. Structure of the proteasome activator REGalpha (PA28alpha). Nature 1997;390:639–643.

63. Walter MR, Cook WJ, Cole LB, Short SA, W. KG, Krenitsky TA, Ealick SE. Three-dimensional structure of thymidine phosphorylase from *Escherichia coli* at 2.8 Å resolution. J Biol Chem 1990;265:1416–1422.

64. Kostrewa D, Gruninger-Leitch F, D'Arcy A, Broger C, Mitchell D, van Loon AP. Crystal structure of phytase from *Aspergillus ficuum* at 2.5 Å resolution. Nat Struct Biol 1997;4:185–190.

65. Siddiqui AS. Computer analysis and classification of protein structural domains [D. Phil thesis]. Oxford, England: University of Oxford; 1997. 422 p.

66. Henrick K, Thornton JM. PQS: A protein quaternary structure file server. Trends Biochem Sci 1998;9:358–361.

67. Velleman PF, Hoaglin DC. Applications, basics and computing exploratory data analysis. Boston: Duxbury; 1981. 354 p.