Population statistics of protein structures: lessons from structural classifications

Steven E Brenner*§, Cyrus Chothia† and Tim JP Hubbard‡

Structural classifications aid the interpretation of proteins by describing degrees of structural and evolutionary relatedness. They have also recently revealed strikingly skewed distributions at all levels; for example, a small number of folds are far more common than others, and just a few superfamilies are known to have diverged widely. The classifications also provide an indication of the total number of superfamilies in nature.

Addresses
*Structural Biology Centre, National Institute for Bioscience and Human-Technology, Higashi 1-1, Tsukuba, Ibaraki 305, Japan and MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK; e-mail: brenner@hyper.stanford.edu
†MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK; e-mail: chc1@mrc-lmb.cam.ac.uk
‡Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SA, UK; e-mail: th@sanger.ac.uk
§Address from August 1997: Department of Structural Biology, Stanford University, Stanford, CA 94305-5400, USA

Current Opinion in Structural Biology 1997, 7:369–376

http://biomednet.com/elecref/0959440X00700369

Abbreviations
FOD frequently occurring domain
PDB Protein Data Bank

Introduction

The past few years have witnessed an astonishing growth in the number of proteins whose structures have been solved by crystallography and NMR. These data, archived by the Protein Data Bank (PDB) [1], grew so dramatically that they risked becoming an embarrassment of riches: so many structures were available that perhaps only one person in the world knew all of them. It was nearly impossible to locate individual structures of interest, much less to see general trends. A bevy of publicly accessible databases [2•,3] have come to the rescue, organizing the proteins and making it possible to understand the relationships between structures. Our review describes some of the interesting characteristics of the distributions of different structures.

Structural and evolutionary relationships

In order to understand the relationships between proteins at different levels, it is crucial to first describe different degrees of similarity. The descriptions used in this review are principally those from the scop database [4••]; however, related levels apply to CATH [5,6••], Entrez [7••], FSSP [8••], DDBASE [9••] and other databases that have been developed. Figure 1 samples a portion of the classification hierarchy and indicates how levels in the different databases relate. See Brenner *et al.* [10•] for a detailed description of the heirarchical levels in the scop classification.

Whereas most proteins in the PDB have just a single domain, many larger proteins have two or more. Most databases split such proteins into their constituent domains, as these are typically the fundamental units of both structure and evolution. Individual domains within a large protein may therefore be considered independently.

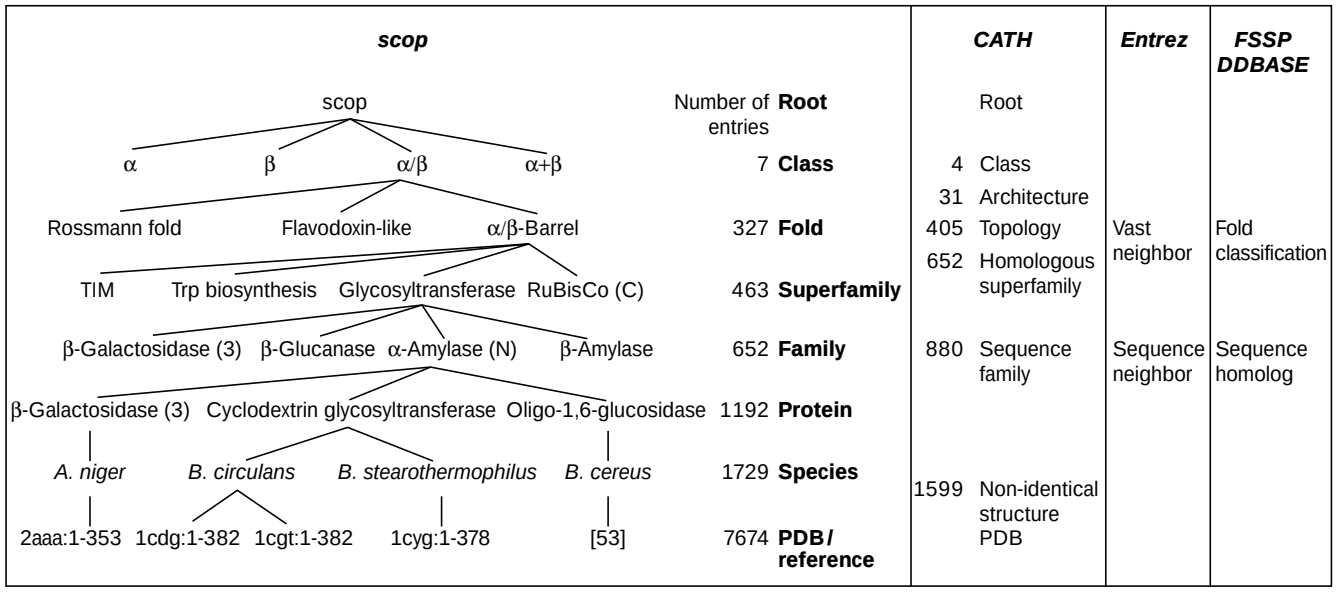
The principal structural similarity between two proteins is the 'fold', which indicates that two proteins (or domains of proteins) share the same core secondary-structure elements in the same arrangement. Although proteins with the same fold must have the same core, the peripheral structural elements may differ greatly, including elaboration and insertion of other domains.

The structures of proteins may be organized by more general degrees of similarity than fold. The CATH database, for example, indicates 'architecture', which reflects the gross arrangement of secondary-structure elements independent of their order. More general yet is the structural class [11], which conveniently, but very crudely, describes the secondary-structure composition of a protein and its most general organizational characteristics.

Proteins can also be classified according to sequence similarities. These sorts of relationships, called 'families', are particularly important because significant sequence similarity implies homology (i.e. an evolutionary relationship), and thus similar structures.

The potentially most interesting and valuable relationships for the structural biologist fall between the structural level of folds and the homology from sequences described by families. Since domains of related globular proteins share the same fold, all the protein domains within a single family must have the same fold. However, the reverse is not true: many proteins have the same fold but very different sequences (which would put them into different families). By a detailed analysis of such proteins, including a consideration of their function and detailed structural similarity, it is sometimes possible to be confident that they are evolutionarily related — even though they do not have any sequence similarity. Actin and the ATPase fragment of 70 kDa heat shock cognate [12] provide a good example of two proteins that share a common fold and whose detailed analysis reveals homology. In the scop database, inferred evolutionary relationships between

Figure 1



A sample region of the scop (version 1.32) hierarchy indicating the number of entries at each level, and the roughly equivalent levels of other databases. Classifications at the fold level and above are purely structural, whereas those at the superfamily level and below are based on evolution. The homologous superfamily (H) level of CATH 1.0 falls in between these two in that it attempts to use solely structural criteria to identify homology. Hyperfamilies (not shown) would be above superfamilies in the hierarchy. FODs and true superfamilies are not hierarchical levels but are special entries at the fold and superfamily levels, respectively. In the scop, CATH, Entrez, and DDBASE classifications, all proteins are divided into domains. In the figure, domains are either indicated in parenthesis or by residue numbers. The scop and CATH databases have hierarchical levels with defined interpretations. Entrez provides two types of links, but these are not necessarily hierarchical nor transitive. The FSSP and DDBASE databases are hierarchical and provide a variety of structural similarity thresholds that do not directly correspond to levels in scop or CATH. Both FSSP and DDBASE contain only a representative set of proteins with dissimilar sequences, each taken from a family of similar sequences. PDB entry codes, 2aaa, 1cdg, 1cgt and 1cyg, and literature reference [53] are classifed in this sample of scop. Adapted with permission from [54].

families that do not exhibit sequence similarity are called 'superfamilies'.

If proteins share a common fold and show some evidence of homology of a degree less discriminating than superfamilies, then they can be grouped into 'hyperfamilies' [13].

Space constraints preclude a discussion of the many other useful classifications of protein structures, especially those based on motifs from large or small elements of structure, or those related to function [14,15•,16–29]. For more details about any of the classifications described in this review, readers should consult the databases and explore the structural information they embody.

Current knowledge and distribution

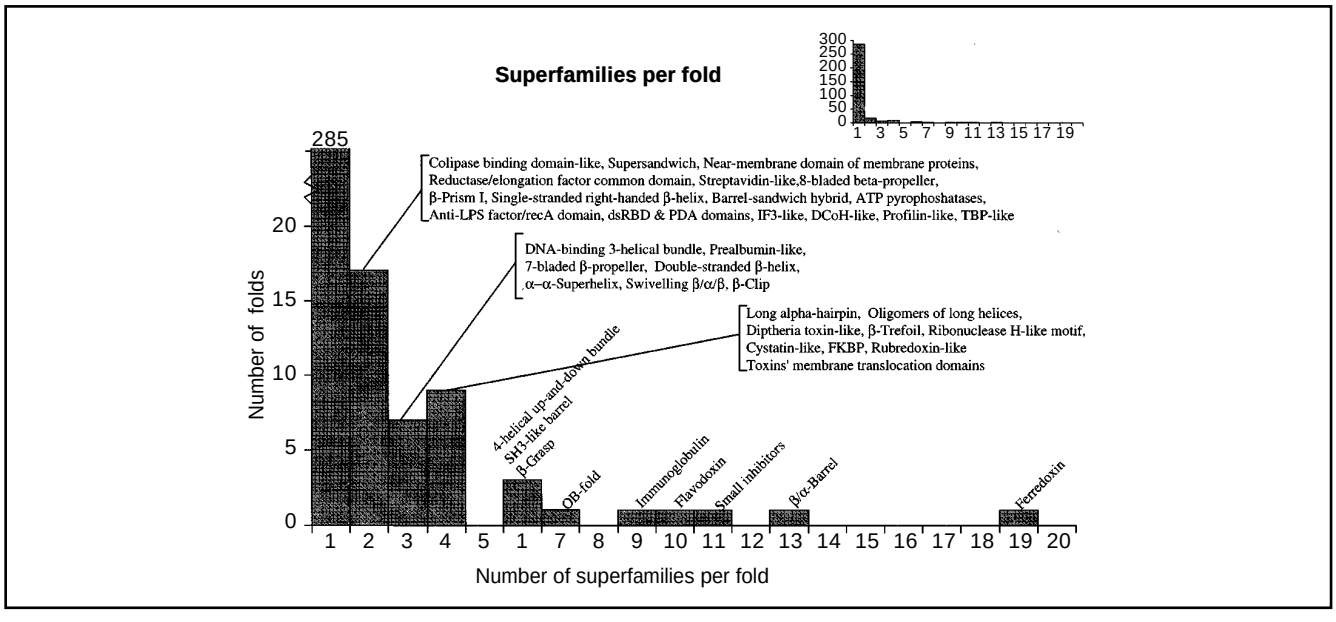
One of the simplest results to come from the construction of databases of relationships between proteins is the number of known protein folds. According to the latest release of the scop database (version 1.32), we now know of 327 protein folds. The numbers provided by other databases are similar (see Fig. 1), but they vary because of their different update dates and their occasional variations in definitions.

Immediately evident from these figures is the high degree of redundancy in the PDB; there are a mean of 4.4 entries for each domain (of a given protein from a single species). The number of structures of very similar proteins has been remarked on many times before [30,31], but examination reveals that their distribution is extremely skewed. A small number of proteins have a huge number of PDB entries (T4 lysozyme has 212), but the majority have one.

Frequently occurring domains

One of the earliest discoveries from scop was that the skew in distributions is a general feature seen at nearly all levels of the hierarchy. We have many representatives of or know much about a small number of items, but we have little data on the majority. Perhaps most interestingly, a small number of folds account for the lion's share of superfamilies. While the numbers in Figure 1 might suggest roughly half of the folds would have one superfamily and half would have two, the actual distribution (Fig. 2) is quite different; it is more extreme than Poisson. Orengo *et al.* [13] recognized this discrepancy and quantified it, using the term superfold to describe over-represented structures that have more than one hyperfamily [13]. In the small database of 80 folds then available, there were nine of these 'frequently occurring

Figure 2



Distribution of superfamilies in each fold. Most protein folds are used by just one known superfamily (far left). The ferredoxin-like fold, however, is adopted by 19 superfamilies and the α/β (TIM) barrel is adopted by 13. Names of all FOD folds are indicated. **(Inset)** The full graph. Adapted with permission from [54].

domains' (FODs) having more than one hyperfamily. As databases have grown, so have the FODs, which now number 42 in scop, or roughly an eighth of all folds.

Whereas the fraction of folds falling into the FOD category has remained roughly unchanged since the survey by Orengo *et al.* [13], the distribution in the current scop appears somewhat different. Orengo *et al.* [13] did not find any folds that contained two hyperfamilies; instead, they found a clear distinction between a very large number that had just one hyperfamily and a small number that had three or more. The overwhelming majority of folds still contain just one superfamily, but the number of folds that have two superfamilies is roughly twice the number that have three or more. Consequently, it is now harder to convincingly distinguish between the ordinary domains and the FODs.

The FODs perform a wide range of functions. For example, consider the ferredoxin-like fold, which is used by 19 superfamilies. Several of these superfamilies contain ribosomal proteins and proteins that bind nucleic acids outside the ribosome. The ferredoxin-like fold also appears in a number of enzymes, especially occurring as domains outside the main catalytic domain of peptidases. Among the remaining proteins with ferredoxin-like folds is, of course, its eponym — the electron transport protein ferredoxin.

All of the four main classes are represented among the FODs. Intriguingly, although the α/β barrel is certainly the best known FOD fold, there is actually a slight bias among the FODs for folds from the α + β or all-β classes.

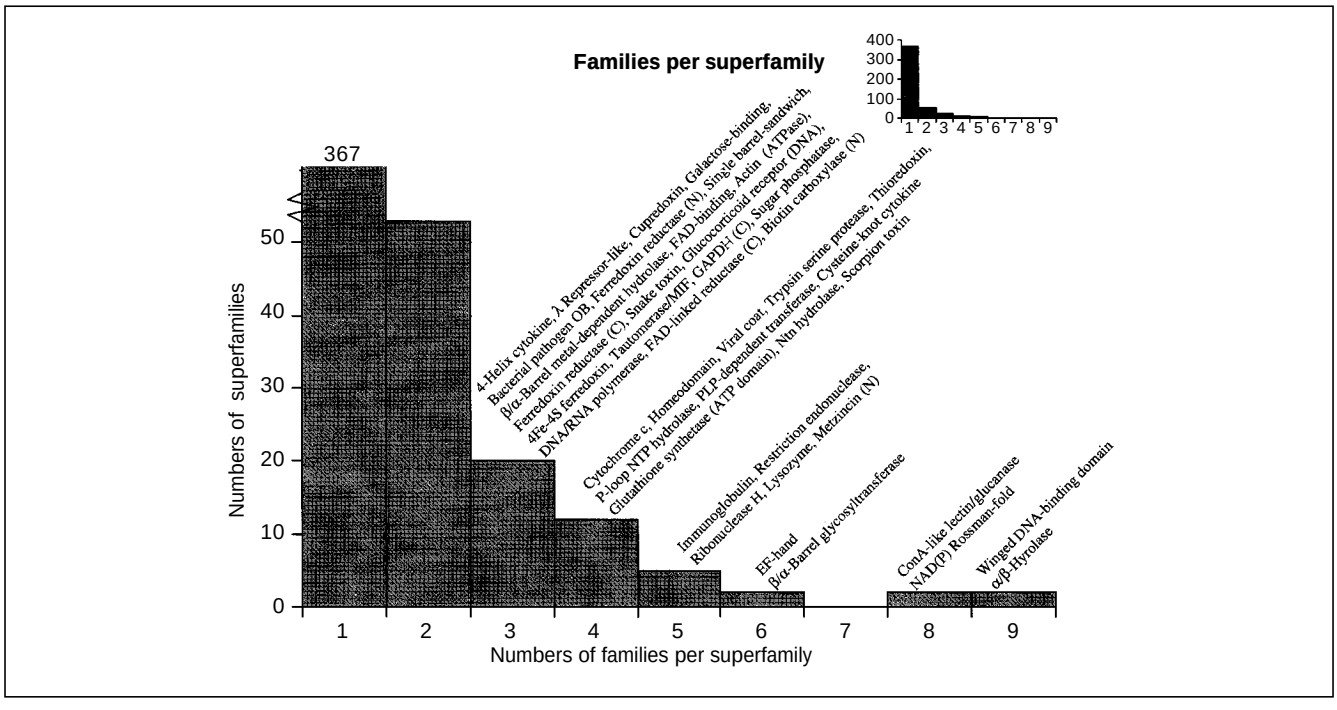
Considerable speculation has arisen concerning the reason for the skewed distribution and existence of FODs. By using lattice models, several groups have suggested that some protein folds can be formed by more sequences than others [32,33].

The PDB is biased in that it contains only those proteins that crystallize or that are suitable for NMR studies; this caveat presently limits nearly all of our current knowledge of protein structure. Within these boundaries, it is important that the distributions described are not the consequence of any deliberate bias among the experimentalists who solve protein structures. A protein that forms a new family or superfamily is — by definition — different in sequence from all proteins of known structure; before solving its structure, therefore, it is impossible to know whether it will have the same fold as other proteins.

True superfamilies

Similar to the situation of degeneracy with folds, an average of 1.4 families in each superfamily exists in scop. As shown in Figure 3, the distribution of these is also skewed; however, it is less skewed than for the folds: 20% (96 out of 463) of all superfamilies have more than one family, and no superfamily has more than nine families. This leaves 367 'potential superfamilies' that, at present, contain just a single family.

Figure 3



Distribution of families within superfamilies. The majority of protein superfamilies have diverged into only one known family. Some, such as the α/β hydrolases and winged-helix domains, have evolved very widely. Names of all superfamilies with more than two families are indicated. **(Inset)** The full graph. Adapted with permission from [54].

The 96 superfamilies that have more than one family are called 'true superfamilies', to reflect their umbrella nature that covers multiple, very distantly related sets of proteins. The true superfamilies play a wide variety of roles, as exemplified by the two superfamilies with the most families. The winged DNA-binding domains are, as their name implies, DNA-binding domains that play non-enzymatic roles in transcription activation and repression, and they are also found in the H1/H5 histones. By contrast, the α/β hydrolases are a broad family of enzymes — many of which have significant structural elaboration — including acetylcholinesterase, lipases, a peptidase and others.

FODs and true superfamilies

Given the skewed populations, it might be expected that true superfamilies would usually occur within FODs, leading to some folds that have enormous numbers of families. Interestingly, however, there is no correlation between the number of superfamilies in a fold and the number of families within a superfamily. Therefore, whereas some true superfamilies belong to FODs, most do not. For example, most superfamilies with the α/β-barrel fold have just one family. Conversely, the α/β-hydrolase fold has just one superfamily, but it contains nine families. In short, no more true superfamilies have FOD structures than would be expected to occur randomly.

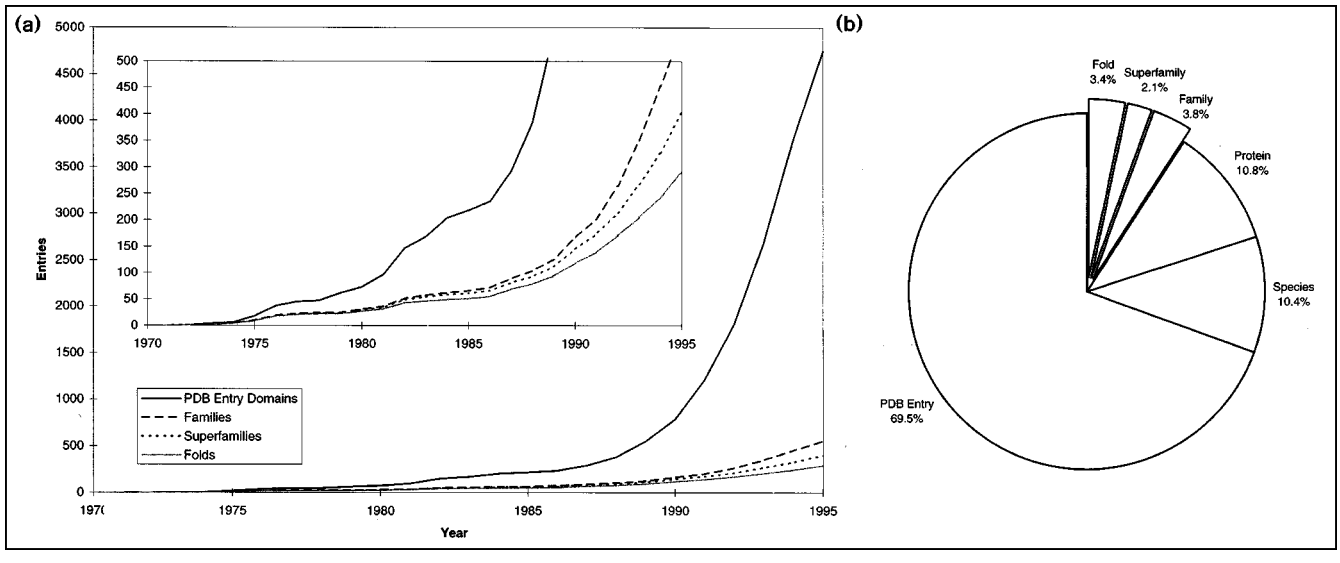
Classes

Fold classes were defined over twenty years ago as very general ways of describing folds that reflect the secondary-structure elements and general aspects of their arrangements [11]. If protein topology were random, most proteins would have mixtures of α helices and β strands. Instead, large numbers of proteins exist that have just helices or just strands, and these form the all-α and all-β classes, respectively. Among those that contain a mixture of α and β segments, there is an unexpected abundance of βαβ′ units, an aspect that recent statistical calculations have confirmed (M Levitt, personal communication). Consequently, proteins with α helices and β stands are divided into α/β for proteins that "have mixed or approximately alternating segments of α helical and β stand secondary structure" [11], and α+β for proteins in which the helical and sheet regions are somewhat segregated.

Because β and α folds are split across two classes, the classes are unique in the hierarchy for being fairly uniform in the number of items at the next lower level (folds) they contain: α, 71; β, 52; α/β: 64; α+β, 79.

These fold classes have taken on additional relevance for gross structure prediction from sequence [34–38]. Recently, Mitchie *et al.* [39•] have extensively re-evaluated

Figure 4



The discovery of structural information. **(a)** Cumulative growth of structural information in the PDB. **(Inset)** A magnified view highlighting the growth in folds, superfamilies, and families. **(b)** The distribution of new PDB entry domains in 1994—the most recent year for which complete information is available. The labels indicate the degree of new information provided by each new domain. For example, nearly 70% of proteins domains submitted to the PDB in 1994 were effectively identical to existing protein sequences, whereas only 3.4% of domains had a new fold. Only those domains in the exploded 9.3% had sequences dissimilar to proteins of known structure, and therefore only those had the potential to form a new fold. Adapted with permission from [54].

the classes and have found that simple procedures can identify a protein class in 90% of cases. However, they found significant overlap between the α+β and α/β classes and therefore argue that these should be merged into a single α&β class. Within this class, subdivisions are made on the basis of β-sheet topology.

Holm and Sander [40•] have also looked at coarse levels of protein similarity and have placed 40% of folds within five major groups they call attractors. The attractors bear very rough similarity to the four classes but focus more heavily on the parallel/antiparallel nature of β sheets, although not in the same way as those defined by Mitchie *et al.* [39•].

Structure information growth

The skewed statistics and many other general aspects of protein structures would be impossible to detect were it not for the meteoric growth in structure information, which is shown in Figure 4a. As recently as a decade ago, the PDB contained only 235 entry domains, 5% of the present number, and the three years from 1993 to 1995 saw a fourfold increase. However, during that later period, the number of folds increased just 70%. Thus, structural information has taken off in the past years, but redundancy (typically owing to more detailed studies of a single protein) has recently been increasing more rapidly than the discovery rate of entirely new structures.

This redundancy is especially clear in Figure 4b, which shows that in 1994 (the most recent year for which complete data are available), nearly 70% of the entries deposited were variants of essentially the same protein chain. A further 21% were clearly related to proteins of known structure. Only proteins in the remaining 9.3% of entries had, on the basis of lack of sequence similarity, the potential to be fundamentally new and unrelated. These entries, representing 103 structures, are distributed between those which defined a new family, superfamily, or fold.

How many superfamilies?

Over two decades ago, Dayhoff [41] and Zuckerkandl [42] independently proposed that the number of protein families was not only finite but was also, probably, a comparatively small number—around 1 000. Since those early speculations, the amount of sequence information has grown by orders of magnitude, leading to changes in our understanding of protein families. Nevertheless, when interest in the finite number of superfamilies in nature was reawakened five years ago [43], the conclusion was similar: the large majority of proteins come from no more than 1 000 families, and from an even smaller number of folds. At odds with this is a computation suggesting there are nearly 8 000 folds [13]. Although Orengo *et al.* [13] believe this number too be an overestimate, it implies a very large number of superfamilies.

The information in hierarchical databases can help address this issue, for it indicates whether a protein domain in a new family (i.e. having no sequence similarity to other proteins of known structure) is homologous, and thus belongs to an existing superfamily. If families

Table 1

Fraction of new families within existing superfamilies and estimate of total superfamilies in nature.

Year	SF* at end of year	Average SF during year†	New families	Families in existing SF	Percentage of families in existing SF	Estimated total SF‡
1989	112	–	–	–	–	–
1990§	146	129	41	7	17	759
1991§	174	160	36	8	22	727
1992	214	194	61	21	34	570
1993	267	240	87	34	39	615
1994	327	297	102	42	41	724
1995§	–	405	129	57	44	920

*SF, superfamily. †Taken from the average of the number of superfamilies at the end of the current and previous years. ‡Obtained by dividing the average number of superfamilies in a year by the fraction of new families that are in an existing superfamily. §Data for 1990 and 1991 should be considered less reliable because of the small number of families in existing superfamilies. Data for 1995, partially from Murzin [52], are approximate.

are uniformly distributed among superfamilies, we can examine the number of superfamilies known at a given point and the fraction of new families that fall into those superfamilies. The ratio will be the total number of superfamilies in nature,

$$\text{Total superfamilies} = \frac{\text{known superfamilies}}{\text{fraction of new families found to be in existing superfamilies}} \quad .$$

So, for example, if there were 500 superfamilies known (in structural detail) and one half of new families fell into an existing superfamily, we would expect the total number of superfamilies to be 1 000.

Table 1 shows the relevant numbers computed for the end of each accession year from 1990 to 1995. While the biased distribution of families in superfamilies complicates interpretation of these data, it is clear that all the large major superfamilies of proteins should soon be known at both the sequence and structural level.

Conclusions

The enormous growth of protein-structure information has been rendered comprehensible by classification databases. Simple statistical analyses of these databases reveal highly skewed distributions among the different levels, which suggests that some folds have independently evolved many times, whereas the vast majority were produced only once. Not correlated with this is the evidence that some superfamilies have diverged widely, whereas the majority now known have been restricted to similar sequences. This may be due to the fact that functional constraints place more restrictions on some families than others [44].

Even though only a tiny fraction of new structure-determination experiments reveal new protein folds, we will probably know most globular folds by the time that the human genome project [45] is completed (although biased distributions indicate that it may be a long time before all the folds are characterized). This provides great impetus and hope for the development of fold-recognition methods [46–48]. They will have to develop quickly, however, if they are not to be precluded by distant-homology recognition procedures [49–51], for by the time that the human genome project is finished, it is probable that structural representatives of most large superfamilies will also be known. Consequently, although the structures of fewer than 100 proteins were known a decade ago, we may know outline structures for the majority of proteins encoded by the human genome in the coming several years.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J: Protein Data Bank. In *Crystallographic Databases: Information Content, Software Systems, Scientific Applications.* Edited by Allen FH, Bergerhoff G, Sievens R. Cambridge: Data Commission of the International Union of Crystallography; 1987:107–132.

2. • Gray PMD, Kemp GJL, Rawlings CJ, Brown NP, Sander C, Thornton JM, Orengo CM, Wodak SJ, Richelle J: Macromolecular structure information and databases. *Trends Biochem Sci* 1996, 7:251–256.

A convenient summary of different protein structure resources, especially those available on the World Wide Web.

3. Wodak SJ: Extending molecular systematics to the third dimension. *Nat Struct Biol* 1996, 3:575–578.

4. •• 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.

This paper introduces the scop database, which hierarchically organizes all proteins of known structure. Scop principally uses manual techniques for domain definition and classification. See also http://scop.mrc-lmb.cam.ac.uk/scop/

5. Orengo CA, Flores TP, Taylor WR, Thornton JM: Identification and classification of protein fold families. *Prot Eng* 1993, 6:485–500.

6. •• Orengo C, Mitchie A, Jones S, Jones D, Swindells M, Thornton J: The CATH classification scheme of protein domain structural families. *Protein Data Bank Quarterly Newsletter* 1996, 78:8–9.

A brief announcement of the CATH database, which organizes protein domains into discrete hierarchical levels using a combination of manual and automated techniques, including the SSAP program described in [5]. See also http://www.biochem.ucl.ac.uk/bsm/cath/

7. •• Hogue CWV, Ohkawa H, Bryant SH: A dynamic look at structures: WWW-Entrez and the Molecular Modeling Database. *Trends Biochem Sci* 1996, 226–229.
This paper describes use of the Entrez Structure and the MMDB database, which provides links from each protein to others of similar structure. It is especially valuable because it is part of the larger Entrez system that encompasses protein and nucleic acid sequences and literature references. See also http://www.ncbi.nlm.nih.gov/Structure/

8. •• Holm L, Sander C: Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res* 1997, 25:231–234.
The FSSP database is an automated hierarchical database of structural relationships and is perhaps the most up to date classification. A particularly useful feature is the ability to find structures similar to those of one's own proteins by submitting coordinates to their server. See also http://www.embl-heidelberg.de/dali/fssp/

9. •• 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.
An automated classification of protein domains. This paper describes the results in useful detail. See also ftp://www.cryst.bbk.ac.uk/pub/ddbase/

10. • Brenner SE, Chothia C, Hubbard TJP, Murzin AG: Understanding protein structure: using scop for fold interpretation. *Methods Enzymol* 1996, 266:635–643.
How to understand a protein structure in the context of the scop database. The classification of the NAD(P)-binding Rossmann fold domains is explained as an example.

11. Levitt M, Chothia C: Structural patterns in globular proteins. *Nature* 1976, 261:552–557.

12. Flaherty KM, McKay DB, Kabsch W, Holmes KC: Similarity of the three-dimensional structures of actin and the ATPase fragment of a 70 kDa heat shock cognate protein. *Proc Natl Acad Sci USA* 1991, 88:5041–5045.

13. Orengo CA, Jones DT, Thornton JM: Protein superfamilies and domain superfolds. *Nature* 1994, 372:631–634.

14. Efimov AV: Common structural motifs in small proteins and domains. *FEBS Lett* 1994, 355:213–219.

15. • Efimov AV: A structural tree for α-helical proteins containing α-α-corners and its application to protein classification. *FEBS Lett* 1996, 391:167–170.
This paper, together with [14], describes a classification of domain structures based on an analysis that is very different from that used in [4••,5,6••–9••,10•]. The structural tree is built from the elaboration of a single α-α-corner motif.

16. Rackovsky S: Quantitative organization of the known protein X-ray structures. I. Methods and short-length-scale results. *Proteins* 1990, 7:378–402.

17. Hoffman DL, Laiter S, Singh RK, Vaisman II, Tropsha A: Rapid protein-structure classification using one-dimensional structure properties on the bioscan parallel computer. *Comp App Biosci* 1995, 11:375–679.

18. Taylor WR, Thornton JM: Recognition of super-secondary structure in proteins. *J Mol Biol* 1984, 173:487–512.

19. Stirk HJ, Woolfson DN, Hutchinson EG, Thornton JM: Depicting topology and handedness in jellyroll structures. *FEBS Lett* 1992, 308:1–3.

20. Pascarella S, Argos P: A data bank merging related protein structures and sequences. *Protein Eng* 1992, 5:121–137.

21. Blundell TL, Zhu ZY: The alpha-helix as seen from the protein tertiary structure: a 3D structural classification. *Biophys Chem* 1995, 55:167–184.

22. Sibanda BL, Thornton JM: β-hairpin families in globular families. *Nature* 1985, 316:170–174.

23. Pavone V, Gaeta G, Lombardi A, Nastri F, Maglio O, Isernia C, Saviano M: Discovering protein secondary structure: classification and description of isolated β-turns. *Biopolymers* 1996, 38:705–721.

24. Prestrelski SJ, Williams AL Jr, Liebman MN: Generation of a substructure library for the description and classification of protein secondary structure. I. Overview of the methods and results. *Proteins* 1992, 14:430–439.

25. Wintjens RT, Rooman MJ, Wodak SJ: Automatic classification and analysis of alpha-alpha-turn motifs in a proteins. *J Mol Biol* 1996, 255:235–253.

26. Martin ACR, Thornton JM: Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *J Mol Biol* 1996, 263:800–815.

27. Wintjens R, Rooman M: Structural classification of HTH DNA–binding domains and protein–DNA interaction modes. *J Mol Biol* 1996, 262:294–313.

28. Suzuki M, Brenner SE: Classification of multi-helical DNA-binding domains and applications to predict the DBD structures of σ factor, LysR, OmpR/PhoB, CENP-B, Rap1, and XylS/Ada/AraC. *FEBS Lett* 1995, 372:215–221.

29. Baxevanis AD, Landsman D: Histone and histone fold sequences and structures: a database. *Nucleic Acids Res* 1997, 25:272–273.

30. Boberg J, Salakoski T, Vihinen M: Selection of a representative set of structures from the Brookhaven Protein Data Bank. *Proteins* 1992, 14:265–276.

31. Hobohm U, Scharf M, Schneider R, Sander C: Selection of representative protein data sets. *Protein Sci* 1992, 1:409–417.

32. Li H, Helling R, Tang C, Wingreen N: Emergence of preferred structures in a simple model of protein folding. *Science* 1996, 273:666–669.

33. Govinarajan S, Goldstein RA: Why are some protein structures so common? *Proc Natl Acad Sci USA* 1996, 93:3341–3345.

34. Muskal SM, Kim SH: Predicting protein secondary structure content: a tandem network approach. *J Mol Biol* 1992, 225:712–727.

35. Dubchak I, Muchnik I, Holbrook SR, Kim SH: Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci USA* 1992, 92:8700–8704.

36. Reczko M, Karras D, Bohr H: An update of the DEF database of protein fold class predictions. *Nucleic Acids Res* 1997, 25:235.

37. Chandonia JM, Karplus M: Neural networks for secondary structure and structural class predictions. *Protein Sci* 1995, 4:275–285.

38. Chou KC: A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. *Proteins* 1995, 21:319–344.

39. • Mitchie AD, Orengo CA, Thornton JM: Analysis of domain structural class using an automated class assignment protocol. *J Mol Biol* 1996, 262:168–185.
An in-depth survey of the structural principles that define classes, and a test of the ability of automated procedures to identify them. This paper also investigates nonstructural characteristics of proteins in the different classes

40. • Holm L, Sander C: Mapping the protein universe. *Science* 1996, 273:595–602.
A broad survey of known protein structures, which introduces the concept of five attractors as very common protein architectural frameworks.

41. Dayhoff MO: Computer analysis of protein sequences. *Feder Proc* 1974, 33:2314–2316.

42. Zuckerkandl E: The appearance of new structures and functions in proteins during evolution. *J Mol Evol* 1975, 7:1–57.

43. Chothia C: One thousand families for the molecular biologist. *Nature* 1992, 357:543–544.

44. Chothia C, Gerstein M: How far can sequences diverge? *Nature* 1997, 385:579–580.

45. Collins FS: Ahead of schedule and under budget: the genome project passes its fifth birthday. *Proc Natl Acad Sci USA* 1995, 92:10821–10823.

46. Lemer CM-R, Rooman MJ, Wodak SJ: Protein structure prediction by threading methods: evaluation of current techniques. *Proteins* 1995, 23:337–355.

47. Bryant SH, Altschul SF: Statistics of sequence–structure threading. *Curr Opin Struct Biol* 1996, 5:236–244.

48. Fischer D, Rice D, Bowie JU, Eisenberg D: Assigning amino acid sequences to 3-dimensional protein folds. *FASEB J* 1996, 10:126–136.

49. Eddy SR: Hidden Markov models. *Curr Opin Struct Biol* 1996, 6:361–356.

50. Gribskov M, Veretnik S: Identification of sequence patterns with profile analysis. *Methods Enzymol* 1996, 266:198–211.

51. Bork P, Koonin EV: Protein sequence motifs. *Curr Opin Struct Biol* 1996, 6:366–376.

52. Murzin AG: Structural classification of proteins: new superfamilies. *Curr Opin Struct Biol* 1996, 6:386–394.

53. Kizaki H, Hata Y, Watanabe K, Katsube Y, Suzuki Y: Polypeptide folding of ***Bacillus cereus*** ATCC7064 oligo-1,6-glucosidase revealed by 3.0 Å resolution X-ray analysis. *J Biochem (Tokyo)* 1993, 113:646–649.

54. Brenner SE: Molecular proprinquity: evolutionary and structural relationships of proteins [PhD Thesis]. Cambridge: Cambridge University; 1996.