Opinion

Progress towards mapping the universe of protein folds

Alastair Grant, David Lee and Christine Orengo

Address: Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK.

Correspondence: Alastair Grant. E-mail: grant@biochem.ucl.ac.uk

Published: 29 April 2004

Genome ***Biology*** 2004, **5:**107

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/5/107

Abstract

Although the precise aims differ between the various international structural genomics initiatives currently aiming to illuminate the universe of protein folds, many selectively target protein families for which the fold is unknown. How well can the current set of known protein families and folds be used to estimate the total number of folds in nature, and will structural genomics initiatives yield representatives for all the major protein families within a reasonable time scale?

In order to attempt predictions of the universe of protein folds - so-called fold space - we need to know how many protein families there are in nature and how many of these are likely to possess a novel fold. Genome sequencing still considerably outpaces the various structural genomics initiatives currently underway in the USA, Canada, Japan, Germany and the UK, with more than 160 completely sequenced genomes yielding about one million protein sequences at the start of 2004 [1]. This contrasts with 24,000 entries of three-dimensional protein structures in the Protein Data Bank (PDB) [2,3], some 500 of which were determined by structural genomics consortia over the last three years. Although this seems a daunting contrast, mounting evidence from the Gene3D (our unpublished data and [4]), SUPERFAMILY [5,6], and Genomic Threading [7,8] databases suggests that a relatively small repertoire of protein folds (around 800) can already be mapped onto about half of all the amino-acid residues encoded in the currently available genome sequences.

Encouragingly, and in parallel with the expansions in the structure and sequence databanks over the last decade, powerful new technologies have been developed for recognizing relationships between proteins on the basis of sequence and/or structural similarity [9]. These allow the universe of protein-family space to be more accurately charted, by allowing recognition of extremely distant homologs.

Estimations of the number of folds

Although Wolf *et al*. [10] attempted to predict the number of folds in individual genomes, most estimates consider the total number of folds in all of nature. Current estimates of the number of folds range from 1,000 to 10,000, depending on the models and approximations applied [11-13]. One of the earliest estimates of fold numbers was a simple approximation by Chothia [14]. This assumed that there is a limited number of folds in nature that sequences can adopt, given the intrinsic physical constraints. If these are randomly sampled in the projects that solve protein structures, then the probability that a new protein sequence has a known fold can be estimated by determining the proportion of unrelated sequences, for example in the structure classifications database SCOP [15,16], that share the same fold as one another and are therefore likely to share that fold with the new sequence. This approach predicted around 1,000 folds, given the proportion of sequences of known structure in SCOP that had unique folds, the fraction of the Swiss-Prot sequence database [17,18] these sequences comprised, and the fraction of new sequences found to be related to sequences already in Swiss-Prot.

A similar model applied by our group [19] also took account of the number of protein families in Swiss-Prot. Using the CATH structure database [20,21], we predicted a higher estimate of around 8,000 folds. Both these simplistic calculations

[14,19] ignore bias in the sequence and structure databases and the fact that some folds, often referred to as superfolds [20], are more highly reused by different protein families in nature than is expected by chance. This uneven fold-family distribution, revealed by several analyses [22-24] can be clearly seen in Figure 1, which shows that a small percentage of fold groups in the CATH domain structure database (54 folds, or 6.6% of the total) are very highly populated, accounting for 76% of domain families for which a structure has been predicted, whilst there are many folds adopted by only a single family.

Although similarity in the folds adopted by different families may reflect folding preferences and convergence to energetically stable folds, it is likely that many of the families that adopt the superfolds are in fact very distantly related, beyond the sensitivity of current algorithms to detect homology at the sequence level. Families adopting the eight-stranded α/β TIM-barrel folds are a case in point, with recent analysis suggesting that many of these families may have evolutionary links - an idea that is supported by unusual sequence signatures and functional properties [25,26].

Since Chothia's early estimates [14], several groups have applied more sophisticated statistical approaches that model the uneven distribution of fold usage in various ways [22,24]. Random sampling of known sequence families and assigning equal likelihood to each fold gives rise to a non-uniform fold distribution which, when further modified to account for the extreme bias of the superfolds and the fact that many folds are only rarely seen in nature, gives an estimate of 4,000 folds [23].

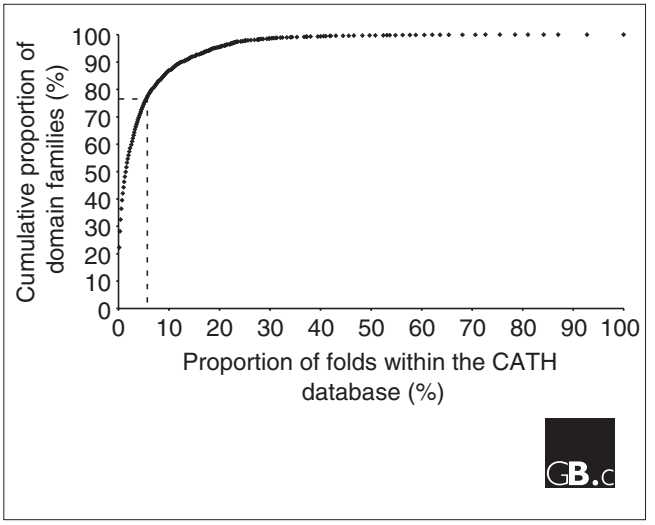

Figure 1
The proportion of domain families represented by CATH fold groups. Within the CATH database [20,21], structures are grouped into fold groups on the basis of both overall shape and connectivity of their secondary structures. Domain families are related at the 35% sequence identity level by complete linkage clustering. The number of domain families within each fold group gives a measure of the sequence diversity of that fold group. A group of 54 CATH fold groups (only 6.6% of the cumulative total of CATH fold groups) accounts for 76% of domain families, as shown by the dotted lines.

Coulson and Moult [12] assume the existence of three types of folds: superfolds, which are adopted by very many protein families and are highly recurrent within proteomes; mesofolds, which have an intermediate number of protein families associated with them; and unifolds, adopted by a single narrow sequence family. On the basis of this assumption, they simulated the expansion of new folds classified in the SCOP structure database over the preceding two years, as a fraction of new sequence families added. Assuming a maximum of 50,000 protein families in nature, this approach predicts up to 400 mesofolds and some 10,000 unifolds in addition to 9 superfolds. Perhaps more importantly, the majority of sequence families belong to superfold and mesofold groups, and for 80% of these families we probably already know the fold.

Several groups have attempted to model the uneven fold-family distribution using power laws. Power law distributions - in which a small number of high-frequency instances occur, but there is a moderate number of common instances and a huge number of very rare instances - appear to be ubiquitous in nature and society, and seem to explain many of the biological trends recently revealed by genome data, such as protein-family distributions, domain associations, and protein-protein interactions [13,27,28]. Karev *et al.* [29,30] model protein-family distributions by simulating the birth (gene duplication), death (gene loss) and innovation (new protein) of different domains in individual genomes. Although this entirely stochastic model fails to account completely for the observed distribution, it shows that a close fit is possible using a model with only three independent parameters. Implicit in the model is the notion that the 'fit' get 'fitter', and domains randomly duplicated early in evolution increasingly dominate the population. None of these models incorporates selection pressures that might operate to favor the retention of duplicated domains performing important biochemical activities. But, in fact, many highly recurrent domains do appear to have important biochemical functions, for example in providing energy or redox equivalents for enzyme reactions, or in responding to cellular signals and binding to DNA [31,32].

These more recent models of the number of folds [12,22-24,29,30] continue to ignore possible biases in the structure and sequence databases. For example, it is likely that proteins sampled for structure determination have been relatively easy to solubilize, purify and crystallize - as shown by the small numbers of transmembrane structures known. Perhaps more worrying are recent analyses suggesting that we have barely sampled sequence and family space, as each new genome adds more families and there is no sign of saturation

in this expansion [33]. Even with the huge advances in genome sequencing, there are still at least ten million organisms as yet uncharacterized [13].

To be more optimistic, though, it is likely that as the sequence and structure databases expand, making it easier to link relatives and also increasing the sensitivity of the profile-based homology search methods and fold-recognition methods, there may be a considerable coalescence of families. Assessment of several widely used homolog-detection methods (such as PSI-BLAST and hidden Markov models, HMMs) using structurally validated homologs has shown significant increases in performance accompanying expansions in the sequence and structure databases [32].

How many protein families are currently recognized?

Given that most estimates of how many folds there are depend heavily on the numbers of protein families that have been identified and their mapping to existing folds, it is useful to briefly consider the current strategies and technical challenges involved in identifying these families. Structural genomics initiatives have promoted several new sequence-based approaches to recognizing protein families. These arose because although there are many well-established protein-family databases (such as PRINTS [34,35], Pfam [36,37], SMART [38,39], ProDom [40,41], InterPro [42,43], TIGRFAMs [44,45] and MIPS [46,47]) most cover only a relatively small proportion of the known sequences. Pfam [36,37], which now includes over 7,000 manually curated families, identifies many of the largest protein families, and any lack of coverage is addressed to a certain extent by InterPro [42,43], which integrates Pfam with several other protein-family resources. The advantage of all these curated databases is that relatives are recognized using family-specific sequence profiles or regular expressions, and there is some degree of manual validation.

Faster approaches for identifying protein families within very large datasets (such as those in non-redundant GenBank [48,49] or Swiss-Prot/TrEMBL [17,18]) often involve aligning the sequences against each other using BLAST and then clustering those with significant similarity [50-54]. The simplest protocols use single-linkage clustering, which often collapses too many families, giving relatives with insufficient global similarity. In ProtoNet [50,51] these effects are robustly handled by permitting alternative user-defined thresholds for clustering that allow granularity to range from families with small closely related proteins to much broader families comprising proteins sharing common sequence motifs. Some of the most promising new methods employ Markov clustering, in particular the TribeMCL [55] implementation developed by Enright and co-workers and used by the TRIBES [56,57] and Gene3D resources (our unpublished data and [4]).

One of the hardest problems in clustering sequences into protein families is handling the similarities between multi-domain proteins and the fact that many different multi-domain proteins share common domains but in different contexts. A significant proportion [58] of proteins are multi-domain - up to 80% in eukaryotes. Furthermore, Teichmann and others [58] have shown that domains have frequently been shuffled and recombined in different ways within genomes, often giving rise to subtly different functions [59].

This recurrence of domains suggests their importance as primary evolutionary units, and although some researchers hypothesize that smaller supersecondary structural motifs may be the building blocks of evolution [60], the majority of globular compact folds characterized to date comprise whole domains. Thus, although some protein-family resources cluster complete gene sequences into families, most attempt to divide proteins into their constituent domains before or after clustering. Recognizing the boundaries of domains is a non-trivial algorithmic challenge, however, particularly if no structural data are available. Even methods based on structures disagree in their assignments 20-40% of the time [61]. The problem is compounded by discontinuities in some domain sequences, whereby the insertion of a second domain disrupts an existing domain within a multi-domain protein. Structural data in the CATH database [20,21] suggest that these discontinuities exist in about 23% of domains occurring in multi-domain proteins [62].

Some of the most successful approaches to the problem of domain-boundary prediction combine sequence data with the propensities of particular amino-acid residues, using neural networks [54,63,64]. Other methods exploit the recurrence of domains in different contexts to identify boundaries from multiple alignments [40,65,66]. The elegant approach of Heger and Holm (named ADDA [66]) exploits graph theory to build networks of domain links in multi-domain proteins from which multiple alignments can be extracted and recursively analyzed and chopped up to yield their single-domain components.

Estimates of the number of protein families that have so far been identified vary substantially, depending on the sequence datasets clustered and the thresholds employed. The ADDA algorithm of Heger and Holm [66] identifies some 34,000 domain families in a combined sequence dataset - derived from Swiss-Prot, TrEMBL, the Protein Information Resource (PIR), PDB, the *Caenorhabditis elegans* protein database Wormpep and Ensembl genome databases - which, after removing redundancy at 40% sequence identity, contained almost 250,000 protein sequences. These are chopped into domains and then clustered into 34,000 domain families. Almost 170,000 domains remain as singletons that are not clustered into any family. Similarly, a recent analysis by Liu and Rost [67], chopping and clustering sequences from eukaryotic genomes,

suggested 17,000 domain-like clusters (regions likely to be domains) in eukaryotes that are likely to have a currently unidentifiable globular structure. Again these represent low estimates, as the eukaryotic genomes currently contribute about half of the total sequences within completed genomes. A more recent publication reports 63,000 domain families from the clustering of 62 complete genomes [68,69].

In our work to develop the Gene3D database of annotated complete genomes [4], we benefited from a number of publicly available algorithms [55,70] and resources [48,49,71]. Our Pfscape protocol (unpublished) first clusters the 600,000 sequences from 120 completed genomes into 59,000 gene families using the TRIBE-MCL algorithm [55], with some 112,000 singleton sequences remaining. Pfscape then maps CATH and Pfam domains onto sequences in these gene families using the SAM-T99 hidden Markov model method [72]. In addition to the 1,277 CATH-domain families and 5,179 (non-overlapping) Pfam-domain families that are recognized, a further 46,000 or so uncharacterized domain families remain, giving a total of almost 53,000 domain families. Figure 2 shows that most of these remaining uncharacterized families (termed NewFam) tend to have far fewer members than the CATH and Pfam families.

Many of the largest families in Gene3D are very sequence-diverse and are perhaps better described as superfamilies, containing some very distant homologs (proteins with less than 20% sequence identity). Thus, although Gene3D identifies almost 53,000 domain superfamilies, these comprise

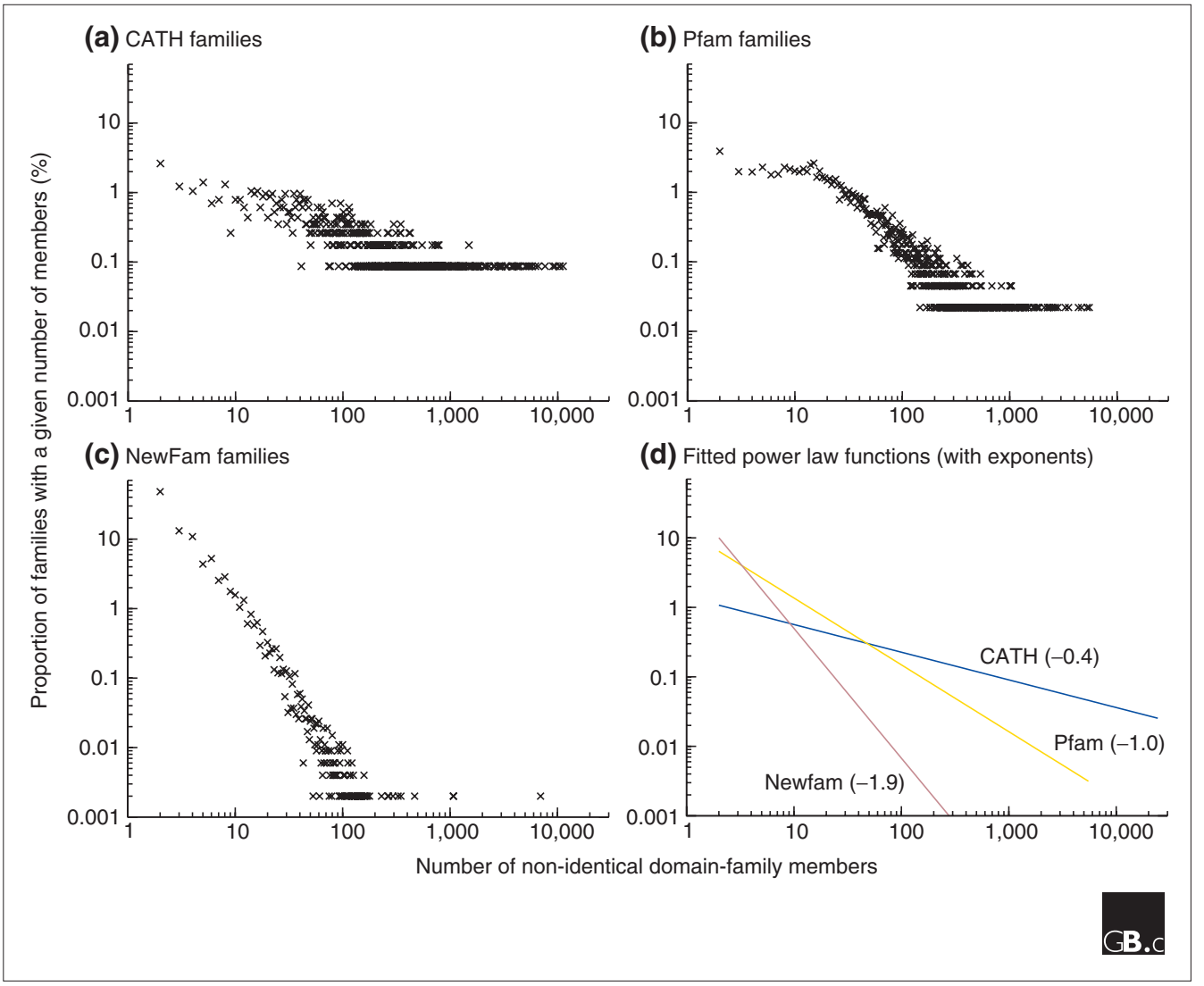

Figure 2
Log-log plots of the sizes of **(a)** CATH, **(b)** Pfam and **(c)** NewFam (uncharacterized) families show power-law-like behavior. **(d)** Fitted power law functions and their exponents are shown for comparison. Most NewFam families have relatively few members. See text for further details.

205,000 close families, in which relatives have 35% or more sequence identity; and at least 20% of these close families have one or more members with at least 35% sequence identity to a known structure. This suggests that structural genomics initiatives would need to target representatives for the remaining 165,000 or so families to obtain good structural models for all families in the examined genomes.

Mapping known protein folds and families onto the genomes

There are now more than 180 completed genomes. What proportion of these are we able to map with the current fold and family classifications? Although estimates of the total number of folds, ranging up to 10,000, suggest that we are a long way from knowing the full fold repertoire, recent analyses of fold and family distributions within sequenced genomes (by SUPERFAMILY [5,6], Gene3D [4] and the Genomic Threading Database [7,8]) using structure-classification databases (SCOP [15,16] and CATH [21,73]) suggest that between one third and two thirds of residues can be assigned to structurally characterized families in SCOP and CATH, adopting around 800 folds in total. Specifically, it is possible to assign folds to between 44% and 81% by HMM, and to achieve 64% average coverage by threading, of sequences in completed genomes; and 26-70% by HMM and 58% average coverage by threading is possible on a residue basis (see Figure 3 for coverage of some representative genomes by Gene3D).

From recent analyses using Gene3D domain families, after exclusion of singleton sequences, 50% of domains can currently be assigned to 1,277 superfamilies (93,571 close families) of known structure in the CATH database (Table 1). A further 33% of domains of no known structure can be assigned to about 1,832 Pfam superfamilies (61,722 close families; see Figure 4). The remaining 17% of domains have been assigned to NewFam uncharacterized domain families (52,973 close families; see Figure 2), most of which are small families.

Several analyses (for example [74,75]) have shown that approximately 22% of predicted protein sequences from genome sequences (which will overlap to some extent with CATH and Pfam assignments) contain transmembrane regions, and about 10-20% of predicted sequences contain long regions (50-100 amino acids) of disorder or low complexity. There is also a significant proportion (around 16%) of small amino-acid sequences with no predicted secondary structure [74].

Are the singletons - of which there are currently 60,000 in Gene3D - in fact distant relatives of existing families that are not recognized by current algorithms, or are they genuinely unique sequences having novel folds? Kunin and co-workers [33] recently showed that although some singletons are reassigned to families as new genomes are completed, there is still an overall gain in the number of singletons with each

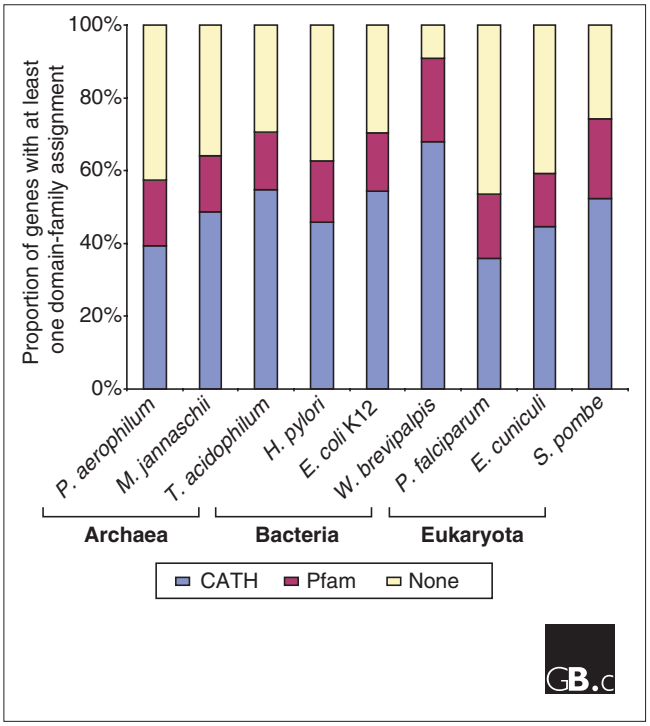

Figure 3
Gene coverage in Gene3D. The chart indicates the percentage of genes in the indicated genome that have at least one non-overlapping assignment from CATH or Pfam. Three representative genomes from each kingdom of life show low, average and high coverage, respectively. The species shown are *Pyrobaculum aerophilum, Methanococcus jannaschii, Thermoplasma acidophilum, Helicobacter pylori, Escherichia coli* K12, *Wigglesworthia glossinidia brevipalpis, Plasmodium falciparum, Encephalitozoon cuniculi* and *Schizosaccharomyces pombe*.

additional sequenced genome. This may change as the databases expand and recognition methods improve. Original estimates of the proportion of singletons in bacterial genomes lay at about 50% [22], but this number has steadily fallen, with average values of 30% for the first release of Gene3D in 2002 [76], and 18% for more recent releases of Gene3D [4]. Some proportion of these proteins may nevertheless represent genuinely new families and folds.

The length distribution of singletons is lower than the length distribution for the average structural domain [74], and many of the very small sequences containing disordered regions may correspond to unstructured proteins existing only as complexes and/or peptides involved in regulation and binding to DNA. These proteins may therefore not fold independently and will lie outside the range of targets amenable to structural genomics.

Revisiting the fold calculations

Using the number of domain families identified by Gene3D (see Figures 1 and 4), we can make a very simple approximation

Table 1

A summary of the families and superfamilies within Gene3D

Type of family	Proportion of non-singleton domains	Number of superfamilies	Number of close families	Number of folds
Known structure (CATH)	50%	1,277	93,571	759 + 54
Superfolds (all of known structure)			*71,080*	*54*
Unknown structure (Pfam)	33%	1,832	61,722	3,871
NewFam	17%		52,973	
Total, excluding singletons:			**208,266**	**4,684**

Data are from [4]; NewFam denotes uncharacterized families. Around 60,000 singletons are excluded from the analysis. See the text for how the number of folds is estimated for the domains of unknown structure.

of the total number of folds in nature by making the following four assumptions. First, we assume that we now know the folds for all the superfolds - defined as folds with three or more homologous superfamilies in CATH (at present this accounts for 71,080 close domain families for 54 highly populated CATH folds; see Table 1). Second, we assume that we have been able to map these folds onto all their relatives in the genome sequences, and so we can remove these folds and families while estimating the remaining numbers of folds. Third, we assume that singletons can be removed from the estimate, as they are probably very distant relatives belonging to known folds that have diverged beyond the sensitivity of current recognition methods, or else they are short sequences unlikely to fold independently but associated with functional complexes. Although singletons could represent novel folds and could therefore skew any estimate of the total number of protein folds, they do not represent a significant proportion of domains. Finally, we assume that non-superfolds and non-singletons have been sampled randomly by families in nature and that there are no biases in their representation within the current sequence and structure databases.

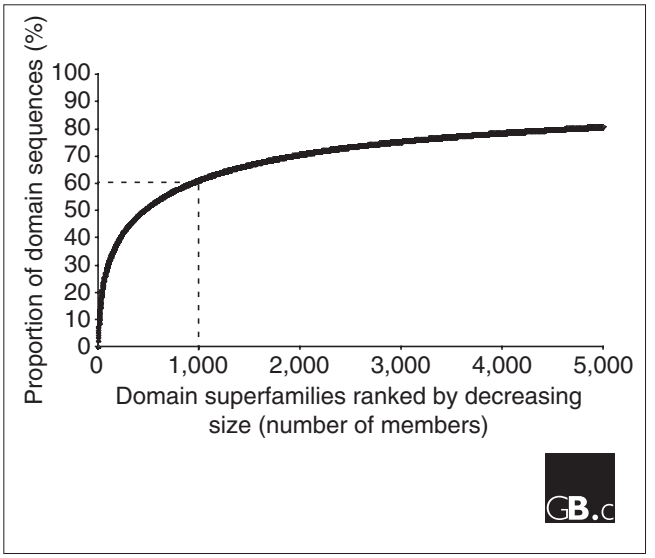

Figure 4
The cumulative number of domains within domain superfamilies (ranked by decreasing size). The 1,000 largest domain superfamilies account for nearly 60% of all domain sequences (see dotted lines). The figure excludes singleton domain families, and is derived from our own unpublished work.

Removing the 54 superfolds from the Gene3D dataset leaves 22,491 close domain families of known structure (see Table 1), which adopt 759 folds in CATH (see Figure 1). We can therefore expect the remaining 114,695 domain families in Gene3D that are of unknown structure (Pfam close domain families plus NewFam close domain families) to adopt (114,695/22,491) x 759, or 3,871 new folds. Adding together the superfolds, known folds and estimated number of new folds (54 + 759 + 3,871) we get an estimate of the number of folds encoded within the 120 genomes included in Gene3D of 4,684. This will probably be a lower bound for the total number of protein folds in nature. But all fold estimates are unsatisfying, in that they necessitate simplified models of fold usage and optimism regarding lack of bias in the databases; whilst our sampling of 'species' space remains so sparse, calculations on the numbers of folds in all of nature seem rather esoteric.

A few large protein families dominate more than half of all predicted protein sequences

Perhaps a more optimistic outlook for the structural genomics initiatives comes from the observation that fewer than 1,000 large CATH and Pfam families map to a significant proportion (around 60%) of all the predicted products of genome sequences, excluding singletons (see Figure 4). What roles are relatives from these large families performing and why are they recurring so frequently within the proteomes?

We used Gene3D to examine the recurrence of structurally characterized families in the predicted proteomes of a set of 56 bacterial genomes [77]. Interestingly, some 274 CATH-defined families are common to a significant proportion of these genomes. Less than 30 of the families are highly duplicated, dominating almost 50% of all the CATH-annotated

genome sequences. In these families, domain recurrence in any proteome correlates with genome size and, in some families, domains are frequently located in proteins with different domain compositions [59]. Many are associated with metabolic pathways, where they perform generic functions such as the provision of energy or redox equivalents for reactions. Frequently some aspect of the chemistry is conserved between paralogs, but substrate specificity may have been modulated by changes in the geometry of active sites. In some cases structural embellishments to the fold cause changes in surface geometry, modulating protein-protein interactions and altering the repertoire of domain associations [59]. A significant proportion adopt a small number of folds, namely TIM-barrel folds, Rossmann-like folds or αβ-plait superfolds. Interestingly, these are among the most ancient folds [78,79]. They all possess simple, regular, layered architectures that might be expected to promote optimal packing of hydrophobic residues in the core of the protein. In support of these hypotheses, Caetano-Anolles and Caetano-Anolles [79] have also proposed that αβ sandwiches and β-barrel-like structures evolved first, with β sandwiches evolving later, predominantly in eukaryotes, where the all-β immunoglobulin superfold recurs extensively. The regularity of their architectures may contribute to the ease with which these folds have been observed to tolerate residue mutations [80], allowing some of the families to diverge further and to adopt a range of different functions.

In addition, functional utility may also contribute to the wide recurrence of these domains [13]. As Koonin and coworkers propose [13], some perform generic functions that are well conserved (for example, nucleotide binding in the Rossmann-like domains) and have been re-used in multiple functional contexts (in different pathways or cellular locations). Alternatively, as in the case of TIM barrels and αβ-plait folds, these architectures possess functional sites (for example the base of the β barrel in the TIMs or the exposed β-sheet surface in the αβ-plaits) that can easily be re-engineered to bring diverse combinations of residues into contact, thereby creating novel catalytic environments.

How unrealistic are fold estimates?

Our estimates here, made using Gene3D, suggest that the largest, most recurrent families encoded within the sequenced genomes have already been characterized in the CATH database and can be expected to adopt about 800 folds. How realistic are our simple estimates of approximately 3,900 folds to be adopted by the remaining families, most of which are characterized in Pfam and some of which are quite small? (For example, Figure 2 shows that the remaining uncharacterized NewFam families are generally much smaller than the CATH and Pfam families.) Small families may turn out to be very distant relatives of superfolds that have diverged beyond recognition, and in acquiring highly specialized functions these now have the narrow sequence constraints observed today [62]. Some may be completely new folds, however, that have arisen by more recent shuffling of subdomains and motifs. Soding and Lupas [60] have presented some intriguing models of evolutionary pathways using diverse recombination of small common submotifs such as α hairpins and αβ motifs. There are fascinating examples of relatives in some families that appear to have acquired new folds through subtle rearrangements within supersecondary motifs [60,81].

It is clear that some common structural motifs are highly reused [82], and this has meant that fold space should perhaps more accurately be viewed as a continuum [83,84], where significant structural overlaps occur in some regions. For the most highly populated architectures within CATH (αβ sandwiches and β sandwiches), folds are often highly 'gregarious' (that is, some subcomponents of the fold are shared with other folds), with at least 40-50% of their structures overlapping structures from other fold groups. Given that the relatives in many large superfamilies adopting these architectures (for example, superfamilies adopting Rossmann-like folds or αβ-plait folds) can be highly structurally divergent, with only 50% of residues in the core remaining structurally conserved during evolution [85], these overlaps can create problems in identifying distinct regions within fold space. The continuous nature of fold space may mean that simulations exploring the number of folds in nature are unrealistic, and that it may be more useful to try to understand the mechanisms by which common motifs can be assembled.

In this context, it is notable that there have recently been some considerable successes in *ab initio* structure prediction, using approaches that assemble proteins from peptide fragment libraries derived from known structures [86]. There now appear to be structural representatives for most 10-15 residue peptides [87], particularly those occurring within secondary structures, and so these advances may become increasingly important for structural modeling of the large number of singletons and 'unifolds' revealed by genome analyses. Such coarse models could help in suggesting the location of an active site or functional interface, perhaps allowing the putative biochemical role of the protein to be modeled in a systems biology context, even if they are not of sufficiently high accuracy to allow drug design.

In summary, attempts to predict the total number of folds in nature are still hampered by uncertainties and approximations. Most calculations predict somewhere in the range of 1,000-10,000 folds. Encouragingly for our understanding of evolution and biological systems, we now know the fold for many of the largest families, in particular those that dominate the genome annotations. Some 800 CATH folds and an additional 1,830 structurally uncharacterized Pfam families can already be assigned to approximately 70% of proteins predicted from genome sequences. Structural genomics initiatives that target the large structurally uncharacterized

families can be expected to succeed in mapping fold space for a significant proportion of sequence space over the coming years.

References

1. Bernal A, Ear U, Kyrpides N: **Genomes OnLine Database (GOLD): a monitor of genomes projects world-wide.** *Nucleic Acids Res* 2001 **29:**126-127.
2. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235-242.
3. **Protein Data Bank** [http://www.rcsb.org/pdb/]
4. **Gene3D** [http://www.biochem.ucl.ac.uk/bsm/cath/Gene3D/]
5. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J: **The SUPERFAMILY database in 2004: additions and improvements.** *Nucleic Acids Res* 2004, **32 Database issue:**D235-D239.
6. **Superfamily** [http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/]
7. McGuffin LJ, Street SA, Bryson K, Sorensen SA, Jones DT: **The Genomic Threading Database: a comprehensive resource for structural annotations of the genomes from key organisms.** *Nucleic Acids Res* 2004, **32 Database issue:**D196-D199.
8. **The Genomic Threading Database** [http://bioinf.cs.ucl.ac.uk/GTD/]
9. Lee D, Grant A, Buchan D, Orengo C: **A structural perspective on genome evolution.** *Curr Opin Struct Biol* 2003, **13:**359-369.
10. Wolf EI, Grishin NV, Koonin EV: **Estimating the number of protein folds and families from complete genome data.** *J Mol Biol* 2000, **299:**897-905.
11. Leonov H, Mitchell JSB, Arkin IT: **Monte Carlo estimation of the number of possible protein folds: effects of sampling bias and folds distributions.** *Proteins* 2003, **51:**352-359.
12. Coulson AFW, Moult J: **A unifold, mesofold and superfold model of protein fold use.** *Proteins* 2002, **46:**61-71.
13. Koonin EV, Wolf YI, Karev GP: **The structure of the protein universe and genome evolution.** *Nature* 2002, **420:**218-223.
14. Chothia C: **Proteins. One thousand families for the molecular biologist.** *Nature* 1992, **357:**543-544.
15. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2004: refinements integrate structure and sequence family data.** *Nucleic Acids Res* 2004, **32 Database issue:**D226-D229.
16. **SCOP: Structural Classification of Proteins** [http://scop.mrc-lmb.cam.ac.uk/scop/]
17. Boeckmann B, Bairoch A, Apweiler R, Blatter M, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I, *et al.*: **The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003.** *Nucleic Acids Res* 2003, **31:**365-370.
18. **Swiss-Prot** [http://www.expasy.org/sprot/]
19. Orengo CA, Jones DT, Thornton JM: **Protein superfamilies and domain superfolds.** *Nature* 1994, **372:**631-634.
20. Orengo CA, Michie AD, Jones S, Jones DTY, Swindells MB, Thornton JM: **CATH - a hierarchical classification of protein domain structures.** *Structure* 1997, **5:**1093-1108.
21. **CATH Protein Structure Classification** [http://www.biochem.ucl.ac.uk/bsm/cath/]
22. Zhang C, DeLisi C: **Estimating the number of protein folds.** *J Mol Biol* 1998, **284:**1301-1305.
23. Govindarajan S, Recabarren R, Goldstein RA: **Estimating the total number of protein folds.** *Proteins* 1999, **35:**408-414.
24. Govindarajan S, Goldstein RA: **Why are some proteins structures so common?** *Proc Natl Acad Sci USA* 1996, **93:**3341-3345.
25. Copley RR, Bork P: **Homology among (betaalpha)(8) barrels: implications for the evolution of metabolic pathways.** *J Mol Biol* 2000, **303:**627-641.
26. Nagano N, Orengo CA, Thornton JM: **One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions.** *J Mol Biol* 2002, **321:**741-765.
27. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model.** *J Mol Biol* 2001, **313:**673-681.
28. Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M: **The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties.** *Genome Biol* 2002, **3:**research0040.1 - 0040.7.
29. Karev GP, Wolf YI, Koonin EV: **Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve?** *Bioinformatics* 2003, **19:**1889-1900.
30. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV: **Birth and death of protein domains: A simple model of evolution explains power law behavior.** *BMC Evol Biol* 2002, **2:**18.
31. Chothia C, Gough J, Vogel C, Teichmann SA: **Evolution of the protein repertoire.** *Science* 2003, **300:**1701-1703.
32. Pawlowski K, Rychlewski L, Zhang B, Godzik A: **Fold predictions for bacterial genomes.** *J Struct Biol* 2001, **134:**219-231.
33. Kunin V, Cases I, Enright AJ, de Lorenzo V, Ouzounis CA: **Myriads of protein families and still counting.** *Genome Biology* 2003, **4:**401.
34. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell AL, Moulton G, Nordle A, Paine AK, Taylor P, *et al.*: **PRINTS and its automatic supplement, prePRINTS.** *Nucleic Acids Res* 2003, **31:**400-402.
35. **PRINTS** [http://bioinf.man.ac.uk/dbbrowser/PRINTS/]
36. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer ELL, *et al.*: **The Pfam protein families database.** *Nucleic Acids Res* 2004, **32 Database issue:**D138-D141.
37. **Pfam** [http://www.sanger.ac.uk/Software/Pfam/]
38. Letunic I, Goodstadt L, Dickens NJ, Doerks T, Schultz J, Mott R, Ciccarelli F, Copley RR, Ponting CP, Bork P: **Recent improvements to the SMART domain-based sequence annotation resource.** *Nucleic Acids Res* 2002, **30:**242-244.
39. **SMART** [http://smart.embl-heidelberg.de/]
40. Servant F, Bru C, Carrère S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: Automated clustering of homologous domains.** *Brief Bioinform* 2002, **3:**246-251.
41. **ProDom** [http://protein.toulouse.inra.fr/prodom/current/html/home.php]
42. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Barrell D, Bateman A, Binns D, Biswas M, Bradley P, Bork P, *et al.*: **The InterPro database, 2003 brings increased coverage and new features.** *Nucleic Acids Res* 2003, **31:**315-318.
43. **InterPro** [http://www.ebi.ac.uk/interpro/]
44. Haft DH, Selengut JD, White O: **The TIGRFAMs database of protein families.** *Nucleic Acids Res* 2003, **31:**371-373.
45. **TIGRFAMs** [http://www.tigr.org/TIGRFAMs/index.shtml]
46. Mewes HW, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, Münsterkötter M, Pagel P, Strack N, Stümpflen V, Warfsmann J, Ruepp A: **MIPS: analysis and annotation of proteins from whole genomes.** *Nucleic Acids Res* 2004, **32 Database issue:**D41-D44.
47. **MIPS** [http://mips.gsf.de]
48. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2003, **31:**23-27.
49. **GenBank** [http://www.ncbi.nlm.nih.gov/Genbank/]
50. Sasson O, Vaaknin A, Fleischer H, Portugaly E, Bilu Y, Linial N, Linial M: **ProtoNet: hierarchical classification of the protein space.** *Nucleic Acids Res* 2003, **31:**348-352.
51. **ProtoNet** [http://www.protonet.cs.huji.ac.il/]
52. Kriventseva EV, Fleischmann W, Zdobnov EM, Apweiler R: **CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins.** *Nucleic Acids Res* 2001, **29:**33-36.
53. **CluSTr** [http://www.ebi.ac.uk/clustr/]
54. Liu J, Rost B: **Domains, motifs and clusters in the protein universe.** *Curr Opin Chem Biol* 2003, **7:**5-11.
55. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families.** *Nucleic Acids Res* 2002, **30:**1575-1584.
56. Enright AJ, Kunin V, Ouzounis CA: **Protein families and TRIBES in genome sequence space.** *Nucleic Acids Res* 2003, **31:**4632-4638.
57. **TRIBES** [http://www.ebi.ac.uk/research/cgg/tribes/]
58. Apic G, Gough J, Teichmann SA: **An insight into domain combinations.** *Bioinformatics* 2001, **17 Suppl 1:**S83-S89.
59. Todd AE, Orengo CA, Thornton JM: **Evolution of function in protein superfamilies, from a structural perspective.** *J Mol Biol* 2001, **307:**1113-1143.
60. Soding J, Lupas AN: **More than the sum of their parts: on the evolution of proteins from peptides.** *Bioessays* 2003, **25:**837-846.
61. 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.
62. Pearl FM, Lee D, Bray JE, Buchan DW, Shepherd AJ, Orengo CA: **The CATH extended protein-family database: providing structural annotations for genome sequences.** *Protein Sci* 2002, **11:**233-244.
63. Rost B: **Did evolution leap to create the protein universe?** *Curr Opin Struct Biol* 2002, **12:**409-416.
64. Yona G, Levitt M: **Towards a complete map of the protein space based on a unified sequence and structure analysis of all known proteins.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**395-406.

65. Park J, Teichmann SA: **DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins.** *Bioinformatics* 1998, **14:**144-150.
66. Heger A, Holm L: **Exhaustive enumeration of protein domain families.** *J Mol Biol* 2003, **328:**749-767.
67. Liu J, Rost B: **Target space for structural genomics revisited.** Bioinformatics 2002, **18:**922-933.
68. Liu J, Rost B: **CHOP proteins into structural domain-like fragments.** *Proteins* 2004, **55:**678-688.
69. Liu J, Rost B: **CHOP proteins into structural domain-like fragments** [http://cubic.bioc.columbia.edu/papers/2004_chop/paper.html]
70. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402
71. Clamp M, Andrews D, Barker D, Bevan P, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, *et al.*: **Ensembl 2002: accommodating comparative genomes.** *Nucleic Acids Res* 2003, **31:**38-42.
72. Karplus K, Barrett C, Hughey R: **Hidden Markov models for detecting remote protein homologies.** *Bioinformatics* 1998, **14:**846-856.
73. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA: **The CATH database: an extended protein family resource for structural and functional genomics.** *Nucleic Acids Res* 2003, **31:**452-455.
74. Rost B, Liu J: **The PredictProtein server.** *Nucleic Acids Res* 2003, **31:**3300-3304.
75. Jones DT: **Do transmembrane protein superfolds exist?** *FEBS Lett* 1998, **423:**281-285.
76. Buchan DW, Rison SC, Bray JE, Lee D, Pearl F, Thornton JM, Orengo CA: **Gene3D: structural assignments for the biologist and bioinformaticist alike.** *Nucleic Acids Res* 2003, **31:**469-473.
77. Ranea JAG, Buchan DWA, Thornton JM, Orengo, CA: **Evolution of protein superfamilies and bacterial genome size.** *J Mol Biol* 2004, **336:**871-887.
78. Buchan DW, Shepherd AJ, Lee D, Pearl FM, Rison SC, Thornton JM, Orengo CA: **Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database.** *Genome Res* 2002, **12:**503-514.
79. Caetano-Anolles G, Caetano-Anolles D: **An evolutionarily structured universe of protein architecture.** *Genome Res* 2003, **13:**1563-1571.
80. Dokholyan NV, Borreguero JM, Buldyrev SV, Ding F, Stanley HE, Shakhnovich EI: **Identifying the importance of amino acids for protein folding from crystal structures.** *Methods Enzymol* 2003, **374:**616-638.
81. Kinch LN, Grishin NV: **Evolution of protein structures and functions.** *Curr Opin Struct Biol* 2002, **12:**400-408.
82. Harrison A, Pearl F, Mott R, Thornton J, Orengo C: **Quantifying the similarities within fold space.** *J Mol Biol* 2002, **323:**909-926.
83. Harrison A, Pearl F, Sillitoe I, Slidel T, Mott R, Thornton J, Orengo C: **Recognizing the fold of a protein structure.** *Bioinformatics* 2003, **19:**1748-1759.
84. Domingues FS, Lackner P, Andreeva A, Sippl MJ: **Structure-based evaluation of sequence comparison and fold recognition alignment accuracy.** *J Mol Biol* 2000, **297:**1003-1013.
85. Orengo CA, Sillitoe I, Reeves G, Pearl FM: **Review: what can structural classifications reveal about protein evolution?** *J Struct Biol* 2001, **134:**145-165.
86. Baker D, Sali A: **Protein structure prediction and structural genomics.** *Science* 2001, **294:**93-96.
87. Du P, Andrec M, Levy RM: **Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update.** *Protein Eng* 2003,**16:**407-414.