A structural perspective on genome evolution

David Lee*, Alastair Grant, Daniel Buchan and Christine Orengo

Protein translations of over 100 complete genomes are now available. About half of these sequences can be provided with structural annotation, thereby enabling some profound insights into protein and pathway evolution. Whereas the major domain structure families are common to all kingdoms of life, these are combined in different ways in multidomain proteins to give various domain architectures that are specific to kingdoms or individual genomes, and contribute to the diverse phenotypes observed. These data argue for more targets in structural genomics initiatives and particularly for the selection of different domain architectures to gain better insights into protein functions.

Addresses
Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, UK
*e-mail: dlee@biochem.ucl.ac.uk

Current Opinion in Structural Biology 2003, **13**:359–369

This review comes from a themed issue on
Sequences and topology
Edited by Mark Gerstein and Janet M Thornton

0959-440X/03/$ – see front matter

DOI 10.1016/S0959-440X(03)00079-4

Abbreviations
HMM hidden Markov model
PDB Protein Data Bank
PSSM position-specific scoring matrix

Introduction

The past few years have brought fascinating insights into the evolution of proteins and biological processes from the analysis of more than 100 completed genomes. This data set, which embraces all kingdoms of life, now includes seven eukaryotes, most notably human, mouse and other well-characterised model organisms, such as *Escherichia coli*, yeast and fly. Some of the most revealing evolutionary analyses exploit the fact that protein structures are more highly conserved than sequences and assign genomic sequences to structural families before performing comparative genome analyses.

However, although improving our ability to trace further back in evolution, these structure-based studies are still limited by the apparent scarcity and bias of the current structural data, as well as by limitations to the sensitivity of available sequence search algorithms. In this review, we consider these challenges, highlighting interesting new developments that improve our ability to map structural domains onto genome sequences. We also survey current levels of structural annotation of completed genomes provided by various public resources and briefly review some interesting new insights these structure-based data provide on the mechanisms of molecular evolution.

How sparse is the structural data? How many domain structure families are known and how many are there likely to be?

In assigning genome sequences to structural families, we are certainly limited by the current set of known structures in the Protein Data Bank (PDB [1]), which contains few transmembrane structures. However, although the rate of structure deposition (~18 000 entries in the PDB, March 2003) still lags significantly behind that of sequence determination (18 million sequences in GenBank, March 2003), various analyses [2,3•] propose that we now have structural representatives of most of the major domain families in nature. It is probable that, for globular proteins, this is the case as fewer than 5% of newly determined structures turn out to have a novel fold [4], despite structural genomics initiatives that explicitly target genes with no apparent structural homologues [5,6].

In fact, Coulson and Moult [2] hypothesise that 80% of sequence families in nature will belong to as few as 400 folds. They also model the existence of a few very highly populated 'superfolds' (approximately nine), which have been confirmed by recent observations that the top ten most frequently recurring folds in completed genomes currently account for nearly 40% of sequence families for known structures [4]. However, a large number (~10 000) of orphan sequences or singletons, described as unifolds, are also proposed. In support of this, in most genomes analysed to date, up to 30% of the sequences appear to be unrelated to any other sequence within their own or any other genome ([7•,8•], see also below). Some of these sequences may correspond to sequencing errors.

However, relationships to known families may have been missed because methods for detecting distant evolutionary relationships from sequence data are still limited (see below). Also, it is already clear that, in some structural families, considerable divergence can occur in paralogues [9,10•] and it can be challenging to recognise these relatives by structure comparison, let alone by methods relying solely on sequence data. The ability of some folds to support a much wider range of diverse sequences has been explored by Shakhnovich and co-workers [11••] in a rigorous statistical manner.

Table 1

Structural predictions for genomes on the World Wide Web.

Name of resource and group	Main method used	URL	Complete genomes			Ranges structurally annotated (where available)		
			E	B	A	% sequences (residues)		
						Min	Max	Ave
SUPERFAMILY; Chothia group	SAM-T99 HMMs	superfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	15	95	17	33 (24)	76 (67)	55 (48)
PEDANT; Frishman group	BLAST	pedant.gsf.de/	7	109	17	–	–	–
Gene3D; Orengo group	pfscape (SAM-T99 HMMs)	www.biochem.ucl.ac.uk/bsm/cath_new/Gene3D/	12	91	16	26 (12)	60 (50)	46 (40)
The Genomic Threading Database; Jones group	GenThreader	bioinf.cs.ucl.ac.uk/GTD/	9	71	15	29	63	46
GeneQuiz; Ouzounis group	BLAST	jura.ebi.ac.uk:8765/ext-genequiz/	8	52	11	–	–	–
3D-GENOMICS; Sternberg group	3D-PSSM (PSI-BLAST)	www.sbg.bio.ic.ac.uk/3dgenomics/	4	7	3	(30)	(49)	(41)
MODBASE; Sali group	Modeller	alto.rockefeller.edu/modbase/	7	2	–	14	71	20
FFAS; Godzik group	FFAS (PSI-BLAST)	bioinformatics.burnham-inst.org/pages/	–	3	–	–	–	–

Although there are now several domain structure classifications and neighbourhood resources ([4], [12] for a review, [13–16]), the number of folds currently identified varies from about 700 (SCOP) to 850 (CATH) because different clustering criteria are used and there is some difficulty in distinguishing folds in some more continuous regions of fold space [17]. The number of structural superfamilies, wherein domains are clearly evolutionarily related, appears to be more consistent at approximately 1800 (± 50).

Current mapping of these structural data onto genome sequences assigns up to three-quarters of genes or partial genes in completed genomes to known structural families ([18,19], see also below); this shrinks by about 10–20% when calculated on a per residue basis. A further 20–30% are probably membrane-associated proteins, for which there are currently few representatives in the PDB. From clustering studies of completed genomes ([7•,20,21], see also below and Table 1), it appears that up to 20–30% of the remaining sequences in each genome belong to genome-specific families or singletons for which no current structural data are available (see Figure 1). Although these may be distant undetected relatives of known structural families, they could alternatively be completely novel structures. Intriguingly, a significant proportion of these are predicted to have low secondary structure content and are probably disordered [7•,22], but may have an important role in regulation.

Encouragingly, Figure 2 shows that many of the known structural families are common to more than one genome, which means that we can start to decipher the mechanisms by which domains have been duplicated and combined to create new protein functions and processes both within and between the genomes (see below).

The advent of structural genomics initiatives brings hope that, by carefully targeting sequence families for which no structural relatives can be detected and coordinating structure determination, we will expand the repertoire of known folds further [5,6,23]. Vitkup *et al.* [24] propose a strategy for the efficient selection of approximately 16 000 domain targets to ensure that most genome sequences are sufficiently close to structural homologues (>25% sequence identity) to allow reliable homology modelling.

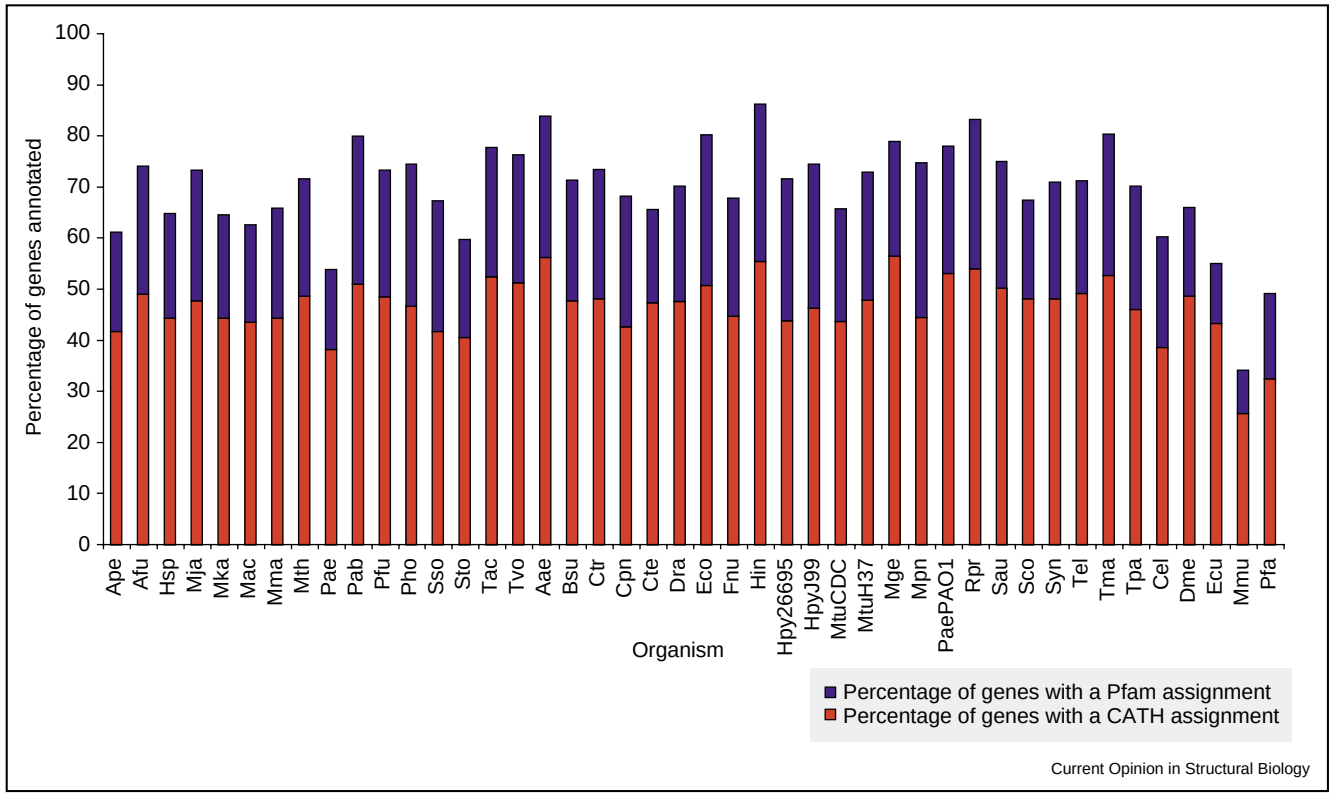
However, the increasing number of orphans detected with the release of each new completely sequenced genome may mean this is an optimistic estimate. Similarly, detailed analyses of enzyme families have suggested that, below 60% identity, the function of paralogues can diverge considerably [25•,26–28] and so considerably more targets may be required to provide a structural rationale for functional modification in some functionally promiscuous protein families.

Exploiting structural data to improve recognition of domains

In addition to detecting more ancient links between sequence families, structural data enable the more reliable identification of domain boundaries in genes. Mapping of structural domains onto whole gene sequences therefore allows us to explore more accurately the domain architecture of genes and how they have evolved by domain duplication and recombination with different domain partners. More importantly, this then allows us to understand how changes in domain architecture modulate function.

It has long been hypothesised that domains are important evolutionary units, speculation that is supported by

Figure 1



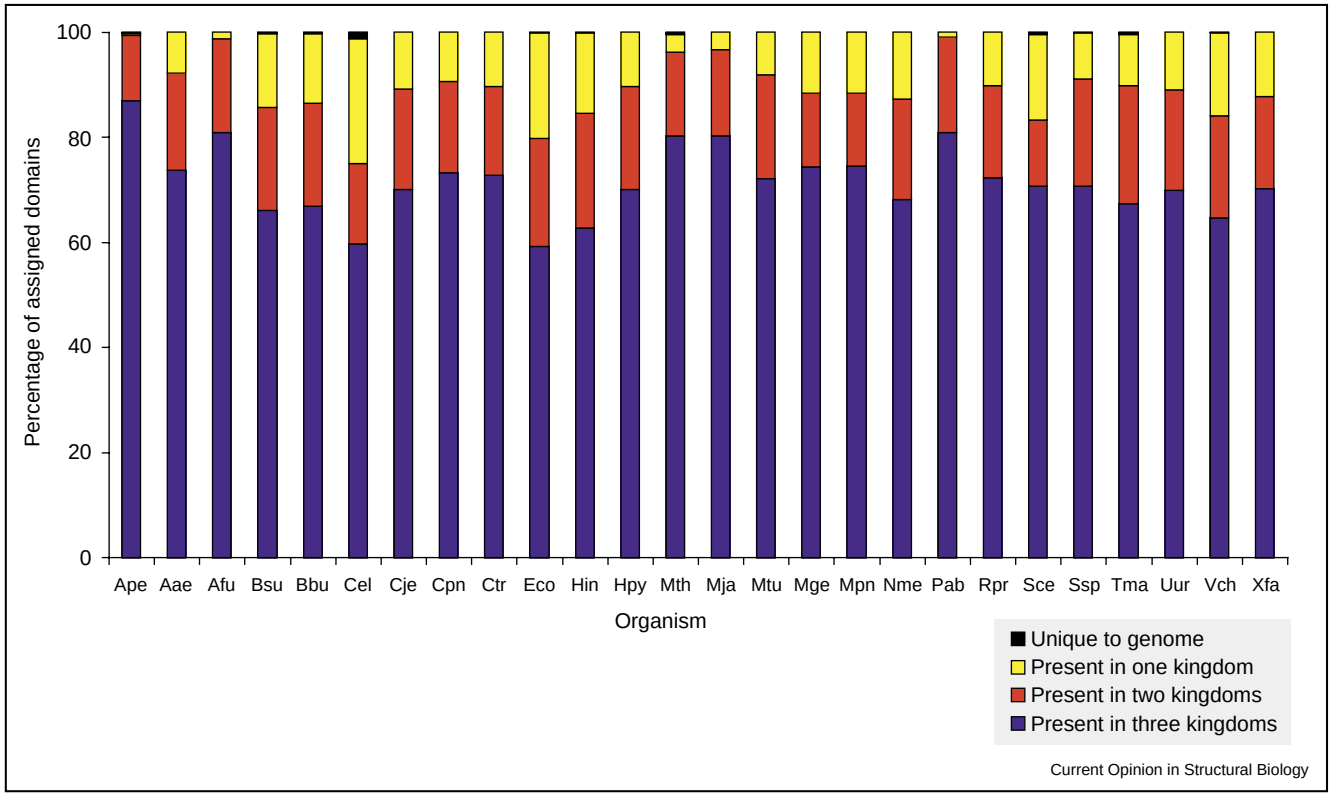
Domain assignments by source — the source of the domain assignments within a genome, as a percentage of the total number of genes. Columns represent domain assignments by CATH and some Pfam (red), and Pfam only (blue). The organism names have been abbreviated to a three-letter code as follows: Ape, *Aeropyrum pernix*; Afu, *Archeoglobus fulgidus*; Hsp, *Halobacterium* sp NRC-1; Mja, *Methanococcus jannaschii*; Mka, *Methanopyrus kandleri* AV19; Mac, *Mathanosarcina acetivorans*; Mma, *Methanosarcina mazei*; Mth, *Methanobacterium thermoautotrophicum*; Pae, *Pyrobaculum aerophilum*; Pab, *Pyrococcus abyssi*; Pfu, *Pyrococcus furiosus*; Pho, *Pyrococcus horikoshii*; Sso, *Sulfolobus solfataricus*; Sto, *Sulfolobus tokodaii*; Tac, *Thermoplasma acidophilum*; Tvo, *Thermoplasma volcanium*; Aae, *Aquifex aeolicus*; Bsu, *Bacillus subtilis*; Ctr, *Chlamydia trachomatis*; Cpn, *Chlamydophila pneumoniae* CWL029; Cte, *Chlorobium tepidum* TLS; Dra, *Deinococcus radiodurans*; Eco, *Escherichia coli* K12; Fnu, *Fusobacterium nucleatum*; Hin, *Haemophilus influenzae* Rd; Hpy26695, *Helicobacter pylori* 26695; HpyJ99, *Helicobacter pylori* J99; MtuCDC, *Mycobacterium tuberculosis* CDC1551; MtuH37, *Mycobacterium tuberculosis* H37; Mge, *Mycoplasma genitalium*; Mpn, *Mycoplasma pneumoniae*; PaePAO1, *Pseudomonas aeruginosa* PAO1; Rpr, *Rickettssia prowazekii*; Sau, *Staphylococcus aureus* N315; Sco, *Streptomyces coelicolor* A3(2); Syn, *Synechocystis* sp PCC 6803; Tel, *Thermosynechococcus elongates*; Tma, *Thermotoga maritima*; Tpa, *Treponema pallidum*; Cel, *Caenorhabditis elegans*; Dme, *Drosophila melanogaster*; Ecu, *Encephalitozoon cuniculi*; Mmu, *Mus musculus*; Pfa, *Plasmodium falciparum* 3D7.

analyses of the completed genome data, which suggest that at least 60% of genes, possibly as high as 80% in eukaryotes, are multidomain proteins ([7•,29••,30]). Domain duplications and recombinations are thought to have occurred extensively. For the limited data set of known protein structures in the PDB, of which only about one-third are multidomain, recent analysis of CATH domains demonstrated that the majority (~85%) recur in different multidomain contexts [4].

Automatic recognition of domains in multidomain proteins can be very difficult, even using structural data, although some interesting new approaches promise much greater accuracy [31]. From a sequence perspective, several resources attempt to locate domains by exploiting domain recurrence and clustering sequences from completed genomes and large sequence databases [20,21,32]. Some of these approaches explicitly combine sequence-based predictions with structural data, which might be expected to help in 'bootstrapping' domain assignments.

Although these resources attempt to identify domain-based families in genomes, the number of families they describe varies considerably from 10 000 [20] to 77 000 [31], with up to 170 000 singletons. The recent clustering of 119 genomes for the Gene3D resource [33•] found approximately 50 000 protein families with potentially unique domain architectures. These comprise different combinations of 1400 structural domains from CATH and 5000 gene families from Pfam (which may contain more than one structural domain) (see Figure 3). There are also ~150 000 singletons unassigned to any family.

Figure 2



Domain assignments by kingdom — the distribution of domain assignments across the three kingdoms of life, as a percentage of the total domain assignments for each genome. Columns represent domains present in all three kingdoms (blue), two out of three kingdoms (red), one kingdom (yellow) or unique to the genome (black). The organism names have been abbreviated to a three-letter code as follows: Ape, *Aeropyrum pernix*; Aae, *Aquifex aeolicus*; Afu, *Archeoglobus fulgidus*; Bsu, *Bacillus subtilis*; Bbu, *Borrelia burgdorferi*; Cel, *Caenorhabditis elegans*; Cje, *Campylobacter jejuni*; Cpn, *Chlamydia pneumoniae*; Ctr, *Chlamydia trachomatis*; Eco, *Escherichia coli*; Hin, *Haemophilus influenzae*; Hpy, *Helicobacter pylori*; Mth, *Methanobacterium thermoautotrophicum*; Mja, *Methanococcus jannaschii*; Mtu, *Mycobacterium tuberculosis*; Mge, *Mycoplasma genitalium*; Mpn, *Mycoplasma pneumoniae*; Nme, *Neisseria meningitidis*; Pab, *Pyrococcus abyssi*; Rpr, *Rickettssia prowazekii*; Sce, *Saccharomyces cerevisiae*; Ssp, *Synechocystis* sp; Tma, *Thermotoga maritima*; Uur, *Ureaplasma urealyticum*; Vch, *Vibrio cholerae*; Xfa, *Xylella fastidiosa*.

Exploiting structural data to detect evolutionary relationships between domains

Sensitive structure comparison algorithms capable of detecting very distant relatives were first developed in the late 1980s and many methods now exist [34], most exploiting common methods and features. Perhaps the most significant recent developments have been improvements to the protocols measuring similarities and assessing statistical significance. Although many resources have long exploited z-scores for highlighting the most significant matches, recent attempts have provided better statistical models of the extreme value distributions that are returned by database scans [17,35].

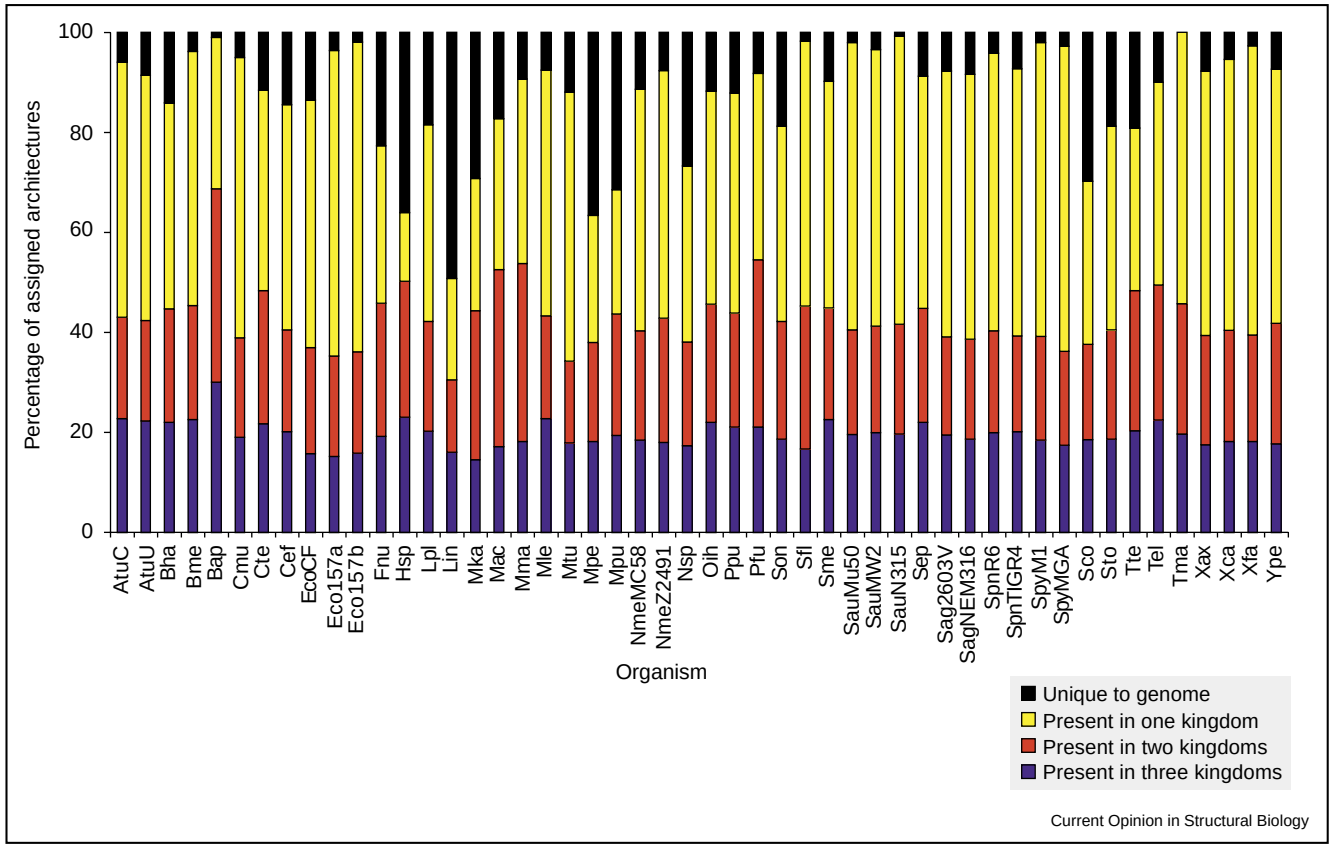
However, structural similarity is not sufficient to guarantee common ancestry, as there may be physical constraints that limit the number of energetically acceptable folds. Evolutionary relationships are usually confirmed by the existence of shared sequence motifs or evidence of functional similarity (see the review by Jackson, Westhead and co-workers in this issue). Holm and co-workers [36] successfully exploit neural networks that combine multiple data (e.g. structure similarity, sequence similarity, matching of functional key words) to perform large-scale automatic identification of homologues in the Dali domain database.

What sequence-based methods are used to assign genome sequences to structural families and how well do these perform?

Although many small bacterial genomes comprise fewer than 5000 genes, larger eukaryotic genomes contain up to 35 000 genes, as in the case of *Takifugu rubripes*. Therefore, methods for scanning genes against structural families need to be sensitive, selective and also very fast.

To evaluate these approaches, several groups have used structural data sets derived from the SCOP and CATH domain classifications to test the sensitivity and selectivity of the methods [18,37–39,40•]. Because structure is so

Figure 3



Domain architectures by kingdom — the distribution of domain architectures across the three kingdoms of life. Columns represent domains present in all three kingdoms (blue), two out of three kingdoms (red), one kingdom (yellow) or unique to the genome (black). The organism names have been abbreviated to a three-letter code as follows: AtuC, *Agrobacterium tumefaciens* str C58 Cel; AtuU, *Agrobacterium tumefaciens* str C58 UW; Bha, *Bacillus halodurans*; Bme, *Brucella melitensis*; Bap, *Buchnera aphidocola*; Cmu, *Chlamydia muridarum*; Cte, *Clostridium tetani* E88; Cef, *Corynebacterium efficiens* YS-314; EcoCF, *Escherichia coli* CFT073; Eco157a, *Escherichia coli* O157a; Eco157b, *Escherichia coli* O157b; Fnu, *Fusobacterium nucleatum*; Hsp, *Halobacterium* sp; Lpl, *Lactobacillus plantarum*; Lin, *Leptospira interrogans*; Mka, *Methanopyrus kandleri*; Mac, *Mathanosarcina acetivorans*; Mma, *Methanosarcina mazei*; Mle, *Mycobacterium leprae*; Mtu, *Mycobacterium tuberculosis* CDC 1551; Mpe, *Mycoplasma penetrans*; Mpu, *Mycoplasma pulmonis*; NmeMC58, *Neisseria meningitidis* MC58; NmeZ2491, *Neisseria meningitidis* Z2491; Nsp, *Nostoc* sp; Oih, *Oceanobacillus iheyensis*; Ppu, *Pseudomonas putida*; Pfu, *Pyrococcus furiosus*; Son, *Shewanella oneidensis*; Sfl, *Shigella flexneri*; Sme, *Sinorhizobium meliloti*; SauMu50, *Staphylococcus aureus* Mu50; SauMW2, *Staphylococcus aureus* MW2; SauN315, *Staphylococcus aureus* N315; Sep, *Staphylococcus epidermidis*; Sag2603V, *Streptococcus agalactiae* 2603V/R; SagNEM316, *Streptococcus agalactiae* NEM316; SpnR6, *Streptococcus pneumoniae* R6; SpnTIGR4, *Streptococcus pneumoniae* TIGR4; SpyM1, *Streptococcus pyogenes* M1; SpyMGA, *Streptococcus pyogenes* MGA; Sco, *Streptomyces coelicolor*; Sto, *Sulfolobus tokodaii*; Tte, *Thermoanaerobacter tengcongensis*; Tel, *Thermosynechococcus elongates*; Tma, *Thermotoga maritima*; Xax, *Xanthomonas axonopodis*; Xca, *Xanthomonas campestris*; Xfa, *Xylella fastidiosa*; Ype, *Yersinia pestis*.

well conserved, these classifications provide data sets of validated structural relatives that are sufficiently distant (e.g. <25% sequence identity) to present a real challenge for sequence-based methods.

Receiver operator curves are usually generated contrasting the selectivity and sensitivity of various methods. These compare the increasing coverage (or number of remote homologues identified) of different sequence search methods with increasing numbers of errors (e.g. 50 false positives in the case of ROC50 curves [41]). Recent implementations of these benchmarking protocols take account of the bias of current classifications towards some structural families by ensuring that the data set contains a single pair of relatives from each family [42].

Although the fastest sequence search methods available (the most widely used of which is the BLAST suite of methods) are based on pairwise alignment, they are not as powerful as 1D-profile-based methods, such as PSI-BLAST [41], or methods employing hidden Markov models (HMMs) [18,43]. These owe their success to their ability to capture information on residue propensities at different positions in the protein. Propensities are derived from statistical analysis of a multiple sequence alignment

containing divergent sequence relatives from the protein family. For example, in a 1D-profile, propensities are frequently encoded in a position-specific scoring matrix (PSSM), which typically contrasts the frequency of amino acid residues at a given position in the alignment with the frequency expected by chance.

Recent developments in this field, and currently among the most popular methods for genome annotation, are powerful iterated search programs such as PSI-BLAST [41], SAM-T99 and SAM-T2K [38,43], which build their own multiple alignment by iteratively searching the database for ever more remote homologues, adjusting the PSSM or state model with each iteration. These programs have been extensively benchmarked using SCOP data sets [37,38]. Currently, about 50–60% of very distant relatives (<25% sequence identity) can be recognised using these approaches [18].

Related algorithms, such as RPS-BLAST [16], have recently emerged that significantly reduce the scanning time for large databases and genomes. By contrast, increased sensitivity at the cost of speed has been achieved by developing profile-profile protocols that compare libraries of profiles to detect evolutionary links [42]. These have also been shown to improve the quality of the alignment, which is important for assessing any residue conservation that may be indicative of functional similarity. Other improvements in the performance of profile-based methods, to date, have been achieved by combining structural data or predicted structural data regarding secondary structure conformation, accessibility or residue contacts [44–46].

Generally, although fold recognition methods such as threading and related approaches have been shown to be more sensitive at recognising remote homologues [47], they are often slower and less amenable to large-scale genome annotation. However, several have been applied to provide annotations particularly for smaller bacterial genomes ([22]; see Table 1). A recent critical assessment of structure prediction (CASP4) in the US demonstrated the value of combining the results from several prediction methods [48], although this is another strategy that is not yet very amenable to whole genome annotation.

Similar problems exist for large-scale homology modelling, another computer-intensive procedure. Although some whole genome scale modelling has been performed [49–51], again this is largely restricted to small bacterial genomes. Of course, the models they provide will be valuable for analysing close relatives in other genomes. As discussed above, an important issue in providing high-quality models for inferring functional properties or predicting ligand binding is the accuracy of the alignment. Threading algorithms often give the best alignments, although some profile and HMM approaches also perform well [22,40•].

It is clear that recent web-based initiatives (such as ENSEMBL [52] and InterPro [53], which display protein family predictions from multiple sources) will be important resources for genome annotation in the future. Both ENSEMBL and InterPro have recently been funded to include domain structure annotations from SCOP and CATH in future versions, displayed using the DAS technology [54].

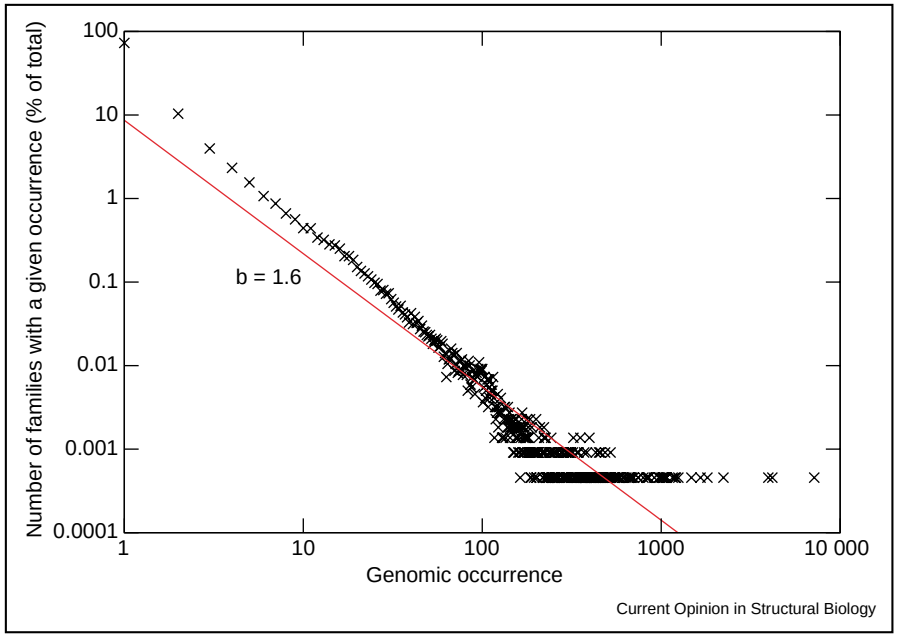
What proportion of genome sequences can be assigned to structural families and how are these families distributed within genomes and across kingdoms?

Table 1 lists some public resources displaying genome annotations and summarises recent coverage of representative genomes in each kingdom. The variation in coverage reflects differences in the methods used to map structural domains onto the genome sequences, the sequence databases used in building the iterated profiles or HMMs, and the structural family classifications and fold libraries employed. Nevertheless, it is very encouraging to see that the SUPERFAMILY resource [18,55••] assigns nearly 50% of the genes (on a per residue basis) of many genomes to known structural families.

Several structural classification resources now provide profiles or HMMs of representative structures for mapping onto completed genomes, together with information on the proportions of genomes that can be annotated in this way (Table 1). For one of the most comprehensive resources, the SUPERFAMILY database, built from SCOP families, genome coverage ranges from 30 to 76% (on a per gene basis) [18] for more than 100 completed genomes. The current statistics for the equivalent Gene3D resource [33•], based on CATH structural families, are shown in Figures 1 and 2, and Table 1. The additional coverage gained by using SAMT-99 models of Pfam families is also plotted in Figure 1 and shows that a significant percentage of genes or partial genes (up to 85%) can now be assigned to a structure- or sequence-based family, from which some functional information can be inferred.

Several groups have published analyses describing and modelling the distributions of structural families within and between genomes [3•,29••,30,33•,56,57•,58–60]. The uneven population of domain families, first noted in the structural classifications, is mirrored in the genomes, whereby most families occur only a few times within a given genome but a few families recur extensively (see Figure 4). This phenomenon, which indicates a power law relationship, was first commented on by Huynen [56] and has been fitted to a range of power law functions [57•,58,59]. It suggests a model whereby the 'fit' get

Figure 4



Power law behaviour is observed for the TribeMCL [71] clustering (inflation value = 3) of 119 complete genomes into families. These families correlate strongly with protein domain architecture. For power law behaviour, the number of families (N) with a given occurrence (F) decays according to the equation $N = aF^{b}$. This distribution has a linear appearance when plotted on double-logarithmic axes, where $-b$ describes the slope. The best-fit power law function is displayed and the value b = 1.6 is typical of the power law behaviour observed for the occurrence of families, superfamilies and folds in genomes.

'fitter' and domains duplicated early in evolution will increasingly dominate the population. Including selectionist pressure in this model would then favour the retention of duplicated domains that perform important biochemical activities.

Wolf *et al.* [58] recently used the Pareto function to simulate a model that captures the birth (gene duplication), death (gene loss) and innovation (new protein) of different domains. However, this entirely stochastic model fails to account completely for the observed distribution, although it comes close. The absence of any selection pressure seems a particular failure, as many of the most recurrent domains have important generic functions (e.g. in providing energy or redox equivalents for catalysis, or in responding to cellular signals and binding to DNA [29••,30]; see Figure 5).

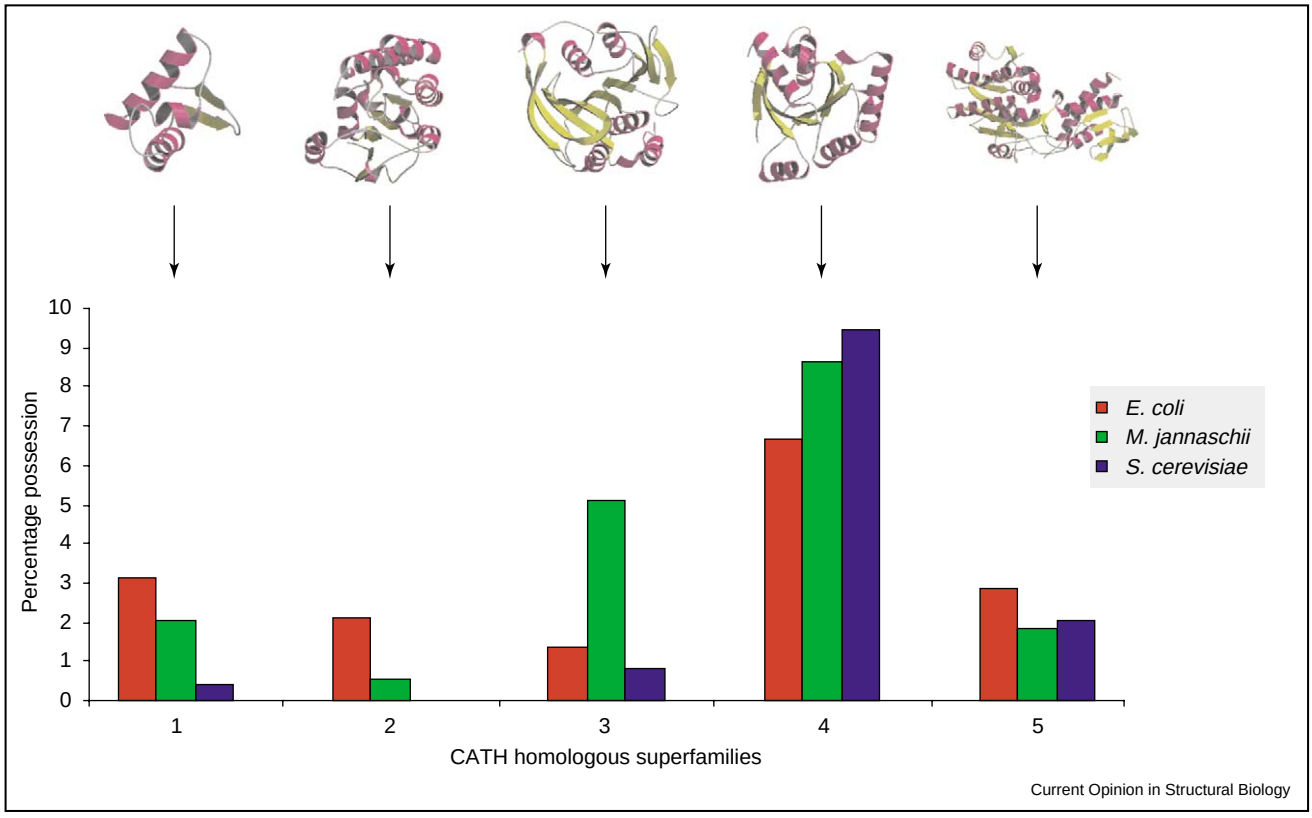
As many of the genome sequences are multidomain proteins and as the function of a protein is usually determined by both the nature of these domains and their assembly in three dimensions, it is also important to consider the distribution of specific domain architectures across genomes. Interestingly, Apic *et al.* [29••] have shown that another power law exists in the pairing of domains — most domain families are partnered with only a few other domains, whereas others are very promiscuous, combining with many different partners (see further discussion below).

What do we learn about the evolution of protein functions and processes by assigning genomes to structural families?

Domain structures can be viewed as a parts list for biology [59], but to really understand how the diverse phenotypes have evolved, we must understand the interactions of these parts and how they are assembled into complex functional units, pathways and signalling processes. This is complicated by the fact that the functional characteristics of these domain modules or 'parts' can sometimes vary considerably between relatives, particularly between paralogous domains [25•,26,27], and relatively high levels of sequence and structural similarity are required to confidently transfer functional properties between relatives in some protein families [25•,26–28,60,61•,62•].

Interestingly, even in enzyme families where functions appear to vary considerably, these changes are more frequently associated with substrate specificity than with fundamental changes in the chemistry of the reaction [25•]. Similar chemical intermediates are often detected. Although catalytic residues enabling the reactions may be contributed by different positions in the protein

Figure 5



The percentage of folds possessed by three model organisms that belong to the five most commonly assigned domains. The different domain types are numbered: domain 1, 'winged helix' repressor DNA-binding domain; domain 2, periplasmic binding protein like-II; domain 3, DNA methylase coenzyme binding domain; domain 4, P-loop-containing nucleotide triphosphate hydrolases; domain 5, NAD(P)-binding Rossmann fold domains.

sequence, they are usually co-located in three dimensions [63]. Structural analyses of functionally variable families reveal that specificities can be modulated by significant structural embellishments in the region of the active site, by changes in the domain architecture and by variations in the oligomerisation states of relatives [25•].

Very frequently, domain function is modulated by changes in a domain's partners. In this context, the elegant analyses of Teichmann and co-workers [29••,30], using SCOP-based annotations, reveal the manner in which the limited repertoire of domain families has been duplicated and combined in many different ways through recombination, fusion and fission. Although there are currently more than 12 000 different domain architectures identified, only a small percentage of these are common to all kingdoms and the phenotypic diversity exhibited across kingdoms correlates with variations in domain architecture. On average, between 60 and 70% of domain architectures identified within a particular genome are unique to the kingdom (see Figure 3). However, these are predominantly assembled from domains that are common across kingdoms [29••]. Only one-third of domain families do not appear to combine with any other domain [29••].

There also appear to be some highly recurrent domain pairs or triplets common across kingdoms, again assembled from the common domain families [29••]. Interestingly, most domain pairs are only seen in one orientation to each other [64•]. The evolution of an interface and function is too costly to evolve twice in two orientations and it is simpler to copy and modify the domain pair in one orientation.

In other words, recombination between common domains has been a major factor in the evolution of kingdom-specific and species-specific functions. These common domains, which recur frequently in genomes, also tend to have multiple domain partners; Teichmann *et al.* propose that they may therefore be more ancient, allowing more evolutionary time for recombination. Amongst the most common domains are the P-loop hydrolases (see Figure 5), which provide energy for motion and reactions by hydrolysing ATP and GTP. Rossmann folds, also common, provide oxidising or reducing energy by oxidation or reduction of

NAD(P). In metazoa, families involved in signal transduction are amongst the most versatile, as transcription/translation is tightly regulated by nucleic-acid-binding domains, which combine with other domains responsible for the specificity of regulation [29••] (see also below).

Only a small fraction of multidomains comprise tandem repeats. Multicellular eukaryotes tend to have much longer repeats than unicellular and most of these either have structural roles or are involved in cell adhesion, or complex signalling and regulatory mechanisms [29••].

Structural annotation of the genomes of some experimentally well-characterised model organisms, most notably *E. coli* and yeast, has also enabled the analysis of pathway evolution [65•,66–68,69•]. The ubiquitous power law prevalent throughout biology is also apparent in pathway evolution. Most families occur on a few related pathways [67,68], whereas a few families occur on many pathways. Again, these are domains supplying generic functions (e.g. ATP binding or provision of redox equivalents for catalysis).

Rison *et al.* [65•] comment that, overall, the data suggest that several pathway evolution mechanisms may occur in concert, although the 'Jensen' patchwork model of evolution is favoured. In this, enzymes are largely recruited to a pathway for the specific chemistry they perform. A very limited 'Horowitz' model is also apparent, involving some serial recruitment of homologues along a pathway. This is supported by Alves and co-workers' analysis [68], in which they reviewed 12 genomes and treated pathways as connected graphs with networks built around metabolites.

For some enzymes, such as transferases and synthetases, Alves and co-workers [68] found quite a high proportion (~60%) of homologues within two steps of each other in the pathway network. Interestingly, in all organisms, there is an association between similar enzyme chemistry and proximity in the network, regardless of homology. Duplicated enzymes are more likely to be added locally in the network, presumably because that will cause less disruption — assuming that the new enzyme has a similar function to the old [68].

Detailed comparison of small-molecule metabolism between yeast and *E. coli* by Jardine *et al.* [69•] revealed that 50–60% of the enzymes are common between the organisms and over 80% of the pathways are shared. Two-thirds of these common enzymes have the same domain architectures, despite at least one billion years of evolutionary separation.

The extent of domain duplication and combination in signalling pathways is revealed by a recent analysis of transcription factors in *E. coli* [70]. This showed that three-quarters have arisen by gene duplication, comprising mostly two-domain proteins with a common DNA-binding domain (of which there are 11 families) combined with a distinct regulatory domain that frequently binds small molecules [70].

Conclusions

In summary, the mapping of structures to completed genome sequences is gradually revealing the intricate mechanisms by which diverse pathways and phenotypes can evolve from an apparently limited repertoire of domain modules. Structural genomics initiatives will help expand our library of these modules. In combination with improvements in homologue detection methods, these data will assist genome analyses not only by revealing more ancient links but also by elucidating domain recombination and the manner by which it has contributed to the glorious panoply of phenotypes in nature.

Update

Since the submission of this review, additional highly relevant papers have been published [72•,73].

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. Westbrook J, Feng Z, Chen L, Yang H, Berman H: **The Protein Data Bank and structural genomics**. *Nucleic Acids Res* 2003, **31**:489-491.

2. Coulson A, Moult J: **A unifold, mesofold, and superfold model of protein fold use**. *Proteins* 2002, **46**:61-71.

3. • Koonin E, Wolf Y, Karev G: **The structure of the protein universe and genome evolution**. *Nature* 2002, **420**:218-223.

The authors review the state of current research in genome evolution. The presence of power law relationships within genome components is discussed. They also present a comprehensive discussion of the potential evolutionary forces driving the emergence of such relationships.

4. Pearl F, Bennett C, Bray J, Harrison A, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo C: **The CATH database: an extended protein family resource for structural and functional genomics**. *Nucleic Acids Res* 2003, **31**:452-455.

5. Sali A: **Target practice**. *Nat Struct Biol* 2001, **8**:482-484.

6. Baker D, Sali A: **Protein structure prediction and structural genomics**. *Science* 2001, **294**:93-96.

7. • Rost B: **Did evolution leap to create the protein universe?** *Curr Opin Struct Biol* 2002, **12**:409-416.

A review of the observed protein structure content of complete genomes. Some marked differences between eukaryotes and prokaryotes are noted with respect to the nature of the folds observed. Rost notes that many proteins do not have close homologues within other genomes and may be orphans.

8. • Liu J, Rost B: **Comparing function and structure between entire proteomes**. *Protein Sci* 2001, **10**:1970-1999.

Liu and Rost survey the state of the art in proteome annotation, focusing on broad functional features of proteins. They observe that, unlike bacteria and archaea, eukaryotic proteins are frequently multidomain and eukaryotes have twice as many coiled-coil proteins. They note that nearly a third of the observed proteins have transmembrane helices and up to a quarter of proteins are secreted. They estimate that there are likely to be between 1200 and 2600 protein folds.

9. Orengo C, Sillitoe I, Reeves G, Pearl F: **Review: what can structural classifications reveal about protein evolution?** *J Struct Biol* 2001, **134**:145-165.

10. Grishin N: **Review: Fold change in evolution of protein**
• **structures**. *J Struct Biol* 2001, **134**:167-185.
An interesting review discussing the similarities between different fold groups and speculating on evolutionary mechanisms giving rise to fold changes.

11. Dokholyan N, Shakhnovich E: **Understanding hierarchical**
•• **protein evolution from first principles**. *J Mol Biol* 2001, **312**:289-307.
Rigorous mathematical analysis of the relationships between protein families sharing common folds, proposing an elegant model for their evolution.

12. Pearl F, Orengo C: **Protein structure classifications.** In *Bioinformatics: Genes, Proteins and, Computers.* Edited by Orengo CA, Jones DT, Thornton JM. Abingdon, UK: Bios; 2003:103-111.

13. Lo Conte L, Brenner S, Hubbard T, Chothia C, Murzin A: **SCOP database in 2002: refinements accommodate structural genomics**. *Nucleic Acids Res* 2002, **30**:264-267.

14. de Bakker PI, Bateman A, Burke D, Miguel R, Mizuguchi K, Shi J, Shirai H, Blundell T: **HOMSTRAD: adding sequence information to structure-based alignments of homologous protein families**. *Bioinformatics* 2001, **17**:748-749.

15. Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L: **A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3**. *Nucleic Acids Res* 2001, **29**:55-57.

16. Marchler-Bauer A, Anderson J, DeWeese-Scott C, Fedorova N, Geer L, He S, Hurwitz D, Jackson J, Jacobs A, Lanczycki C *et al.*: **CDD: a curated Entrez database of conserved domain alignments**. *Nucleic Acids Res* 2003, **31**:383-387.

17. Harrison A, Pearl F, Mott R, Thornton J, Orengo C: **Quantifying the similarities within fold space**. *J Mol Biol* 2002, **323**:909-926.

18. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure**. *J Mol Biol* 2001, **313**:903-919.

19. Pawlowski K, Rychlewski L, Zhang B, Godzik A: **Fold predictions for bacterial genomes**. *J Struct Biol* 2001, **134**:219-231.

20. Krause A, Stoye J, Vingron M: **The SYSTERS protein sequence cluster set**. *Nucleic Acids Res* 2000, **28**:270-272.

21. Yona G, Linial N, Linial M: **ProtoMap: automatic classification of protein sequences and hierarchy of protein families**. *Nucleic Acids Res* 2000, **28**:49-55.

22. Pawlowski K, Rychlewski L, Zhang B, Godzik A: **Fold predictions for bacterial genomes**. *J Struct Biol* 2001, **134**:219-231.

23. Liu J, Rost B: **Target space for structural genomics revisited**. *Bioinformatics* 2002, **18**:922-933.

24. Vitkup D, Melamud E, Moult J, Sander C: **Completeness in structural genomics**. *Nat Struct Biol* 2001, **8**:559-566.

25. Todd A, Orengo C, Thornton J: **Evolution of function in protein**
• **superfamilies, from a structural perspective**. *J Mol Biol* 2001, **307**:1113-1143.
Comprehensive analysis of the mechanisms by which biochemical functions have evolved in enzyme families with structural representatives.

26. Devos D, Valencia A: **Practical limits of function prediction**. *Proteins* 2000, **41**:98-107.

27. Wilson C, Kreychman J, Gerstein M: **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores**. *J Mol Biol* 2000, **297**:233-249.

28. Todd A, Orengo C, Thornton J: **Sequence and structural differences between enzyme and nonenzyme homologs**. *Structure* 2002, **10**:1435-1451.

29. Apic G, Gough J, Teichmann S: **Domain combinations in**
•• **archaeal, eubacterial and eukaryotic proteomes**. *J Mol Biol* 2001, **310**:311-325.
Analysis of the SCOP database. The authors observe power law relationships in the networks of domain partners. They show that many domains do not have domain partners and those that do often only have partners from one or two families. They show that some domains are highly promiscuous and have many partners (e.g. the P-loop nucleotide triphosphate hydrolases).

30. Teichmann S, Murzin A, Chothia C: **Determination of protein function, evolution and interactions by structural genomics**. *Curr Opin Struct Biol* 2001, **11**:354-363.

31. Guo J, Xu D, Kim D, Xu Y: **Improving the performance of DomainParser for structural domain partition using neural network**. *Nucleic Acids Res* 2003, **31**:944-952.

32. Servant F, Bru C, Carrere S, Courcelle E, Gouzy J, Peyruc D, Kahn D: **ProDom: automated clustering of homologous domains**. *Brief Bioinform* 2002, **3**:246-251.

33. Buchan D, Shepherd A, Lee D, Pearl F, Rison S, Thornton J,
• Orengo C: **Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database**. *Genome Res* 2002, **12**:503-514.
The authors present a database of structural assignments to genes of unknown structure within complete genomes. Structural assignments are based on sequence identity using PSI-BLAST. The database currently contains data for 66 genomes.

34. Sillitoe I, Orengo C: **Protein structure comparison**. In *Bioinformatics: Genes, Proteins and Computers*. Edited by Orengo CA, Jones DT, Thornton JM. Abingdon, UK: Bios; 2003:81-102.

35. Levitt M, Gerstein M: **A unified statistical framework for sequence comparison and structure comparison**. *Proc Natl Acad Sci USA* 1998, **95**:5913-5920.

36. Dietmann S, Holm L: **Identification of homology in protein structure classification**. *Nat Struct Biol* 2001, **8**:953-957.

37. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, Chothia C: **Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods**. *J Mol Biol* 1998, **284**:1201-1210.

38. Madera M, Gough J: **A comparison of profile hidden Markov model procedures for remote homology detection**. *Nucleic Acids Res* 2002, **30**:4321-4328.

39. Pearl F, Orengo C, Pearl F, Lee D, Bray J, Sillitoe I, Todd A, Harrison A, Thornton J, Orengo C: **Assigning genomic sequences to CATH**. *Nucleic Acids Res* 2000, **28**:277-282.

40. Lindahl E, Elofsson A: **Identification of related proteins on family,**
• **superfamily and fold level**. *J Mol Biol* 2000, **295**:613-625.
The authors compare the performance of the predominant sequence searching methods for use in fold, family and superfamily recognition. PSI-BLAST, HMM and threading methods are compared.

41. Schaffer A, Aravind L, Madden T, Shavirin S, Spouge J, Wolf Y, Koonin E, Altschul S: **Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements**. *Nucleic Acids Res* 2001, **29**:2994-3005.

42. Yona G, Levitt M: **Within the twilight zone: a sensitive profile-profile comparison tool based on information theory**. *J Mol Biol* 2002, **315**:1257-1275.

43. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R: **What is the value added by human intervention in protein structure prediction?** *Proteins* 2001, **(suppl 5)**:86-91.

44. Hargbo J, Elofsson A: **Hidden Markov models that use predicted secondary structures for fold recognition**. *Proteins* 1999, **36**:68-76.

45. Panchenko A, Marchler-Bauer A, Bryant S: **Combination of threading potentials and sequence profiles improves fold recognition**. *J Mol Biol* 2000, **296**:1319-1331.

46. Rychlewski L, Jaroszewski L, Li W, Godzik A: **Comparison of sequence profiles. Strategies for structural predictions using sequence information**. *Protein Sci* 2000, **9**:232-241.

47. Jones DT: **Protein structure prediction**. In *Bioinformatics: Genes, Proteins and Computers.* Edited by Orengo CA, Jones DT, Thornton JM. Abingdon, UK: Bios; 2003:135-150.

48. Sippl M, Lackner P, Domingues F, Prlic A, Malik R, Andreeva A, Wiederstein M: **Assessment of the CASP4 fold recognition category**. *Proteins* 2001, **(suppl 5)**:55-67.

49. Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom M, Madhusudhan M, Mirkovic N, Sali A: **Protein structure modeling for structural genomics**. *Nat Struct Biol* 2000, **7(suppl)**:986-990.

50. Marti-Renom M, Stuart A, Fiser A, Sanchez R, Melo F, Sali A: **Comparative protein structure modeling of genes and genomes**. *Annu Rev Biophys Biomol Struct* 2000, **29**:291-325.

51. Guex N, Diemand A, Peitsch M: **Protein modelling for all**. *Trends Biochem Sci* 1999, **24**:364-367.

52. Brooksbank C, Camon E, Harris M, Magrane M, Martin M, Mulder N, O'Donovan C, Parkinson H, Tuli M, Apweiler R *et al.*: **The European Bioinformatics Institute's data resources**. *Nucleic Acids Res* 2003, **31**:43-50.

53. Mulder N, Apweiler R, Attwood T, 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.

54. Hubbard T: **Biological information: making it accessible and integrated (and trying to make sense of it)**. *Bioinformatics* 2002, **18(suppl 2)**:S140.

55. •• Gough J, Chothia C: **SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments**. *Nucleic Acids Res* 2002, **30**:268-272.
SUPERFAMILY is a database of HMM models of the SCOP structural domain families. Many of the SAMT98 HMM models are refined manually to generate highly sensitive and accurate models. Structural assignments to genes and genomes are available from the database.

56. Huynen M, van Nimwegen E: **The frequency distribution of gene family sizes in complete genomes**. *Mol Biol Evol* 1998, **15**:583-589.

57. • Qian J, Luscombe N, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model**. *J Mol Biol* 2001, **313**:673-681.
The authors show that the occurrence of protein folds and the use of protein families by extant genomes follow a power law behaviour. They show that, although the majority of families are utilised very seldom, some families are highly over used.

58. Wolf Y, Grishin N, Koonin E: **Estimating the number of protein folds and families from complete genome data**. *J Mol Biol* 2000, **299**:897-905.

59. Luscombe N, 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**:40.

60. Shakhnovich B, Dokholyan N, DeLisi C, Shakhnovich E: **Functional fingerprints of folds: evidence for correlated structure-function evolution**. *J Mol Biol* 2003, **326**:1-9.

61. • Rost B: **Enzyme function less conserved than anticipated**. *J Mol Biol* 2002, **318**:595-608.
Rost demonstrates that recent work implying that enzyme function is highly conserved within protein superfamilies may overestimate the level of conservation.

62. • Anantharaman V, Aravind L, Koonin EV: **Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins**. *Curr Opin Chem Biol* 2003, **7**:12-20.
A comprehensive review of current work in assessing the structural basis of functional evolution. Common folds are observed to play roles in a variety of biochemical reactions and such ubiquity appears to involve generic, symmetrical substrate-binding sites.

63. Todd AE, Orengo CA, Thornton JM: **Plasticity of enzyme active sites**. *Trends Biochem Sci* 2002, **27**:419-426.

64. • Bashton M, Chothia C: **The geometry of domain combination in proteins**. *J Mol Biol* 2002, **315**:927-939.
This paper presents an analysis of the conservation of domain order within multidomain proteins. The authors observe that sequential order is the same in 98% of domain pairings. They note that sequential order in protein sequence appears not to affect the spatial arrangement of the protein structure. They conclude that sequential order is most probably conserved as it represents fewer chromosomal recombination events.

65. • Rison S, Teichmann S, Thornton J: **Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli***. *J Mol Biol* 2002, **318**:911-932.
This paper deals with a subset of enzymatic *E. coli* genes. The thorough analysis shows that domains within a pathway are unlikely to be homologues of each another. Domains are most likely to be recruited between pathways and are recruited on the basis of reaction mechanism, rather than substrate-binding characteristics.

66. Teichmann S, Rison S, Thornton J, Riley M, Gough J, Chothia C: **Small-molecule metabolism: an enzyme mosaic**. *Trends Biotechnol* 2001, **19**:482-486.

67. Tsoka S, Ouzounis C: **Functional versatility and molecular diversity of the metabolic map of *Escherichia coli***. *Genome Res* 2001, **11**:1503-1510.

68. Alves R, Chaleil R, Sternberg M: **Evolution of enzymes in metabolism: a network perspective**. *J Mol Biol* 2002, **320**:751-770.

69. • Jardine O, Gough J, Chothia C, Teichmann S: **Comparison of the small molecule metabolic enzymes of *Escherichia coli* and *Saccharomyces cerevisiae***. *Genome Res* 2002, **12**:916-929.
The authors compare the small-molecule enzymes of *E. coli* and *S. cerevisiae*. They note that 271 families are common to both genomes. These involve 384 gene products and 390 gene products in *E. coli* and *S. cerevisiae*, respectively. This accounts for half of the *E. coli* and two-thirds of the *S. cerevisiae* small-molecule metabolic enzymes. Of the 271 common enzymes, over 66% have 30–50% sequence identity. Around this core set of enzymes, the organisms have very different extensions to small-molecule metabolism.

70. Babu M, Teichmann S: **Evolution of transcription factors and the gene regulatory network in *Escherichia coli***. *Nucleic Acids Res* 2003, **31**:1234-1244.

71. Enright AJ, Van Dongen S, Ouzounis CA: **An efficient algorithm for large-scale detection of protein families**. *Nucleic Acids Res* 2002, **30**:1575-1584.

72. • Heger A, Holm L: **Exhaustive enumeration of protein domain families**. *J Mol Biol* 2003, **328**:749-767.
An algorithm, ADDA, is presented for domain decomposition and clustering of protein domain families. This is applicable on the scale of all available protein sequences and allows a global survey of protein domain space.

73. 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**:0040.