Exploiting protein structure data to explore the evolution of protein function and biological complexity

Russell L. Marsden*, Juan A. G. Ranea, Antonio Sillero, Oliver Redfern, Corin Yeats, Michael Maibaum, David Lee, Sarah Addou, Gabrielle A. Reeves, Timothy J. Dallman and Christine A. Orengo

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

New directions in biology are being driven by the complete sequencing of genomes, which has given us the protein repertoires of diverse organisms from all kingdoms of life. In tandem with this accumulation of sequence data, worldwide structural genomics initiatives, advanced by the development of improved technologies in X-ray crystallography and NMR, are expanding our knowledge of structural families and increasing our fold libraries. Methods for detecting remote sequence similarities have also been made more sensitive and this means that we can map domains from these structural families onto genome sequences to understand how these families are distributed throughout the genomes and reveal how they might influence the functional repertoires and biological complexities of the organisms.

We have used robust protocols to assign sequences from completed genomes to domain structures in the CATH database, allowing up to 60% of domain sequences in these genomes, depending on the organism, to be assigned to a domain family of known structure. Analysis of the distribution of these families throughout bacterial genomes identified more than 300 universal families, some of which had expanded significantly in proportion to genome size. These highly expanded families are primarily involved in metabolism and regulation and appear to make major contributions to the functional repertoire and complexity of bacterial organisms.

When comparisons are made across all kingdoms of life, we find a smaller set of universal domain families (approx. 140), of which families involved in protein biosynthesis are the largest conserved component. Analysis of the behaviour of other families reveals that some (e.g. those involved in metabolism, regulation) have remained highly innovative during evolution, making it harder to trace their evolutionary ancestry. Structural analyses of metabolic families provide some insights into the mechanisms of functional innovation, which include changes in domain partnerships and significant structural embellishments leading to modulation of active sites and protein interactions.

Keywords: protein structure; function; evolution; genome analysis

1. INTRODUCTION

We are at the dawn of an exciting era in biology. Thanks to the genome initiatives, there is now a large complement of completed genomes (greater than 260), with a significant number (greater than 30) from eukaryotic organisms, including human. The last decade has also seen the development of some highly sensitive methods for comparing protein sequences and structures to reveal evolutionary relationships (see Grant *et al.* 2004; Redfern *et al.* 2004 for reviews). This means that many proteins from completed genomes can be assigned to an evolutionary family that is already well characterized in a public database (e.g. Pedant, Frishman *et al.* 2003; InterPro, Mulder *et al.* 2003; ProtoNet, Sasson *et al.* 2003; Pfam,

Bateman *et al.* 2004; SCOP, Andreeva *et al.* 2004; CATH, Pearl *et al.* 2005), telling us much about their putative functions. Thus, we can examine the distribution of different families among the various species and kingdoms and also consider the ways in which differential expansions of families and modifications of function in paralogues have influenced the functional repertoires of these organisms (see Orengo & Thornton *in press*, for a review).

It is clear from recent analyses of this genomic data that a large proportion of proteins are multi-domain, particularly in eukaryotes where figures as high as 80% have been suggested (Apic *et al.* 2001). Analyses have also confirmed the extent to which some very common domains are duplicated, shuffled within a genome and combined in different ways (Vogel *et al.* 2004a). This follows a mosaic model of constructing new proteins (Vogel *et al.* 2004b) that can subsequently evolve modified functions, thereby expanding the functional repertoire of the organism.

* Author for correspondence (marsden@biochem.ucl.ac.uk).

One contribution of 15 to a Discussion Meeting Issue 'Bioinformatics: from molecules to systems'.

Protein family resources have played an important role in helping to organize this data and analyses exploiting these resources are beginning to shed light on the mechanisms by which proteins evolve and acquire new functions. Because a protein's structure is much more highly conserved than its sequence (Chothia & Lesk 1986; Sillitoe & Orengo 2003), structural data allows us to track further back in evolution, detecting very remote homologies for which all similarity has been washed from the sequence—in the Midnight zone as coined by Rost (1997). Although there is still a vast gap in the numbers of structures (approx. 21 000 unique protein chains in the protein data bank (PDB, Deshpande *et al.* 2005)) and sequences (greater than 2.4 million non-redundant sequences in GenBank (Benson *et al.* 2005), greater than 1.7 million non-redundant sequences in UniProt (Apweiler *et al.* 2004)) determined, at least 60% of sequences or partial sequences from completed genomes can now be assigned to domain structure families in one of the major classifications (Madera *et al.* 2004; McGuffin *et al.* 2004; Lee *et al.* 2005). This is a significant proportion and means that we now have sufficient structural annotations to start exploiting these data to probe evolutionary relationships more deeply.

First, we must apply sensitive bioinformatics algorithms that allow us to detect these evolutionary relationships between proteins, both from the structural data and, where there is no information on structure, by exploiting sensitive sequence profiles. Below we review some recent approaches but concentrate mainly on describing the approaches we have developed for classifying evolutionary relatives in our two in-house protein and domain family resources (Gene3D (Buchan *et al.* 2002; Lee *et al.* 2005) and CATH (Pearl *et al.* 2005)).

2. USING STRUCTURAL DATA TO IDENTIFY REMOTE HOMOLOGIES

Many different approaches have been used to compare protein structures (see reviews Orengo *et al.* 2001; Sillitoe & Orengo 2003; Redfern *et al.* 2005). In 1993, we used the SSAP algorithm developed by Taylor & Orengo (1989) to align structural domains and classify them into evolutionary superfamilies in the CATH domain database (Orengo *et al.* 1993; Orengo *et al.* 1997). SSAP compares residue structural environments between domains and uses double dynamic programming to handle insertions and deletions, allowing even very remote homologues (less than 20% sequence identity) to be detected. CATH now contains 67 054 domain structures clustered into approximately 1600 evolutionary superfamilies. These can be collapsed into approximately 900 'fold groups' as some superfamilies are found to adopt similar folding arrangements, although they share no other common feature. Lack of other similarities may suggest these proteins have diverged significantly during evolution, changing their sequences and function considerably. Alternatively, convergence to the same fold may have occurred during evolution, for energetic reasons.

Over 5 years ago, several worldwide structural genomics initiatives were funded, aiming to target structurally uncharacterized protein families for structure determination. These initiatives hope to reduce the amount of redundancy prevalent in the current PDB, whereby approximately 90% of sequences are nearly identical, often because the same structures have been targeted to determine the impact of a mutation or conformational changes that occur following ligand binding. By targeting unique families, unrelated to those for which representative structures had already been solved, the aim of the structural genomics initiatives was to increase the coverage of 'fold space' and determine the complete repertoire of structural types.

To keep pace with these initiatives several groups developed very fast structural search methods that allowed crystallographers to assign newly solved structures to existing families or determine whether they did indeed represent unique folds. For example, the CE method developed by Bourne and co-workers (Shindyalov & Bourne 1998) at the PDB and the secondary structure matching (SSM) method for the macromolecular structure database (Velankar *et al.* 2005) allow you to rapidly search the PDB with a new structure to find structural relatives. Other fast, publicly available methods of high accuracy include VAST for the Entrez database (Gibrat *et al.* 1996) and the STRUCTAL/LSQMAN method, developed by Levitt and co-workers (Kolodny *et al.* 2005; see Redfern *et al.* 2005 for a review).

We have also developed a very fast algorithm, CATHEDRAL (CATH's existing domain recognition algorithm), to increase the speed and ease with which new structures can be classified in the CATH domain database. Furthermore, because technological improvements in crystallography have increased the numbers of multi-domain structures selected by crystallographers for structure determination and since a large proportion of genome sequences, more than two-thirds, are expected to be multi-domain, our aim was to develop a method that combined speed and accuracy in recognizing domain folds, with accuracy in determining the boundary of that fold within the multi-domain structure.

CATHEDRAL first performs a very rapid search by simply comparing secondary structure elements between proteins, rather than residues, as there are generally an order of magnitude fewer of these. Graph theory algorithms are exploited, originally pioneered by Artymiuk and co-workers (Mitchell *et al.* 1990). These reduce the three-dimensional information in a protein structure to a two-dimensional graph in which the nodes represent secondary structures and can be labelled according to their type. Edges connecting the nodes are labelled according to information on the angles and distances separating the secondary structures. Graphs can be easily compared using the Bron–Kerbosch algorithm, which detects the largest common subgraph with nodes of the same type, sharing distances and angles within an allowed threshold. These approaches give very high sensitivity, up to 98%, depending on the structural family, for a small error rate of 0.1% (Harrison *et al.* 2002).

Having benefited from a fast initial comparison to identify the putative fold of a domain structure, CATHEDRAL then refines the alignment by performing a slower more accurate alignment of all relatives within that fold group using the residue based double dynamic programming algorithm, SSAP (Taylor & Orengo 1989). CATHEDRAL shows good performance in fold recognition, comparable to other publicly available algorithms, and can be used to provide both local alignments of well superposed fragments as well as global alignments corresponding to a significant proportion of the query.

Because the fold recognition stage is so rapid, very large database searches can be performed, allowing a sound statistical framework based on the extreme distribution of scores obtained for random matches. This allows us to assess the significance of any structural match, an important feature when attempting to find the best fold matches to a multi-domain protein, so that domain boundaries can be most accurately established. Nearly one-third of structural domains characterized within CATH are discontinuous and the domain is assembled from chain segments that are disconnected in the sequence but co-located in three-dimension. This occurs following the insertion of another domain or structural motif by DNA shuffling, during evolution. To improve boundary recognition of these domains, CATHEDRAL performs iterative searches whereby large, contiguous domains scoring above an optimized threshold are identified first and excised to improve the recognition of the remaining discontinuous domains.

For 74% of a large dataset of 2079 domains within 890 multi-domain proteins, CATHEDRAL could locate the boundary within ± 10 residues, an encouraging performance given that in many highly populated families in CATH, structural relatives show considerable structural diversity (Sillitoe & Orengo 2003; Reeves *et al.* in preparation) and so known structural relatives may vary quite considerably from any new instance of the family being determined. For classification in CATH, manual inspection allows rapid refinement of the domain boundaries. Currently, CATHEDRAL is the only publicly available structure comparison method that performs both fold group recognition and domain boundary assignment. A CATHEDRAL web server (<http://cathwww.biochem.ucl.ac.uk/cgi-bin/cath/CathServer.pl>) allows structural biologists to scan the CATH fold library with newly determined structures so that the boundaries of any domain regions matching existing folds within the CATH database can be returned.

3. EXPLOITING SEQUENCE PROFILES TO RECOGNIZE REMOTE HOMOLOGUES

Structure comparison methods allow very remote relatives to be recognized, since the structural environments of at least 50% of residues in the domain core are conserved during evolution, even when the sequence similarity falls below 15% (Sillitoe & Orengo 2003). However, the repertoire of known structures is still 30 fold smaller than known sequences, as structure determination remains a more difficult technology

(Sali 1998), and therefore, distant relationships between many proteins can only be established using sequence comparisons (see Nagl 2003; Redfern *et al.* 2005 for a recent review). Of these, the profile-based methods are the most powerful. These perform a multiple sequence alignment of family relatives to determine the frequencies of residue types at different positions in the protein and thus the likelihood of finding these patterns in a newly sequenced relative. Hidden Markov models (HMMs) are among the most sensitive of these approaches, performing well in the critical assessment of techniques for protein structure prediction competitions (Aloy *et al.* 2003).

To recognize sequence relatives for structural families in CATH, we optimized parameters used by the SAM-T HMM protocol developed by Karplus *et al.* (1998). The best way of doing this is to construct a dataset of very remote relatives (less than 35% sequence identity) that have already been recognized by structural comparisons, and then determine the percentage of this dataset that can be recognized by the sequence based methods. This benchmarking strategy, pioneered by Chothia and co-workers (Park *et al.* 1998), gives a stringent test for assessing the recognition of remote homologues.

For CATH, we have built a library of 6003 HMMs, one for each sequence family, and found that this library can recognize 76% of remote sequence homologues. This is a nearly 20% improvement in performance compared to that obtained using a similar approach nearly 2 years ago (see Sillitoe *et al.* 2005), and is largely due to the increase in the number of relatives in both the structural and sequence databases over that time. These increases have derived from both an expansion in the fold library and also an increase in HMM sensitivity as more divergent sequence relatives from each family are sampled. Gough and co-workers have obtained similar performances using HMM libraries built from representatives in the SCOP domain database (Gough *et al.* 2001).

Further improvements in performance can be achieved by exploiting a profile-profile protocol, whereby HMMs are constructed using the query sequences as seed sequences and then scanned against an existing HMM library made from SCOP families (Yona & Levitt 2002). Several groups have attempted to increase sensitivity by building HMMs that exploit structural data in some way, for example by seeding the HMM from a multiple structural alignment of remote relatives in a family. By including more distant relatives but ensuring they are carefully aligned through consideration of structural data, the residue frequency pattern captured by the HMM should represent even the most divergent members of the family and anchor the positions most important for the folding, stability or function of the domain. While early attempts using profile methods showed considerable promise (Kelley *et al.* 2000), more recent analyses have revealed relatively small improvements in performance of less than 5% (Griffiths-Jones & Bateman 2002; Sillitoe *et al.* 2005).

Another more successful approach for stretching performance is to exploit an intermediate sequence library. For CATH an expanded HMM library of

Figure 1. Cumulative residue annotation coverage of 30 representative completed genomes; 10 archaeal (top), 10 bacterial (middle) and 10 eukaryotic (bottom). Annotation coverage is calculated on a hierarchical basis, such that the percentage of residues assigned to CATH domains are calculated first, followed by calculation of the percentage of remaining residues assigned to Pfam domains, and so on.

27 161 models (CATH-ISL HMM) is built by adding HMMs seeded from additional CATH domain sequences identified in UniProt. These were recognized using the original CATH HMM library built from structural relatives. By scanning against this expanded library, 86% of remote homologues can be recognized in the benchmark dataset for the same low error rate of 0.1%, a 10% increase in performance, equivalent to that observed for profile–profile based methods. However, speed is obviously an issue. As with profile–profile based methods, scanning the CATH-ISL HMM can take ten times longer. It is not feasible to perform these scans on all non-redundant sequences in UniProt or GenBank and so they are currently only being undertaken for representative genomes from each kingdom to assist in comparative genome analyses.

4. IDENTIFYING DOMAIN FAMILIES IN THE GENOME

(a) *Targeting structurally uncharacterized families for structure determination*

Many international structural genomics initiatives, currently underway in the USA, Europe and Asia (see Brenner 2001; Todd *et al.* 2005), are attempting to increase our knowledge of ‘structure space’. Since 2000, the NIH in the United States have funded a series of pilot projects, with the primary aim of targeting families containing no close structural relatives. It is hoped that each newly solved structure will allow homology models to be built for new areas of sequence space. We have participated in the Midwest Consortium for Structural Genomics headed by Andrzej Joachimiak at the Argonne National Laboratory.

In order to identify suitable targets and prioritize certain types of families, we have developed a new resource, Gene3D, for comparing sequences from completed genomes (Lee *et al.* 2005). By performing comparative genome analysis we can identify families that are highly specific, for example, to a particular pathogenic organism or alternatively those families common across many species, that may perform essential functions in humans and other organisms, but because they are present in bacteria are more amenable to expression and crystallization.

We first identify domain sequences in the genomes that can be assigned to families of known structure to determine whether they have close structural relatives (i.e. having $\geq 30\%$ sequence identity), thus allowing a homology model to be constructed (Bray *et al.* 2004). This sequence identity cutoff is often considered to relate to a lower bound of reliability for homology models. More remote relatives may be suitable for target selection if their predicted secondary structure or functional annotation suggest they perform a very different functional role and that there are likely to be considerable structural differences to the relative that has already been structurally characterized.

To date, we have scanned protein sequences from 203 completed genomes against the CATH HMM library to identify domain regions that can be assigned to CATH structural families. Approximately 50% of domain sequences are recognized as relatives of CATH families, corresponding to nearly 40% of residues within genomes (figure 1), though the average levels of annotation tend to be lower for eukaryotic organisms.

These levels of annotation are slightly lower than those reported for other public resources. For example,

Figure 2. (a) Assignment of CATH, Pfam and NewFam regions to genome sequences in the Gene3D database. A hierarchical scheme is used, where CATH domains are first assigned, followed by non-overlapping Pfam domain assignments. We label regions with 50 or more residues, without a CATH or Pfam assignment, as NewFam regions. (b) Residue coverage of CATH, Pfam and NewFam annotations across genomes in the Gene3D database. The coverage of each annotation type is given as a percentage of residues, excluding sequences belonging to single member CATH, Pfam and NewFam families.

the SCOP based superfamily resource assigns on average 60% of sequences or partial genome sequences to SCOP superfamilies (Gough *et al.* 2001). For assigning CATH annotations a strict threshold requires that 50% of residues the CATH domain are matched to residues in the genome sequence before accepting the match. This is a rather conservative estimate that no doubt misses some very remote homologues but does guarantee a very low error rate of less than 0.1%.

Higher levels of structural annotation are recorded by resources exploiting threading-like algorithms. The genome threading database of Jones and co-workers (McGuffin *et al.* 2004) cites average sequence based annotations of 64% and residue based annotations of 58% obtained using GenTHREADER (Jones 1999). The 3D-GENOMICS database of Sternberg and co-workers (Fleming *et al.* 2004) also provide genome-scale domain annotations, in this case using the HMMer algorithm (Eddy 1998). These approaches are able to recognize a higher proportion of remote homologues in the midnight zone (less than 20% sequence identity) by combining sequence information with some structural data, e.g. on solvent accessibility or residue contacts to 'thread' a sequence through a domain structure in an energetically favourable manner.

(b) Providing domain family assignments for structurally uncharacterized domain sequences

Thus, even these extremely powerful approaches assign less than two-thirds of genome sequences to structural families in SCOP or CATH. To recognize other types of domains, and resolve the domain composition of the proteins, we have annotated the remaining sequences with Pfam domains. Pfam is an excellent protein family resource developed by Bateman *et al.* (2004). It currently contains 7677 domain families that were initially seeded using domain sequences identified by

the ProDom resource of Corpet *et al.* (2000). Pfam families are further populated by matching UniProt sequences to HMMs built automatically for each family, using the HMMer algorithm (Durbin *et al.* 1998). Considerable manual validation of alignments and of the assigned relatives is performed and the sequences are also matched against SCOP entries, where possible, to resolve domain boundaries.

(c) Assigning domain sequences in the genomes to CATH and Pfam families

To maximize domain annotations in Gene3D, genome sequences are scanned against both CATH and Pfam HMM model libraries, both built using SAM-T99 for consistency. Assignments to CATH domains take precedence as structural data allows more reliable assignment of domain boundaries. The DomainFinder algorithm is used to resolve conflicting matches (Pearl *et al.* 2001). Approximately 1432 Pfam superfamilies overlap completely with 803 CATH superfamilies resulting in an almost twofold collapse of these families. In some cases CATH domains overlap a portion of a Pfam domain, but provided the remaining Pfam assignment is at least 50 residues this is counted as a separate Pfam domain (see figure 2a). By adding Pfam assignments, over three quarters of the residues can be assigned to a well-characterized domain family in either CATH or Pfam (figure 2b).

In providing annotation for genome sequences, we have also performed several analyses to identify regions associated with transmembrane helices, coiled coils, N-terminal signal peptides and low complexity regions. On average these comprise about 14% of residues in a genome within a range of 10 to 22%, values similar to those calculated by Liu & Rost for a much smaller set of genomes (Liu & Rost 2001).

To provide some characterization for sequence regions not assigned to CATH or Pfam, we considered

Figure 3. The length distribution of NewFam regions in Gene3D sequences compared to domains in the CATH domain database. A large proportion of NewFam sequences appear domain-like in their length.

all unassigned regions comprising more than 50 residues that could thus be treated as putative domains. By comparing the sequences of these segments using BLAST and clustering related sequences into families using a robust clustering tool (TRIBE-MCL) developed by Ouzounis and co-workers (Enright *et al.* 2003), we detect a further 42 883 domain families, which we describe as NewFam families, the majority of which are very small (89% of families contained less than five relatives) and comprise fewer than 200 residues. Figure 3 shows that the length distribution of NewFam sequences is shifted to lower values than domain structures in CATH, suggesting that most of these families do indeed correspond to single domains. Figure 4 shows that CATH domains comprise the largest families found in the genomes, followed by Pfam, then NewFam. This is not surprising as structural data allow much more distant evolutionary relatives to be traced.

It can be seen from figure 2*b* that, on average, 45% of residues in completed genomes are assigned to CATH structural families, 31% to Pfam families and the remaining 24% to the uncharacterized NewFam families. We are currently analysing some of the larger NewFam families to determine whether they represent species diverse and functionally interesting families that may be good targets for the structural genomics initiatives. Information on the domain assignments to CATH, Pfam and NewFam families can be viewed and searched in our Gene3D resource, established in 2002 (Buchan *et al.* 2002; Lee *et al.* 2005; <http://bsmmac1.biochem.ucl.ac.uk:8080/Gene3D>).

By mapping domain sequences to CATH, Pfam and NewFam families and sorting these in order of the largest families, we see that approximately 70% of the domain sequences in completed genomes can be assigned to fewer than 2000 of the largest CATH and Pfam families. Seven hundred and thirty four of these are CATH families containing structural representatives. Some of these are extremely large families, in which homology models can only be reliably built for less than 10% of relatives. For many of the remaining relatives, predicted secondary structures and functional annotations suggest that there has been considerable divergence in structure and function and that targeting some of these relatives would be an important goal for

the structural genomics initiatives and would help illuminate the structural mechanisms by which new functions evolve in protein families (see also below).

The structurally uncharacterized Pfam families would also make excellent targets for structural genomics initiatives, providing structural data for families for which none currently exists and ensuring that at least three quarters of the genome sequences can be assigned to a characterized domain family of known structure. Although targeting the largest families first enables a higher proportion of genome sequences to be accurately modelled, this is not to discourage the careful selection of targets from small families, e.g. NewFam families, particularly those specific to human or pathogenic organisms.

5. ANALYSING THE DOMAIN REPERTOIRE IN COMPLETED GENOMES AND HOW IT INFLUENCES BIOLOGICAL COMPLEXITY

There are now more than 260 completely sequenced genomes, from all kingdoms of life. Although many of these are bacterial (greater than 200), there is a slowly increasing proportion of eukaryotic genomes, currently over 30, and with at least 21 archaea, and so we are now equipped to compare the gene complements of these diverse organisms and examine the manner in which changes in the gene repertoire influence the functional repertoire and biological complexity of the organisms. Insights have already been obtained by several groups examining sequence similarities between proteins in the different organisms and observing trends in the distributions of specific families. Koonin *et al.* (2002) have identified a core set of families common to all kingdoms of life and mainly associated with protein biosynthesis, supporting Woese's hypothesis that these essential proteins would be expected to 'crystallize' early in evolution and remain unchanged once they had evolved to a certain level of efficiency and robustness (Woese 1998).

Clearly, these analyses are limited by the current dataset, which though enriched with each new genome completed still represents a tiny proportion of all species in Nature. Venter *et al.* (2004), extracting organisms from the Sargasso Sea, stunned biologists with the revelation that 1 kg of sea water contains more than 1800 species. However, by comparing across all the kingdoms, we should discern some trends, valuable to our understanding of how complex life emerged. Furthermore, by exploiting the structural data we employ a more powerful lens into the past, as the structures of the proteins change much less and even the most diverse paralogues frequently retain structural similarity in at least half their residues (Chothia & Lesk 1986; Orengo *et al.* 2001*b*). These are usually in the core of the structure, which may then be embellished in rather different ways in diverse relatives to engineer a variety of protein interactions and modulated functions (see also below).

(a) Recruiting domain families to different biochemical pathways

Several groups have already started examining the structural data to answer some key questions in

Figure 4. Power-law distributions of (a) CATH, (b) Pfam and (c) NewFam domain families. Power-laws are fitted to the data and the gradients are compared in (d).

evolution. For example, how have different biochemical pathways evolved. Various hypotheses had been proposed (Todd *et al.* 2001; Teichmann & Babu 2004), whereby gene duplication plays a large role in expanding functional repertoire, since the additional gene copy or paralogue is free to evolve new functional roles. By analysing the distribution of relatives from CATH and SCOP families in metabolic pathways, for which comprehensive data has now been assembled by the Kegg Database in Kyoto (Rison *et al.* 2002; Kanehisa *et al.* 2004) demonstrated that there was a significant tendency for paralogues to be recruited to different pathways within an organism, where their roles were frequently determined by the type of chemistry they could perform. This finding supported a 'patchwork' theory of evolution initially proposed by Jensen (1976).

An alternative mechanism first proposed by Horowitz (1945), whereby pathways expand by duplication of their component genes, so that the resulting paralogues evolve to perform different catalytic steps along the pathway and thereby improve the efficiency of the biochemical transformation, appears less likely. Analyses of functional diversity among structural paralogues in CATH also supported the Jensen theory, revealing that some aspect of the chemistry was often highly conserved, e.g. a chemical intermediate, while substrate specificity frequently changed considerably as paralogues participated in diverse pathways and were involved in transforming very different substrates (Todd *et al.* 2001).

(b) How differential domain family expansion can influence phenotypes

Other insights provided by the structural data probe more complex levels of function, revealing changes in

phenotypes associated with differences in the domain repertoire. Chothia and other researchers in Cambridge have used structural data to illustrate the manner by which differential duplication of domains in different species can influence functions and phenotypes exhibited by the organisms. For example, studies of the immunoglobulin superfamily that perform important roles in cell-cell communication, such as cell-surface receptors during embryonal development, have revealed dramatic changes occurring in different eukaryotic organisms following differential gene expansions (Vogel *et al.* 2003).

In the fruitfly, specific expansions mainly affect cell-cell communication, e.g. as receptors that are involved in neuronal development. In the worm, the relatives from this family are largely muscle or extracellular matrix proteins. The fly specific expansion appears to contribute to higher physiological complexity in the fly by equipping it, for example, with a more intricate nervous system (Vogel *et al.* 2003). In their analysis Vogel and co-workers, therefore, define two types of expansions in protein family repertoires. Conservative expansion increases an organism's ability to adapt to its environment but does not significantly affect physiology; this is seen in the expansion of two chemoreceptor families in the worm but not in the fly, and also in the dramatic expansions in some metabolic families in bacteria, described below. Progressive expansion leads to significant changes in form that may promote an organism's ability to adapt and may also contribute to divergence of species.

(c) Domain family expansions influencing complexity in bacterial genomes

Information on domain structure repertoires, in each species, can easily be extracted from our Gene3D

resource and compared both within and across kingdoms. Since bacterial organisms are so highly represented within the dataset, we have first examined trends in domain distributions within this kingdom. There are more than 940 domain structure families from CATH identified in a dataset of 100 bacterial genomes, accounting for nearly two-thirds of the superfamily domains. Of these, 359 are common to at least 70% of the species, suggesting that they are very likely to be universal to all bacterial genomes. This threshold is reasonable given the sensitivities of the current profile-based methods used to assign sequences to CATH families (see above). The number of universal domains may even represent a conservative estimate. If function changes and the structural constraints associated with the original function are removed, relatives can diverge considerably in sequence and structure (see also below) beyond the limits of our detection methods.

We examined the behaviour of these 359 common or 'universal' domains further, to understand their contribution to and importance for genome evolution. By determining the extent to which the domain families had expanded within each organism, i.e. by counting the number of relatives within the genome, we could discern several different types of behaviour. Information on family expansion was captured in an 'occurrence profile', inspired by the work of Eisenberg and co-workers, who derived 'phylogenetic profiles' to detect the presence or absence of orthologues in completed genomes and thereby infer functional associations (Pellegrini *et al.* 1999). For our analysis, the occurrence profiles determines the number of relatives or degree of expansion of the family, within each organism.

Three types of behaviour could be clearly discerned using the Gene3D occurrence profiles. The functions of these families were analysed using data from the literature and also the COGs database, established by Koonin and co-workers (Tatusov *et al.* 2001). This has four main categories; information storage and processing, cellular processes, metabolism and poorly characterized. Within each of these categories there are further subdivisions describing function more precisely, e.g. the information class is subdivided into translation, RNA processing and transcription. One set of universal domains, found to have similar numbers of relatives in each genome, tend to be small families principally involved in translation and protein biosynthesis and can be described as 'evenly distributed', reflecting the fact that their numbers change little between different bacterial organisms.

The two other types of families show considerable expansions with increasing genome size. Standard residual analyses of these genome size correlations revealed one group of about 38 superfamilies that have expanded linearly with increasing genome size. While the other group of about 20 families have expanded nonlinearly with genome size, following a power law of rank close to two. The first type of family, which we describe as 'linearly increasing', has been the most extensively duplicated of all families in bacterial genomes and although there is some diversity in function observed, the majority of domain families

are principally involved in metabolism, with some relatives assigned to the poorly characterized class in COGs. The second set, which we describe as 'nonlinearly increasing', also have a relatively high duplication rate, though on average not as high as the linearly increasing domains, and the dominant function observed for these families is gene transcription regulation. The domains associated with these two types of expanding families contribute to nearly 50% of domain structure annotations in these bacterial genomes.

(d) *The balance between metabolic and regulatory domain families in bacteria influences genome size*

While the 'evenly distributed' domain families, associated with translation, show no differential expansion in any of the organisms studied, by contrast, the size dependent families are clearly making a significant contribution to genome size. Since in bacteria the number of open reading frames (ORFs) can be related to genome complexity (Mira *et al.* 2001), these domains are, therefore, also contributing to genome complexity.

Intriguingly, as bacterial genome size increases, there is significant expansion occurring in the number of nonlinearly increasing domains relative to the linearly expanding domains. This behaviour suggests an analogy with models that describe the microeconomics of factories. We can imagine that the linearly dependent families, associated with metabolism, are contributing to the 'wealth' of the organism. This is because expansion of these families is often accompanied by functional divergence. Since bacterial survival depends on fast replication of the genome and high levels of gene redundancy are not tolerated, so paralogues retained within the genome frequently evolve modified functions increasing the organism's ability to benefit from different nutrients and substrates in the environment.

However, there is a cost associated with this, as paralogous domains retain some similarities with the original orthologue and therefore, to control this expanded repertoire efficiently, it is necessary to expand the number of regulatory proteins. Since the scale of this will reflect the increase in the number of different interactions and pathways made possible by the expansion in the metabolic domains, it is reasonable to expect this to be a nonlinear effect as multiple interactions may be involved. As with an efficient factory, a successful bacterial organism would be expected to balance the 'revenue' acquired by increase in certain metabolic domains against the 'cost', in terms of slower replication rate caused by the increasing numbers of regulatory domains needed to control the new metabolic processes.

This is analogous to a factory branching out to produce new products by designing new ways of exploiting existing components. Studies have shown that as more product lines are designed, at some stage the number of managers required to ensure smooth running of the production process starts to affect profit, since managers are much more expensive than the workers producing the goods. At some point it is no

longer profitable to introduce new lines and thus an optimum number of lines is reached. Similarly with bacteria, we can imagine that the benefit to the organism realized by expanding its metabolic repertoire is offset by the necessary increase in regulatory genes. Provided the benefit outweighs the cost the metabolic families will continue to expand. By taking the difference between the expansions in these families at increasing genome sizes, an optimal size can be predicted where these effects are balanced. Furthermore, the optimum calculated, of 4800 genes, coincides with the statistically most frequent genome size observed in non-specialized bacteria, i.e. bacteria with no specific environmental requirements, suggesting that although we are only sampling a subset of bacterial genomes and only those families for which structural data exists, the trends we are observing are clearly contributing in some major way to bacterial genome size and complexity.

(e) What can we learn about the last universal common ancestor (LUCA) of all kingdoms of life from the domain structure annotations

Although we have fewer completed genomes for eukaryotes (greater than 30) and archaea (21), because they are from dispersed phylogenetic branches we can start to compare the domain repertoires across all kingdoms of life and draw some tentative conclusions on the nature of domains present in early ancestral organisms. Again, by using Gene3D, and reasonably conservative thresholds, we can identify a subset of about 140 domains common to all kingdoms of life and therefore highly likely to have been present in the last universal common ancestor (LUCA). These ancient domain families account for nearly 50% of domain annotations, on average, in each genome, indicating that at least half of the structurally annotated genes in Gene3D come from a relatively few phylogenetic lineages that originated before the separation of the major kingdoms.

These ancestral domains have representatives in the four major functional groups from COGs, supporting the theory of a genetically and functionally evolved LUCA. Translation and metabolic functional groups comprise the majority of ancestral families, having similar numbers of ancestral domain families (48 and 46, respectively). Metabolism has undergone a higher expansion than translation during evolution so that metabolic families in LUCA comprise only 12% of the total number of all metabolic families in COGs. This proportion is 53% for families involved in translation. Families involved in cellular processes are the third most represented group in LUCA with almost double the number of families than replication (21 and 12, respectively). These data suggest a more complex picture of LUCA than emerges from analyses exploiting sequence information alone, which paint an image of a LUCA primarily composed of families involved in protein biosynthesis.

It is unlikely that genes involved in universal and, therefore, essential cellular functions would be horizontally transferred and therefore, massive horizontal gene transfer (HGT) of these essential genes seems improbable (Kurland *et al.* 2003). Although HGT

probably played an important role at the very beginning of evolution, the high ubiquity and occurrence of the ancestral superfamilies suggests Darwinian evolution as a major process (Koonin *et al.* 2002; Chothia *et al.* 2003).

(f) How comprehensive is our picture of LUCA?

Previous analyses of completed genomes, while detecting common domains involved in translation (see Koonin (2003) for a review), have had difficulty tracking other types of families. However, it seems unlikely that early organisms would evolve a robust system of building proteins but few biological processes that use these proteins to exploit the organism's environment. Could it be that in some families gene duplication, followed by modification and acquisition of new functions in paralogues, has resulted in some relatives diverging beyond the limit of detection using profile-based methods so that new families appear to have arisen during evolution?

We can consider the universality of a domain family as the percentage of organisms from all kingdoms of life in which it can be detected. If we compare the different types of domain families, again restricting ourselves largely to data from the bacterial genomes, which are sufficiently large to allow statistical trends to be detected, we can calculate the degree of universality for each domain family. Families involved in metabolism show a wide range of universality values reflecting the important role of metabolism in bacterial adaptation, as discussed above, and the expansion in metabolic domains as a means of generating new functional variants (Ranea *et al.* 2004, 2005).

(g) How can we measure the degree of innovation of domain families?

Woese introduced the term 'evolutionary temperature' to describe whether systems had effectively crystallized their functions (cold), so no or little further functional change is observed, or whether systems had been more suitable to accept genetic variants (hot; Woese 1998). He described families involved in translation as being much 'cooler' than other families.

We examined a number of properties that could be used to measure this evolutionary parameter and ascertain whether the genetic and functional diversity of the family suggested an evolutionary 'cold' or 'hot' family. We found that duplication rates, which are highly correlated with the number of functional variants in a family, vary considerably for different families. On average, families involved in translation have low duplication rates and also tend to be universal to all species. By contrast families involved in metabolism or regulation and poorly characterized domains have much higher duplication rates, accompanied by functional diversity, and they exhibit much wider ranges of universality.

When we consider the correlation of family occurrences with genome size, we find that families involved in translation tend to have low correlation with genome size, with a distribution close to that expected for random. By contrast families involved in metabolism are much more likely to have occurrences highly correlated with genome size, with a distribution far

Figure 5. Size correlation analysis. For the 727 functionally annotated superfamilies: superfamily percentages (y-axis) distributed by size correlation coefficient classes (x-axis). Displayed distributions: random distribution: J, translation; M, metabolism.

removed from random (see figure 5). Furthermore, these families tend to have highly varying numbers of relatives in each organism, reflecting their versatility to expand in a specific manner for each organism, enabling efficient exploitation of the organism's environment.

Finally, we can infer an 'innovation rate' that reflects the extent to which families within a given functional group (e.g. metabolic) are distributed among the different kingdoms. Most families involved in translation have high universality (see figure 6). By contrast, it can be seen that the poorly characterized families and metabolic families show a very wide range of universality, with some families common to all species and others highly specific for organisms or subspecies.

All these data suggest that families involved in metabolism have remained highly innovative and thus 'hotter' during evolution. Following duplication, paralogues diverge to acquire completely new functions, leading to further divergence from the ancestral domain and thus, in many cases, giving rise to a 'new' domain family.

These findings help to explain why previous analyses, using sequence data alone, mostly detect families associated with translation in LUCA. By exploiting the structural data we trace more distant relationships to reveal that families involved in other cellular processes such as metabolism and regulation were present in LUCA but have remained highly innovative throughout evolution. Because this erodes the common sequence signal from ancestrally related domains and since it is unlikely that we will acquire structural data for all relatives, we will always be restricted in our speculations on the domain repertoires in early life until we develop new methods for stepping back through evolution. However, it is likely that we have underestimated the contributions to LUCA from families in the functional groups (e.g. metabolism) that occupy a functional niche wherein it benefits to remain highly innovative during evolution.

Figure 6. Analysis of the superfamilies innovation rates of the functional groups. For the all functional groups, the plot displays the superfamily universal distribution percentage (y-axis) against the accumulated percentage of superfamilies (x-axis). One letter code: J, translation; L, replication; M, metabolism; C, cellular processes; P, poorly characterized; K, transcription.

6. IDENTIFYING PROTEIN FAMILIES TO EXAMINE THE DISTRIBUTION OF DOMAIN COMPOSITION REPERTOIRES IN COMPLETED GENOMES

Although our attempts to trace LUCA expose the current limitations in tracing evolutionary relationships given the high 'innovativity' or 'versatility' of certain functional groups, they provide important insights into families that have diverged significantly. How do novel functions arise in these domain families? Analysis of approximately 170 enzyme families of known structure in CATH revealed the extent to which function was changing in some highly versatile families. For example, in the P-loop hydrolase family more than 200 different functions can be detected. By mapping the structural domains to related sequences in GenBank, UniProt and completed genomes, using the technologies described above, we could extend the populations of the CATH superfamilies and explore the range of functions exhibited across a family more comprehensively. In some cases, functional change is brought about by mutation of a single residue, modifying or removing catalytic activity. However, more dramatic effects also occur. In more than 90% of families, showing high functional plasticity, changes in the domain partners were impacting on function. Effects due to varying partners ranged from modifications to the active site geometry, changes in the oligomerization state and in the surfaces involved in protein-protein interactions (Todd *et al.* 2001, 2002). This frequently modulated substrate specificity and the functional complexes in which the domains participate. By contrast, some aspect of the chemistry was often conserved, e.g. a shared chemical intermediate.

(a) Clustering sequences from completed genomes into protein families

To explore these effects further and provide comprehensive information on domain compositions that would help improve target selection in structural genomics, we have clustered the sequences from 203 completed genomes into protein families. This was done using a conservative protocol (PFscape, Lee *et al.* 2005)

that ensures that multi-domain sequences with the same domain composition, i.e. the same domains occurring in the same order along the sequence, are clustered. PFscape benefits from the robust clustering algorithm (TRIBE-MCL) devised by Enright *et al.* (2002), which analyses a matrix of pairwise sequence similarities obtained by BLAST, and then divides sequence space into clusters depending on the 'flux' of similarities between clusters. The algorithm is based on a sophisticated, mathematical approach developed by Van Dongen (2000).

Many other robust protocols have been applied to cluster protein sequences and provide information on protein families (e.g. SYSTERS, Krause *et al.* 2000; TRIBES, Enright *et al.* 2003; PROTONET, Sasson *et al.* 2003; see Redfern *et al.* 2005 for a review). However, because some domains have been very extensively duplicated and shuffled in genomes, it can be difficult to adjust parameters to prevent 'proteins' containing common 'domains' from clustering together, even though they have different domain partners. All resources suffer from this 'chaining' effect, to some extent. To try to minimize this, and maintain consistent domain compositions in each protein family, we optimized the parameters for the clustering protocol by exploiting structural data from CATH (Lee *et al.* 2005).

(b) Identifying the domain composition of protein families

Applying PFscape to 777 363 sequences from 203 genomes identifies 58 292 protein families, each containing two or more sequences, and 197 864 singleton sequences remain. Data on protein families are stored in the Gene3D database. CATH and Pfam domain annotations are obtained by mapping these domains onto the genome sequences using HMM technology. To obtain more comprehensive domain coverage any remaining regions are assigned to NewFam families using the protocol described above. Providing complete domain coverage enables domain composition to be characterized for each Gene3D family, so that outlying proteins can be removed to improve the cluster consistency.

(c) Updating protein families and providing functional annotations

A related update protocol, PFupdate, allows new genomes to be scanned against the protein families and matching sequences are merged into the families using conservative thresholds (Marsden *et al.* 2006). Currently, between two-thirds and three quarters of sequences in new genomes can be assigned to existing protein families in Gene3D, some matching more than one family, suggesting more distant relationships. The remaining sequences are unique to the genome and give rise to new Gene3D protein families or remain as singletons. Information on the protein families in each genome and their domain composition is available from the Gene3D website (see above).

To improve functional annotations for each protein family in Gene3D, sequences from the COG (Tatusov *et al.* 2001), GO (<http://www.geneontology.org>), EC (Webb 1992), KEGG (Kanehisa *et al.* 2004), DIP (Salwinski *et al.* 2004), BIND (Alfarano *et al.* 2005),

Figure 7. Average proportions of genome sequences in common protein (a) and domain (b) families.

and MINT (Zanzoni *et al.* 2002) databases were scanned against the genome sequences and clustered using the PFupdate protocol. These resources provide a range of information on biochemical function, cellular process, biochemical pathways and protein-protein interactions. Information on the protein family, domain compositions and protein functions is also available from the Sesami website (Marsden *et al.* 2006), which has been developed to aid selection of domain families for structure determination. Structurally uncharacterized domains in Pfam or NewFam can be chosen using criteria based on species distribution, family size and probable functions. Multiple representatives from large domain families can be selected depending on the different multi-domain contexts in which they occur.

(d) Protein and domain family distributions in the genomes

With the genome sequences organized into protein families, we can compare across the species to identify common or unique protein families. On average, only 10% of sequences in each genome, depending on the organism, belong to protein families that are common to species from all kingdoms of life. Again, common families are those that can be detected in at least 70% of species in each kingdom, a threshold chosen to reflect the sensitivities of the HMM technologies used to search for remote homologues. Nearly one-half of the sequences in each genome are unique to the kingdom in which the organism belongs and between 10 and 20% of the remaining sequences are unique to the organism (see figure 7a). Rost has shown that a significant proportion of these remaining singletons are small sequences, which tend to have low predicted secondary structure, and may be regulatory sequences that adopt conformations on binding specific proteins.

In contrast to the protein family distribution, the perspective for domain distribution appears rather different. Using the domain mappings to structurally characterized families in CATH, as this allows us to detect more ancient relationships, between 60 and 70% of domain structure annotations in genomes are common to all kingdoms of life (see figure 7b). Less

Figure 8. Distribution by fold of (a) CATH (b) 203 completed genomes. The angles subtended by the sectors correspond to the frequency of occurrence of the protein folds, measured by the number of close sequence families within each fold group.

than 20% are unique to the kingdom or individual species. These common domains, previously identified from our preliminary analysis of LUCA (see above), are some of the most highly populated domain families in CATH. About one-third of them adopt one of the superfolds, (Orengo *et al.* 1994), which are clearly dominating the genome sequences (figure 8).

A significant proportion of the large domain families common to all kingdoms of life are associated with metabolism or regulation and perform generic functions (e.g. providing energy or redox equivalents, binding to DNA). The size of the family correlates well with the number of different domain partners and different functions exhibited by relatives in the family.

Our data, therefore, suggests that although most sequences in a genome are from 'protein' families unique to the organism or kingdom to which the organism belongs, a large proportion of the 'domains' within these proteins are common to all kingdoms of life. These observations support a mosaic theory of protein evolution suggested by Teichmann and co-workers (Apic *et al.* 2001), whereby common domains have been extensively duplicated during evolution and shuffled in the genomes. Functions can be modified by changes in domain partners, often because this modulates the geometry of the active site and thereby the nature of the substrate binding there (Todd *et al.* 2001).

7. STRUCTURAL ANALYSIS OF LARGE DOMAIN FAMILIES WILL REVEAL MECHANISMS FOR MODIFYING FUNCTIONS AND EXPANDING THE FUNCTIONAL REPERTOIRE OF ORGANISMS

Less than 2000 domain families account for 70% of domain family annotations in the genomes (see figure 9) of which 1266 belong to structurally uncharacterized families in Pfam. Although the remaining 734 families contain one or more structural representatives, for

Figure 9. Coverage of protein sequences from 203 completed genomes by CATH, Pfam and NewFam families, ordered by size on the x-axis. The largest 2000 families cover approximately 70% of domain sequences.

many of the sequences in these families there is no structural relative close enough to use as a parent in homology modelling. Calculations using Gene3D data show that in order to ensure that families have enough structural relatives to build homology models for all the remaining relatives, at least 90 000 domain sequences would need to be targeted for structure determination. Although, there are several international structural genomics initiatives, and though the NIH funded protein structure initiative (PSI, see Burley *et al.* 2000; Brenner 2001) is about to be renewed for a further 5 years, it is unlikely that they will solve this number of structures in a reasonable time. Therefore, we need to rationalize our target selection further and perhaps select those representatives that provide most information on the range of functions found across the family.

Since a change in domain partnership can clearly modify function, the information in Gene3D can be used to select different multi-domain contexts for a targeted domain, choosing relatives that are likely to

Figure 10. The correlation between the number of COG functional annotations and the number of structural sub-groups for each CATH superfamily. The data points are coloured by the percentage coverage of functional annotation across sequence space within a superfamily. For example, 0–25 represents those superfamilies where less than 25% of the sequence families contributed to a COG annotation, so the sample can be seen to be sparse, in comparison to 75–100, where the functional sampling is near complete across sequence space.

occupy different regions of ‘function’ space because their multi-domain contexts differ.

Previous analyses of function in CATH enzyme families of known structure showed that in the absence of information on multi-domain context, domains from multi-domain proteins could only be assumed to possess similar functions provided they shared high sequence similarity ($\geq 60\%$ identity; Todd *et al.* 2001). However, domain annotation information in Gene3D now allows us to classify domain sequences according to their multi-domain contexts in Gene3D. Revisiting our analysis of function conservation with this data, we observe that provided domain composition is conserved, domains having quite low sequence identity, down to 20%, frequently share common functions. Thus sub-classification of domain families, by their domain partnerships will help to rationalize target selection in large, functionally promiscuous families.

Many large families in CATH, highly populated with genome relatives, are observed to be structurally very diverse. Functional diversity in these families correlates well with structural diversity (see figure 10). We have used a number of measures to analyse structural diversity in CATH families (Reeves *et al.* in preparation), including percentages of residues with conserved structural environments and variations in secondary structure separations and orientations. We have also developed a new algorithm, 2DSEC, to identify common secondary structures and characterize the degree to which the common structural core has been embellished by secondary structure insertions. Analysis of all structurally variable families in CATH revealed that although these insertions are frequently discontinuous in sequence, they are often co-located in three-dimension. In some cases they assemble close to the active site where they modify the geometry and thus

influence substrate binding. In other cases, embellishments influence surface geometry and properties, causing changes in domain partnerships and/or protein–protein associations. Changes in oligomerization state are also mediated by these effects.

Therefore, large differences in the secondary structure compositions of relatives may be a useful indicator of significant changes in domain or protein associations and, therefore, modifications in function. To help in further rationalizing target selection, reducing the estimate of 90 000 domain sequences quoted above, secondary structure predictions can be made for domain sequences in each targeted domain family to identify subsets of sequences with significant differences in the numbers of secondary structures predicted.

We have used the P-loop hydrolases to explore this strategy for target selection. This is an extremely large superfamily with 7943 sequence families in Gene3D (sharing more than 30% sequence identity) and with 270 different COGs functions and 40 different enzyme classifications currently known. A simple target selection protocol would target a representative from each sequence family (relatives clustered at 30% sequence identity) to ensure that good homology models could be built for all relatives in the family. However, only 4328 different domain combinations are identified across these families, reducing the number of targets about twofold. However, it would be hoped that the careful identification of distinct functional groups could further reduce number of targets to approximately 310.

It must also be noted that if the structures generated through the structural genomics projects are intended to be used as homology modelling templates, enabling genomic-scale levels of structural and functional annotations, it is vital that we first provide functional characterization of each newly solved structure. With this in mind it is essential that structural genomics projects work closely with the wider experimental community, expert in assigning function to proteins, to provide and validate the functional characterizations of newly solved proteins.

8. CONCLUSION

The wealth of genome data now available, combined with expansions in the fold library thanks to structural genomics and improvements in the technologies for detecting distant homologues, is giving fascinating insights into how domain and protein families have evolved. We are still limited in our ability to trace all ancestral domains but by carefully organizing the sequence data from completed genomes into domain and protein families we can begin to explore how divergence in these families enriches the functional repertoire of an organism. In the future, these insights can help guide the selection of representative domain sequences from structurally uncharacterized families so that by acquiring this three-dimensional-data we will glean more profound insights into the structural mechanisms by which new domain partnerships and functions evolve.

R.L.M. and D.L. were supported by the NIH funded PSI structural genomics initiative, J.A.G.R. by the EU funded temblor grant. A.S. was supported by the Instituto de Salud Carlos III, RMN C03/08. O.R. and G.A.R. by the BBSRC, C.Y. by the EU funded Biosapiens project, M.M. was funded by Wellcome and S.A. and C.A.O. by the MRC.

REFERENCES

- Alfarano, C. *et al.* 2005 The biomolecular interaction network database and related tools 2005 update. *Nucleic Acids Res.* **33**, D418–D424. 33 Database issue.
- Aloy, P., Stark, A., Hadley, C. & Russell, R. B. 2003 Predictions without templates: new folds, secondary structure and contacts in CASP5. *Proteins* **53**, 436–456. (doi:10.1002/prot.10546)
- Andreeva, A., Howorth, D., Brenner, S. E., Hubbard, T. J., Chothia, C. & Murzin, A. G. 2004 SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.* **32**, D226–D229. (doi:10.1093/nar/gkh039)
- Apic, G., Gough, J. & Teichmann, S. A. 2001 Domain combinations in archaeal, eubacterial and eukaryotic proteins. *J. Mol. Biol.* **310**, 311–325. (doi:10.1006/jmbi.2001.4776)
- Apweiler, R. *et al.* 2004 UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* **32**, D115–D119. (doi:10.1093/nar/gkh131)
- Bateman, A. *et al.* 2004 The Pfam protein families database. *Nucleic Acids Res.* **32**, D138–141. (doi:10.1093/nar/gkh121)
- Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Wheeler, D. L. 2005 Genbank. *Nucleic Acids Res.* **33**, D34–D38. (doi:10.1093/nar/gki063)
- Bray, J. E., Marsden, R. L., Rison, S. C., Savchenko, A., Edwards, A. M., Thornton, J. & Orengo, C. A. 2004 A practical and robust sequence search strategy for structural genomics target selection. *Bioinformatics* **20**, 2288–2295. (doi:10.1093/bioinformatics/bth240)
- Brenner, S. E. 2001 A tour of structural genomics. *Nat. Rev. Genet.* **2**, 801–809. (doi:10.1038/35093574)
- Buchan, D. W., Shepherd, A. J., Lee, D., Pearl, F. M., Rison, S. C., Thornton, J. M. & Orengo, C. A. 2002 Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. *Genome Res.* **12**, 503–514. (doi:10.1101/gr.213802)
- Burley, S. K. 2000 An overview of structural genomics. *Nat. Rev. Genet.* **2**, 801–809.
- Chothia, C. & Lesk, A. M. 1986 The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.
- Chothia, C., Gough, J., Vogel, C. & Teichmann, S. A. 2003 Evolution of the protein repertoire. *Science* **300**, 1701–1703. (doi:10.1126/science.1085371)
- Corpet, F., Servant, F., Gouzy, J. & Kahn, D. 2000 ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res.* **28**, 267–269. (doi:10.1093/nar/28.1.267)
- Deshpande, N. *et al.* 2005 The RCSB protein data bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res.* **33**, D233–D237. (doi:10.1093/nar/gki057)
- Durbin, R., Eddy, S., Krogh, A. & Mitchison, G. 1998 *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.
- Eddy, S. R. 1998 Profile hidden Markov models. *Bioinformatics* **14**, 755–763. (doi:10.1093/bioinformatics/14.9.755)
- Enright, A. J., Van Dongen, S. & Ouzounis, C. A. 2002 An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30**, 1575–1584. (doi:10.1093/nar/30.7.1575)
- Enright, A. J., Kunin, V. & Ouzounis, C. A. 2003 Protein families and TRIBES in genome sequence space. *Nucleic Acids Res.* **31**, 4632–4638. (doi:10.1093/nar/gkg495)
- Fleming, K., Muller, A., MacCallum, R. M. & Sternberg, M. J. 2004 3D GENOMICS: a database to compare structural and functional annotations of proteins between sequenced genomes. *Nucleic Acids Res.* **32**, D245–D250. (doi:10.1093/nar/gkh064)
- Frishman, D. *et al.* 2003 The Pedant genome database. *Nucleic Acids Res.* **31**, 207–211. (doi:10.1093/nar/gkg005)
- Gibrat, J. F., Madej, T. & Bryant, S. H. 1996 Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**, 377–385. (doi:10.1016/S0959-440X(96)80058-3)
- Gough, J., Karplus, K., Hughey, R. & Chothia, C. 2001 Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.* **313**, 903–919. (doi:10.1006/jmbi.2001.5080)
- Grant, A., Lee, D. & Orengo, C. 2004 Progress towards mapping the universe of protein folds. *Genome Biol.* **5**, 107. (doi:10.1186/gb-2004-5-5-107)
- Griffiths-Jones, S. & Bateman, A. 2002 The use of structure information to increase alignment accuracy does not aid homologue detection with profile HMMs. *Bioinformatics* **18**, 1243–1249. (doi:10.1093/bioinformatics/18.9.1243)
- Harrison, A., Pearl, F., Mott, R., Thornton, J. & Orengo, C. 2002 Quantifying the similarities within fold space. *J. Mol. Biol.* **323**, 909–926. (doi:10.1016/S0022-2836(02)00992-0)
- Horowitz, N. H. 1945 On the evolution of biochemical synthesis. *Proc. Natl Acad. Sci. USA* **31**, 153–157.
- Jensen, R. A. 1976 Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409–425. (doi:10.1146/annurev.mi.30.100176.002205)
- Jones, D. T. 1999 GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* **287**, 797–815. (doi:10.1006/jmbi.1999.2583)
- Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. & Hattori, M. 2004 The KEGG resources for deciphering the genome. *Nucleic Acids Res.* **32**, D277–D280. (doi:10.1093/nar/gkh063)
- Karplus, K., Barrett, C. & Hughey, R. 1998 Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856. (doi:10.1093/bioinformatics/14.10.846)
- Kelley, L., MacCallum, R. & Sternberg, M. 2000 Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**, 499–522. (doi:10.1006/jmbi.2000.3741)
- Kolodny, R., Koehl, P. & Levitt, M. 2005 Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.* **346**, 1173–1188. (doi:10.1016/j.jmb.2004.12.032)
- Koonin, E. 2003 Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* **1**, 127–136. (doi:10.1038/nrmicro751)
- Koonin, E. V., Wolf, Y. I. & Karev, G. P. 2002 The structure of the protein universe and genome evolution. *Nature* **420**, 218–223. (doi:10.1038/nature01256)
- Krause, A., Stoye, J. & Vingron, M. 2000 The SYSTERS protein sequence cluster set. *Nucleic Acids Res.* **28**, 270–272. (doi:10.1093/nar/28.1.270)
- Kurland, C. G., Canback, B. & Berg, O. G. 2003 Horizontal gene transfer: a critical view. *Proc. Natl Acad. Sci. USA* **100**, 9658–9662. (doi:10.1073/pnas.1632870100)
- Lee, D., Grant, A., Marsden, R. L. & Orengo, C. 2005 Identification and distribution of protein families in 120 completed genomes using Gene3D. *Proteins Struct. Funct. Bioinform.* **59**, 603–615.

- Liu, J. & Rost, B. 2001 Comparing function and structure between entire proteomes. *Protein Sci.* **10**, 1970–1979. (doi:10.1110/ps.10101)
- Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C. & Gough, J. 2004 The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.* **32**, D235–D239. (doi:10.1093/nar/gkh117)
- Marsden, R. L., Lee, D., Maibaum, M., Yeats, C. & Orengo, C. A. 2006 Comprehensive analysis of 203 structural genomics with provides structural genomics with new insights into protein family space. *Nucleic Acids Res.* Submitted.
- McGuffin, L. J., Street, S., Sorensen, S. A. & Jones, D. T. 2004 The genomic threading database. *Bioinformatics* **20**, 131–132. (doi:10.1093/bioinformatics/btg387)
- Mira, A., Ochman, H. & Moran, N. A. 2001 Deletion bias and the evolution of bacterial genomes. *Trends Genet.* **10**, 589–596. (doi:10.1016/S0168-9525(01)02447-7)
- Mitchell, E. M., Artymiuk, P. J., Rice, D. & Willett, P. 1990 Use of techniques derived from graph theory to compare secondary structure motifs in proteins. *J. Mol. Biol.* **212**, 151–166. (doi:10.1016/0022-2836(90)90312-A)
- Mulder, N. J. *et al.* 2003 The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31**, 315–318. (doi:10.1093/nar/gkg046)
- Nagl, S. 2003 Profile based methods for sequence analysis. In *Bioinformatics genes, proteins & computers* (ed. C. A. Orengo, D. T. Jones & J. M. Thornton). Oxford, UK: BIOS.
- Orengo, C. A. & Thornton, J. M. In press. Protein families and their evolution—a structural perspective.
- Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. 1993 Identification and classification of protein fold families. *Protein Eng.* **6**, 485–500.
- Orengo, C. A., Jones, D. T. & Thornton, J. M. 1994 Protein superfamilies and domain superfolds. *Nature* **372**, 631–634. (doi:10.1038/372631a0)
- Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. 1997 CATH—a hierarchical classification of protein domain structures. *Structure* **5**, 1093–1108. (doi:10.1016/S0969-2126(97)00260-8)
- Orengo, C. A., Sillitoe, I., Reeves, G. & Pearl, F. M. 2001 Review: what can structural classifications reveal about protein evolution? *J. Struct. Biol.* **134**, 145–165. (doi:10.1006/jsbi.2001.4398)
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. 1998 Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201–1210. (doi:10.1006/jmbi.1998.2221)
- Pearl, F. M. *et al.* 2001 A rapid classification protocol for the CATH domain database to support structural genomics. *Nucleic Acids Res.* **29**, 223–227. (doi:10.1093/nar/29.1.223)
- Pearl, F. *et al.* 2005 The CATH domain structure database and related resources Gene3D and DHS provide comprehensive domain family information for genome analysis. *Nucleic Acids Res.* **33**, D247–D251. (doi:10.1093/nar/gki024)
- Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D. & Yeates, T. O. 1999 Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA* **96**, 4285–4288. (doi:10.1073/pnas.96.8.4285)
- Ranea, J. A., Buchan, D. W., Thornton, J. M. & Orengo, C. A. 2004 Evolution of protein superfamilies and bacterial genome size. *J. Mol. Biol.* **336**, 871–887. (doi:10.1016/j.jmb.2003.12.044)
- Ranea, J. A., Buchan, D. W. A., Thornton, J. M. & Orengo, C. A. 2005 Factory economics can explain optimum genome size in bacteria. *Trends Genet.* **21**, 21–25. (doi:10.1016/j.tig.2004.11.014)
- Redfern, O., Bennett, C. & Orengo, C. 2004 Structural classifications of proteins. In *Bioinformatics encyclopaedia* (ed. R. Apweiler). London: Wiley.
- Redfern, O., Bennett, C. & Orengo, C. 2005 Survey of current protein family databases. *J. Chromatogr. A* **815**, 97–107. (doi:10.1016/j.jchromb.2004.11.010)
- Reeves, G. A., Dallman, T., Redfern, O. & Orengo, C. A. In preparation. Structural diversity of domain superfamilies in the CATH database.
- Rison, S., Teichmann, S. & Thornton, J. 2002 Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*. *J. Mol. Biol.* **318**, 911–932. (doi:10.1016/S0022-2836(02)00140-7)
- Rost, B. 1997 Protein structures sustain evolutionary drift. *Fold. Des.* **2**, S19–S24. (doi:10.1016/S1359-0278(97)00059-X)
- Sali, A. 1998 100,000 protein structures for the biologist. *Nat. Struct. Biol.* **5**, 1029–1032. (doi:10.1038/4136)
- Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. 2004 The database of interacting proteins: 2004 update. *Nucleic Acids Res.* **32**, D449–D451. (doi:10.1093/nar/gkh086)
- Sasson, O., Vaaknin, A., Fleischer, H., Portugaly, E., Bilu, Y., Linial, N. & Linial, M. 2003 ProtoNet: navigating the hierarchical clustering of the protein space. *Nucleic Acids Res.* **31**, 348–352. (doi:10.1093/nar/gkg096)
- Shindyalov, I. N. & Bourne, P. E. 1998 Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739–747. (doi:10.1093/protein/11.9.739)
- Sillitoe, I. & Orengo, C. 2003 Protein structure comparison. In *Bioinformatics genes, proteins & computers* (ed. C. A. Orengo, D. T. Jones & J. M. Thornton). Oxford, UK: BIOS.
- Sillitoe, I., Dibley, M., Bray, J., Addou, S. & Orengo, C. 2005 Assessing strategies for improved superfamily recognition. *Protein Sci.* **14**, 1800–1810.
- Tatusov, R. L. *et al.* 2001 The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22–28. (doi:10.1093/nar/29.1.22)
- Taylor, W. R. & Orengo, C. A. 1989 Protein structure alignment. *J. Mol. Biol.* **208**, 1–22. (doi:10.1016/0022-2836(89)90084-3)
- Teichmann, S. A. & Babu, M. M. 2004 Gene regulatory growth by duplication. *Nat. Genet.* **36**, 492–496. (doi:10.1038/ng1340)
- Todd, A. E., Orengo, C. A. & Thornton, J. M. 2001 Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143. (doi:10.1006/jmbi.2001.4513)
- Todd, A. E., Orengo, C. A. & Thornton, J. M. 2002 Plasticity of enzyme active sites. *Trends Biochem. Sci.* **27**, 419–426. (doi:10.1016/S0968-0004(02)02158-8)
- Todd, A. E., Marsden, R. L., Thornton, J. M. & Orengo, C. A. 2005 Progress of structural genomics initiatives: an analysis of solved target structures. *J. Mol. Biol.* **348**, 1235–1260. (doi:10.1016/j.jmb.2005.03.037)
- Van Dongen, S. 2000 Graph clustering by flow simulation. Ph.D. thesis, University of Utrecht. (<http://www.library.uu.nl/digiarchief/dip/diss/1895620/inhoud.htm>)
- Velankar, S., McNeil, P., Mittard-Runte, V., Suarez, A., Barrell, D., Apweiler, R. & Henrick, K. 2005 E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res.* **33**, D262–D265. (doi:10.1093/nar/gki058)
- Venter, J. C. *et al.* 2004 Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74. (doi:10.1126/science.1093857)
- Vogel, C., Teichmann, S. A. & Chothia, C. 2003 The immunoglobulin superfamily in *Drosophila melanogaster*

- and *Caenorhabditis elegans* and the evolution of complexity. *Development* **130**, 6317–6328. (doi:10.1242/dev.00848)
- Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C. & Teichmann, S. A. 2004a Structure, function and evolution of multidomain proteins. *Curr. Opin. Struct. Biol.* **14**, 208–216. (doi:10.1016/j.sbi.2004.03.011)
- Vogel, C., Berzuini, C., Bashton, M., Gough, J. & Teichmann, S. A. 2004b Supra domains: evolutionary units larger than single domain protein domains. *J. Mol. Biol.* **336**, 809–823. (doi:10.1016/j.jmb.2003.12.026)
- Webb, E. C. 1992 *Enzyme nomenclature*, Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology. New York: Academic Press.
- Woese, C. 1998 The universal ancestor. *Proc. Natl Acad. Sci. USA* **95**, 6854–6859. (doi:10.1073/pnas.95.12.6854)
- Yona, G. & Levitt, M. 2002 Within the twilight zone: a sensitive profile-profile comparison tool based on information theory. *J. Mol. Biol.* **315**, 1257–1275. (doi:10.1006/jmbi.2001.5293)
- Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. & Cesareni, G. 2002 MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140. (doi:10.1016/S0014-5793(01)03293-8)