The CATH protein family database: A resource for structural and functional annotation of genomes

Over the last decade, there have been huge increases in the numbers of protein sequences and structures determined. In parallel, many methods have been developed for recognising similarities between these proteins, arising from their common evolutionary background, and for clustering such relatives into protein families. Here we review some of the protein family resources available to the biologist and describe how these can be used to provide structural and functional annotations for newly determined sequences. In particular we describe recent developments to the CATH domain database of protein structural families which have facilitated genome annotation and which have also revealed important caveats that must be considered when transferring functional data between homologous proteins.

Keywords: Protein structure classification / Structural genomics / Sequence comparison / Structure comparison / Functional annotation / Review

1 Introduction

The international genome sequencing projects have expanded the sequence databases considerably and there are now more than half a million non-redundant sequences deposited in GenBank [1] and more than 50 complete genomes sequenced. Although a large proportion of these are from microbial organisms, several eukaryote genomes have also been determined, including the fly (Drosophila melanogaster), the worm (Caenorhabditis elegans), and yeast (Saccharomyces cerevisiae). The human genome is almost complete and it is likely that the mouse genome will be finished within the next few years. These data and the promising new technologies associated with transcriptomics and proteomics will enable biologists to revolutionise their research by investigating whole organisms rather than focusing on single proteins or families of proteins.

One of the major challenges in this post genome era will be to assign functional properties and biological roles to all these sequenced genes. Bioinformatics will play an important role through the design of protocols for inheriting such functional properties from evolutionarily related and experimentally characterised proteins, thereby
obviating the need for direct characterisation or enabling more directed experiments to be designed. This is because structural and functional properties of evolutionary related proteins are often very well conserved and many research groups have now developed classification methods for clustering relatives into protein families in order to facilitate the annotation of new relatives identified in the genomes.

1.1 Sequence databases

There are several protein family databases based largely on sequence data (e.g. Pfam [2], PRINTS [3], PROSITE [4], SMART [5]) and constructed by comparing proteins deposited in the sequence databanks (e.g. GenBank [1], SWISS-PROT [6], TrEMBL [6]). These resources use a variety of different methods for detecting relatives. For example, the Pfam database contains ~2800 protein sequence families characterised by specific profiles generated using Hidden Markov Models [7], see below). Other resources apply 1-D sequence profiles or more flexible pattern recognition methods for detecting relatives (see [8] for a review). A recent initiative (InterPro, [9]) seeks to integrate these databases, providing common views of the families thereby enabling biologists to benefit more readily from all the available functional annotations provided by each. Sequence based protein family databases are currently much more extensive than those constructed from structural data, since experimental methods for determining protein structures are still considerably more challenging than sequence determination. In contrast to the GenBank sequence databank with its 12 million entries, the protein structure databank (PDB) now held at Rutgers University in the United States [10], contains only ~23,000 protein chains.

However, structural genomics initiatives now being funded internationally, should help to reduce the gap between sequence and structure data. These initiatives aim to determine representative structures for all known protein sequence families. Furthermore, since a protein’s structure is much more highly conserved than its sequence, fewer structural representatives will be required within each protein family. Improvements in technologies used for structure determination and international collaboration on target selection [11, 12] should hopefully provide much of these data within the next five years.

1.2 Structure databases

There are also several structural classifications, assembled using a variety of different philosophies for recognising structural relatives. The SCOP database [13, 14] contains 900 evolutionary protein families or homologous superfamilies and is largely constructed by careful manual evaluation of structural relationships, whilst the CATH database [15, 16] employs a combination of automated and manual approaches (see below). CATH currently contains 1200 evolutionary superfamilies. A functional dictionary (DHS [17]) collates functional information for these superfamilies extracted electronically from various sources. Other resources include the HOMSTRAD [18] and CAMPASS [19] databases which contain functional information for nearly 850 protein superfamilies extracted from the SCOP database together with alignments for each. The 3D Database resource [20] also has domain structure families, again assembled using a variety of manual and automated protocols. In contrast, the DALI Domain Dictionary [21] and the VAST structural neighbours list [22] are generated completely automatically and do not specifically cluster related proteins into structural families, yet they provide excellent facilities for searching for relatives.

Although the structural resources are less complete than the sequence classifications, they provide additional information. Since structure is much more highly conserved during evolution than the sequence, these resources can identify very distant relatives, undetected in the sequence based resources. Furthermore, most of the structural databases are domain based and again this can facilitate the recognition of distant relationships. Domain boundaries are much more readily identified from structural data, either by manual inspection or using automated approaches [23]. The domain has long been proposed as an important evolutionary unit and recent analyses of some completed genomes have indeed suggested that domain recurrence is even more extensive than previously thought, with 52% of the domains in the Mycoplasma genitalium genome recur- ring, whilst in the C. elegans genome an even higher proportion 96% recur (see [24] for a review). Thus it is highly likely that domain duplication and shuffling within the genome are important mechanisms for expanding the functional repertoire of an organism.

Structural data also provides essential insights into the biochemical mechanism of a protein’s biological role not readily obtained from the sequence alone. Many algorithms have been written for identifying those protein residues in contact with ligands or substrates within the active site (see for example LIGPLOT [25], TESS [26], ASSAM [27], Relibase [28]). A recent resource (PROCAT, Porter, C., Thornton, J. M., personal communication) provides a catalogue of over 178 active site templates describing the specific constellations of active site residues essential for catalytic activity, identified using a robust approach based on a geometric hashing algorithm.
1.3 Improvements in recognising distant homologues

Significant advances in bioinformatic approaches over the last few years have been based on the development of powerful new methods for recognising evolutionary relationships even in very distant homologues. For example, profile based sequence search methods such as PSI-BLAST [29] can now recognise up to one-third of distant evolutionary relationships previously only detectable by comparing the 3-D structures of the proteins [30]. These approaches have been designed for database searching and iteratively identify increasingly distant sequence relatives. Following each scan of the database a multiple alignment of relatives is constructed and used to generate a sequence pattern or profile for the family describing specific residue preferences at each position in the protein. Similarly, Hidden Markov Models have also been used to characterise sequence preferences for specific protein families [2, 7, 31]. These use probabilistic approaches for modelling allowed residue substitutions, insertions and deletions. In a recent competition held every two years in the US, such approaches have been shown to be amongst the most sensitive currently available (CASP, Competitive Assessment of Structure Prediction [32]). However, the most powerful recognition methods for mapping between structures and sequences and therefore for finding structural relatives to genome sequences, are those based on threading (see [8] for review). These approaches include empirical potentials describing the preferences for specific residue pairs to be in contact (e.g., Threader [33]). Threading algorithms tend to be much more computationally expensive and therefore not so widely used for database scans.

In recent attempts to assign structural data to genome sequences between 40–50% of small bacterial genomes could be assigned to known structural families in the SCOP or CATH databases (see [24] for a review), although in many cases only part of the gene could be matched. On a residue basis, the coverage is closer to 30–40% (Buchan, D., personal communication). The Ensembl database (http://www.ensembl.org) based at the European Bioinformatics Institute in Cambridge, UK, provides structural and functional annotations for genes in the human genome, although again this represents a smaller percentage of residues assigned. Below we review recent developments of the CATH structural domain database, involving both new protocols for classifying known structures and robust strategies for recognising sequence relatives to these families in the genomes.

Expansion of the database with genome sequences both improved the sensitivity of the methods used for recognising homologues and also enabled considerable expansion of the functional dictionaries compiled for each homologous superfamily. Analyses of functional variation in enzyme superfamilies in CATH subsequently revealed substantial divergence of function in some and has provided general guidelines for reliable inheritance of functional characteristics within these families.

2.1 Overview of the CATH domain database

As with other major structural databases, the first level in the CATH hierarchy describes the protein (C)lass; that is whether the structure comprises mainly \( \alpha \)-helices, mainly \( \beta \)-strands or a mixture of both. In SCOP this latter class is divided into alternating \( \alpha \)/\( \beta \) proteins and those proteins containing segregated \( \alpha \) and \( \beta \) regions. However, there is considerable overlap between the two as many proteins consist of both alternating regions and clearly separated \( \alpha \) and \( \beta \) regions and in CATH, these divisions become more apparent lower in the hierarchy. At the next (A)rchitectural level, proteins are grouped according to the orientations of their secondary structures in 3-D. A large proportion of structures adopt very simple layered architectures such as sandwiches (e.g., two or three-layer \( \alpha \)-/\( \beta \) proteins (Fig. 1)) or barrel-like arrangements. In the current version of CATH, nearly 50% of the superfamilies adopt five simple and rather symmetric architectures (Fig. 3A). The (T)opology level or fold group then discriminates according to differences in the connectivities between the secondary structures in these architectures. The three-layer \( \alpha \)-/\( \beta \) architecture is one of the most favoured architectures and there are currently 71 different topologies observed within it, although one of these (the Rossmann fold, see Fig. 3B) appears to be particularly favoured. The architectural level is unique to CATH and is assigned by manual inspection of the structures when it cannot be inferred by homology. However, other databases, such as SCOP, often describe architectural features for some individual protein families.

Within each fold group, proteins are classified according to whether they appear to be related by divergent evolution. Relationships are established using a variety of automated algorithms and protocols for manual validation. These analyse sequence, structural and functional data to infer homology and are described below. However, over the last five years there have been an increasing number of analogous structures detected which have similar folds but no other evidence to suggest homology. Several groups have suggested that the structural simi-
Similarity between these analogous proteins arises due to physical-chemical constraints on the packing of \( \alpha \)-helices and \( \beta \)-strands in 3-D and that consequently there may be a limited number of folds available in nature. Estimates range from about one thousand folds [34] to several thousand folds [35, 36]. However, the relatively high proportion of orphan genes currently observed in the genomes may suggest that the larger estimates are more accurate.

Analyses of both the SCOP and CATH databases have revealed that some folding arrangements are very highly favoured, described as superfolds in CATH [15, 16] and frequently occurring domains (FODS) in SCOP [37]. In CATH, 20\% of the homologous superfamilies are found to adopt as few as five highly recurring superfolds (Fig. 2). Recent calculations (Moult, personal communication) which take account of these biases in the structure databases and also the number of known sequence families have suggested that up to 90\% of protein superfamilies may adopt fewer than 500 folds with the remaining superfamilies resulting in a large tail of several thousand "unifolds". Analyses of the number of unique folds identified annually certainly suggest that the number of major fold groups known is close to saturating, with only 10\% of new structures found to have novel folds over the last year [38]. However, there are undoubtedly biases in both the structure and sequence databases, reflecting particular biological and medical interests and, in the structural resources, the difficulties involved in crystallising or solving certain classes of proteins (e.g. membrane proteins).

CATH Version 2.3 (November 2000) currently contains 27 000 domain entries which have been classified into 28 regular architectures, with 37\% of the superfamilies assigned to complex architectural bins as their architecture was too irregular to be classified. Nearly 700 fold groups are currently identified containing 1200 homologous superfamilies and 1800 sequence families are clustered at 35\% sequence identity. Figure 3a shows the number of superfamilies (245) adopting the top five superfolds. Figure 3b shows that nearly half the fold groups comprise simple regular architectures, in particular two-layer \( \alpha \beta \) sandwiches, three-layer \( \alpha \beta \) sandwiches, two-layer \( \beta \) sandwiches, \( \alpha \)-bundles and \( \beta \)-barrels.
2.2 Methods used for classifying new structures and sequences in the CATH domain database

2.2.1 Improvements in sequence based protocols

Structure genomic initiatives have promised thousands of new structures within the next five years. Already, nearly 200 new structures are added to the protein databank each month. Rapid protocols are needed to keep pace with this data. Since sequence based methods are much faster than comparing 3-D structures, often by two or three orders of magnitude, these techniques are applied first to recognise homologues. The recent improvements in profile based methods have enabled a much higher proportion of relatives to be identified [39].

The PSI-BLAST algorithm (see Section 1.3) is amongst the most powerful and easy to implement and we have recently optimised this method for database searching, to rapidly assign new homologues on the basis of sequence similarity. This involved benchmarking the performance obtained for different PSI-BLAST thresholds using a dataset of distant homologues clearly identified by structural methods. We selected 1351 representative structures from the CATH database, none of which had more than 35% sequence identity to one another, but of which many were evolutionary relatives clearly recognisable by structure comparison methods and which had been previously classified into 906 superfamilies in CATH. These relationships lie in the twilight zone of sequence similarity and simple pairwise based methods often fail to detect them [39]. By varying the thresholds for accepting PSI-BLAST matches, we were able to recognise 26% of the relatives, tolerating a small error rate of 0.06% per query.

A simple protocol (DomainFinder, [40]) was designed for identifying domain regions within the genes matched. Where several structural relatives matched a similar region on the gene sequence, the consensus residue range of the hit was recorded together with the minimum and maximum residue positions and the range for the best hit. Validation against domain boundaries assigned in CATH structurally, showed that in 71% of the domains identified, the boundaries identified by the best hit were optimal and were within 10 residues of the structural assignments.

A further protocol (Drange, Buchan et al., manuscript submitted) enabled resolution of conflicts between matches of different homologous superfamilies. Sometimes these were examples of very distant homologues which had diverged structurally quite considerably so that often only a small structural motif remained in common between the two. However, if functional properties had been conserved, there often remained sequence propensities sufficient for PSI-BLAST to detect the relationship. In other cases homologues had been assigned to the same fold group in CATH but to different superfamilies because the functions of the proteins were not sufficiently similar and no sequence similarity could be detected using pairwise methods.

Using these approaches it was possible to assign some 250 000 domain sequences from GenBank, into sequence and structural families in CATH, expanding the database nearly tenfold. Figure 4 shows that the sequence families fall into the three protein classes in roughly the same proportion as the sequence families derived from structural relatives in CATH. It can also be seen that only a small proportion (~15%) of these sequence families have a close structural homologue (> 35% sequence identity). Below this level of sequence
Figure 4. Expansion of the CATH database obtained by incorporating sequence relatives from GenBank. Relatives were identified using PSI-BLAST and the Domain-Finder protocol. Populations are shown for the few-secondary structures, mainly-alpha, mainly-beta and alpha/beta classes in CATH (V1.6). In light grey are the sequence families with sequences >35% sequence identity to a structural relative, in dark grey are the sequence families in which no member has a structural relative with >35% sequence identity.

It is often difficult to accurately predict the structure of the protein using homology modelling. However, even an approximate model of the protein fold, which can be generated using threading techniques, can often be useful in suggesting the location of the active site or in rationalising the effect of a residue mutation on the activity of the protein.

In order to extend the database still further, a related algorithm, IMPALA [41], was also benchmarked and applied. This takes the output from an iterative database search using PSI-BLAST and builds a multiple alignment and 1-D sequence profile from the final list of matches. Using IMPALA, we built a library of 8000 profiles for CATH, one for each non-identical structure in the database, including multidomain proteins and partially classified structures. It can be seen from Fig. 5 that the IMPALA profiles do not perform better than PSI-BLAST. However, scanning new structures against the IMPALA profiles is considerably faster than scanning the GenBank sequence database with PSI-BLAST. Therefore, the sequences of newly determined structures and also those from completed genomes, will be routinely scanned against the CATH IMPALA profiles, in order to regularly recruit new structures and gene relatives into the database.

Several other resources also link sequence and structure relatives in a similar way. For example the PDB-ISL [42], set up by the developers of SCOP and coworkers, provides lists of sequence relatives for each nonredundant structural entry in the SCOP database. These were identified using both BLAST and PSI-BLAST. In a related resource, SUPERFAMILY, gene relatives identified using Hidden Markov Models [31] have been explicitly integrated into homologous superfamilies in SCOP [13, 14]. The Entrez resource at the NCBI in the US [43] also provides lists of gene sequence relatives for each nonredundant structure in the Protein Databank identified using PSI-BLAST.

As well as providing additional annotations for gene sequences, explicitly integrating sequences into known structural families, in this way, significantly increases the sensitivity of methods such as PSI-BLAST, by exploiting established relationships. Figure 5 shows that over one-quarter of known structural relatives can be identified using PSI-BLAST if query sequences are scanned against a sequence library containing the nonredundant GenBank database and nonredundant sequences from the PDB. This proportion increases to over one half if the PDB sequences within the library are grouped into their appropriate superfamilies. Furthermore, by grouping GenBank sequences into these superfamilies, the sensitivity of the methods increases to 82% for PSI-BLAST and 76% for IMPALA.

A specific web site, Gene3D (Buchan et al., manuscript submitted), linked to CATH [15, 16], has been set up to provide structural and functional annotations for all the
completed genomes. This currently provides data for 38 complete genomes, 28 microbial, 8 archaeal and 2 eukaryotic. Protein family assignments are made using the DomainFinder and Drange algorithms (see Section 2.2.1) and for each gene within a given genome, protein family matches are highlighted and links provided to the relevant superfamly in CATH and Dictionary of Homologous Superfamilies (DHS). General statistics are also provided for each genome on the numbers and classes of all superfamily matches identified within the genome. Figure 6, constructed using all the data available in Gene3D, Version 1.0, shows the recurrence of CATH superfamilies in all the completed genomes. Interestingly, some of the superfamilies recur extensively, in some cases three orders of magnitude more frequently than other superfamilies. Most of these highly recurring superfamilies are those adopting superfolds, in particular the Rossmann fold and the triosephosphate isomerase (TIM) barrel fold, which are known to be extensively used by enzymes [44]. However, the predominance of these superfamilies in the PSI-BLAST matches also reflects to a certain extent their preponderance in CATH.

2.2.2 Improvements in structure based protocols

2.2.2.1 Domain fold recognition

Despite significant improvements in sequence based protocols, very distant evolutionary relatives can only be detected by comparing the protein structures. The CATH classification was initially established using a sensitive and reliable structure comparison method (SSAP, [45]) which had been optimised to recognise homologous proteins with a low error rate. However, although it is robust, the method is computationally very expensive as it uses dynamic programming to compare residue structural environments between proteins. Therefore, in order to cope with the flood of new structures anticipated from the international structure genome initiatives, we have developed a fast prescreen which identifies putative relatives which can then be accurately compared using the slower SSAP algorithm.

The new method (GRATH, (G)raphical (R)epresentation of (A)rchitecture, (T)opology and (H)omology) represents protein structures as simple graphs, where each node in the graph corresponds to secondary structures in the protein (i.e. β-strands or α-helices) and the edges between the nodes contain information on the distances and angles between the secondary structures (Harrison et al., manuscript submitted). The method is a very effective filter which can be used to screen the CATH library of representative structures with a newly determined structure. For all three protein classes, mainly-α, mainly-β or α-β it identifies the correct structural family within the top ten matches 98% of the time, with the top fold being identified as the best match 90% of the time.

2.2.2.2 Domain boundary identification

In order to facilitate domain boundary identification, a statistical approach has also been developed for GRATH, to assess the significance of a match. It is possible to calculate the frequency with which a particular score is obtained by chance based on the extreme value distribution obtained from database searches. This has enabled

---

**Figure 6.** Distribution of fold families and the repetition of their use as defined by the number of occurrences.
the method to be adapted to search for domain folds within multidomain structures. Nearly one-third of the protein structures in CATH contain multiple domains and the fact that nearly a third of these are discontinuous often makes it difficult to identify the domain boundaries correctly. Although there are many automated approaches for recognising boundaries, based mainly on chemical or physical properties (e.g. searching for large hydrophobic clusters, maximal intradomain residue contacts etc) most succeed only about 70–80% of the time (see [23] for a review). In CATH we have adopted a consensus approach seeking agreement between three different algorithms. However, consensus is reached in only about 10–15% of cases and a considerable amount of manual validation is required, measurably increasing the time required for classification of multidomain proteins.

Since domain duplication and shuffling occurs frequently within genomes (see above), many protein family classifications use the notion of domain recurrence to improve recognition of domain boundaries. This is particularly helpful for recognising boundaries when only sequence data is available (see MKDOM, [46]). In CATH, approximately 83% of domain folds currently recur and it is likely that this number will increase as more multidomain structures are solved with the improvements in crystallographic techniques. Therefore, we recently developed a simple iterative approach whereby a new multidomain structure is scanned against the CATH library of representative domain folds using GRATH. The best domain matches, determined statistically, are iteratively removed from the structure till three or fewer secondary structures remain. This prevents small domains from being split by the method and considerably reduces the error rate. The method was able to recognise 87% of domain folds in a nonredundant dataset of 345 multidomains, previously classified and manually validated in CATH (Harrison et al., manuscript submitted). Once the putative domain region has been recognised using GRATH, an accurate residue alignment is obtained using SSAP in order to obtain a more precise residue range for the domain.

2.2.2.3 3-D templates for homologous superfamilies

Although GRATH and SSAP can be used to identify the fold group for a new structural entry in CATH, these methods often have difficulty in distinguishing between homologous and analogous proteins. Homologues tend to have a higher degree of structural similarity than analogues, but this is not always the case for extremely distant relatives or for more structurally constrained folds (e.g. β-barrels). However, it has been shown that residue contacts within the core of the protein fold are often very specific for a particular homologous superfamily and are well conserved [47]. Therefore, in order to improve the recognition of evolutionary structural relatives, we have generated 3-D profiles or structural fingerprints for each superfamily in CATH, encoding information about conserved residue contacts (CORA, [48]). We have also developed a method for using these templates to recognise distant homologues (CONALIGN [49]). In some structurally, very diverse superfamilies multiple templates are required to represent all the structural subgroups within the family. Significant improvements can be obtained using the templates over searches with individual structures from the superfamily. A CATH server has been established (http://www.biochem.ucl.ac.uk/bsm/cath-new) which provides facilities for searching a new structure against the CATH library using GRATH, SSAP and CONALIGN and returns the structural family, together with links to other useful structural and functional resources (e.g. PDBsum [50], DHS [17]).

3 Insights into functional evolution from analysis of the CATH protein family database

Traditionally, bioinformatics techniques have been applied to detect evolutionary relationships between gene sequences in order to transfer functional information. This often involves searching protein family databases (e.g. Pfam, PRINTS, SCOP, CATH) for putative relatives. However, although orthologous gene relatives often retain similar functions because they are vital to the integrity of the organism, paralogous genes which arise from duplications of the orthologue, are free to evolve new functions. New functions can arise through various evolutionary mechanisms, such as mutations in the active site of an enzyme. These may enable the protein to bind a different substrate or to catalyse a different chemical reaction using the same substrate.

Over the last few years, sequence analyses on complete genomes have confirmed that domain duplication is a common event leading to evolution of new functions [24, 51]. Another important mechanism is domain shuffling between multidomain proteins. This can affect the structural environment of a domain by changing its domain partners, thereby modulating its function for example by affecting the accessibility of the active site. Figure 7 illustrates various mechanisms of evolving new functions.

Possible variations in function between paralogous domains should obviously be taken into account when transferring biochemical properties between relatives, especially as most protein family databases do not distin-
Methods of evolving new functions. In practice, new functions often evolve via a combination of mechanisms. In particular, gene duplication provides two identical copies of the same gene and one, free of functional constraints, can assume a new biological role through incremental mutations, gene fusion or oligomerisation. Note, however, that a change in oligomerisation state without prior duplication or mutation events can provide a route to an alternative gene function and this represents a method by which proteins moonlight (gene recruitment).

distinguish between orthologues and paralogues and techniques for determining these relationships are difficult, requiring complete genome data. We recently examined all the enzyme superfamilies in the CATH database, to determine how reliably functional properties could be inherited between relatives [51]. Enzyme superfamilies were chosen, as these are generally well annotated with consistent functional descriptions (i.e. from the Enzyme Commission (EC) classification). These descriptions can be accessed electronically via the Enzyme Database [52] and are associated with a numeric identifier which is very amenable to computational analysis. The EC classification contains four levels, the first of which describes the type of reaction, for example hydrolase or isomerase. Subsequent levels then depend to some extent on the type of reaction being described. However, enzymes which share the first three levels in the classification catalyse the same chemical reaction, although this may be performed on different substrates or involve different cofactors.

167 enzyme superfamilies were selected, having sufficient structural relatives, and the analyses were extended to include all the gene sequence relatives which could be detected for the superfamily using the PSI-BLAST protocols described in Section 2.2.1. Many CATH superfamilies are quite small, and expanding these superfamilies with gene relatives enabled a much more extensive analysis of functional properties within the superfamily. Figure 8 shows that only about one-third of these 167 superfamilies exhibit complete functional conservation of all four levels in the EC classification. In a further quarter, functions were conserved to the third EC level, implying similarity in reaction chemistry but variation in substrate or cofactor. However, in a significant proportion (26%) completely different functions were identified within the same superfamily. Extensive analysis of a selection of 31 of the most diverse of these superfamilies, revealed various features responsible for the functional changes.

In 80% of the 31 superfamilies examined, substrate specificity was diverse, in that the substrates bound varied in size, chemical properties and structural scaffolds. Interestingly, reaction chemistry was much more conserved than the substrate. In 93% of superfamilies some similarity in reaction chemistry, shared by all enzyme members, could be identified. In 75% of cases this similarity was limited to the initial reaction step or to

Figure 7. Methods of evolving new functions. In practice, new functions often evolve via a combination of mechanisms. In particular, gene duplication provides two identical copies of the same gene and one, free of functional constraints, can assume a new biological role through incremental mutations, gene fusion or oligomerisation. Note, however, that a change in oligomerisation state without prior duplication or mutation events can provide a route to an alternative gene function and this represents a method by which proteins moonlight (gene recruitment).

Figure 8. Conservation of enzyme function, as defined by EC number (x-axis), in homologous superfamilies. Shown in grey is the conservation of EC number between non-identical CATH representatives in 167 CATH homologous superfamilies containing two or more enzymes. In black is the conservation of EC number between all non-identical representatives within these superfamilies (i.e. sequence relatives also included). Enzymes having multiple or incomplete EC numbers are ignored. Not shown is the number of superfamilies containing both enzymes and nonenzymes. Of the 167 CATH superfamilies included in this chart, 13 contain nonenzymes and with the inclusion of sequence relatives, this number increases to 59.
stabilisation of a common reaction intermediate. Only in 7% of cases is the reaction chemistry completely unconserved. This is perhaps unsurprising, reflecting the difficulty in engineering a completely new chemistry compared to modifying the shape of the active site to suit a new substrate. A very high proportion, nearly 90% of superfamilies, contained examples of variation in domain organisation or of members existing as hetero-oligomers, reinforcing the notion of extensive domain shuffling as a useful mechanism for engineering new functions.

The correlation between pairwise sequence identity and functional conservation was also investigated using CATH superfamilies, combined with PSI-BLAST data, containing two or more enzymes. The conclusions are summarised in Table 1. For single domain proteins, there is a high probability of functional similarity for relatives sharing 40% or more sequence identity. However, for domains within multidomain proteins, a higher pairwise sequence identity (>60%) is required to ensure reliable functional annotation. Below 40% identity, the proportion of relatives with similar functional properties diminishes rapidly. Therefore, when considering very distant homologues, it is important to search for more specific sequence motifs to support the inheritance of functional information. Some databases (e.g. PRINTS) specifically identify sequence fingerprints associated with function. Various search methods are also available for detecting sequence or structural motifs characteristic of function. It is hoped that the accuracy of these types of approaches will improve over the next decade. Many databases display extensive functional annotations for families and superfamilies (e.g. InterPro, PRINTS). In CATH, a web-based Dictionary of Homologous Superfamilies [17] has been set up which displays information for each superfamily, extracted from a range of functional sources (e.g. SWISS-PROT [6], Enzyme Database [52] etc). This resource can be used to determine whether a particular superfamily exhibits a wide range of functions necessitating caution in functional inheritance or whether the superfamily appears to support a single function, even in very diverse relatives.

<table>
<thead>
<tr>
<th>Table 1. Confidence levels of functional inheritance with respect to pairwise sequence identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All domains</td>
</tr>
<tr>
<td>&gt; 96% accuracy in correct prediction of EC number</td>
</tr>
<tr>
<td>Seq ID &gt; 40% A.B.C.-prediction</td>
</tr>
<tr>
<td>Seq ID &gt; 30% A.B.C.-prediction</td>
</tr>
</tbody>
</table>

4 Concluding remarks

The development of powerful new techniques for recognising very distant homologues promises to reveal the biochemical properties of many sequences in the completed genomes. Protein family databases such as CATH, built from these methods, can therefore provide essential functional data for interpreting the experimental data being generated by transcriptomics and proteomics. However, it is clearly important to bear in mind the extent to which function can vary between paralogues when using this information to interpret the experimental data. Furthermore, resources describing protein-protein interactions and cellular pathways, not reviewed here, will also be crucial in unravelling the networks of interactions and functional complexes occurring in a given cellular context, in order to understand the phenotype of a particular gene.

We acknowledge the support from the Medical Research Council. AET and JEB are in receipt of studentships from the BBSRC.

Received May 29, 2001

5 References
