Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography B, 815 (2005) 97–107
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Review

Survey of current protein family databases and their application in comparative, structural and functional genomics

Oliver Redfern, Alastair Grant, Michael Maibaum, Christine Orengo*

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

Received 31 August 2004; accepted 1 November 2004
Available online 30 November 2004

Abstract

The last two decades have witnessed significant expansions in the databases storing information on the sequences and structures of proteins. This has led to the creation of many excellent protein family resources, which classify proteins according to their evolutionary relationship. These have allowed extensive insights into evolution and particularly how protein function mutates and evolves over time. Such analyses have greatly assisted the inheritance of functional annotations between experimentally characterised and uncharacterised genes. Moreover, the development of bioinformatics tools acts as a companion to the new technologies emerging in biology, such as transcriptomics and proteomics. The latter enable researchers to analyse gene expression profiles and interactions on a genome-wide scale, generating vast datasets of proteins, many of which include experimentally uncharacterised proteins. Protein family/function databases can be used to help interpret this data and allow us to benefit more fully from these technologies. This review aims to summarise the most popular sequence- and structure-based protein family databases. We also cover their application to comparative genomics and the functional annotation of the genomes.

Keywords: Databases; Proteomics; Genomics; CATH; Structural classifications

Contents

* Corresponding author. Tel.: +44 207 679 3284; fax: +44 207 679 7193.
E-mail address: orengo@biochem.ucl.ac.uk (C. Orengo).

1570-0232/$ – see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2004.11.010

1. Introduction

Over the course of evolution, mutations in the nucleotide bases of genes can result in numerous changes to the polypeptide chains they encode. In addition, substantial insertions and deletions of residues may occur by various recombination processes. These evolutionary mechanisms have given rise to families of proteins, which share a common ancestor but often exhibit considerable divergence in their sequence and structure. Frequently, this diversity is accompanied by changes in protein function. However, some protein families, such as the globins, retain a specific biological function despite high sequence diversity.

The earliest database of protein families was pioneered in the late 1970s by Margaret Dayhoff at the MIPS Institute in Germany and used to model the tolerances to different amino acid substitutions occurring through evolution. A plethora of protein family classifications have arisen over the subsequent 20 years, affording interesting insights into evolutionary processes. Moreover, they have provided comprehensive bioinformatics resources that may be used for inheriting functional information from experimentally characterised genes to their sequence or structural relatives. Powerful homologue detection algorithms are combined with manual validation to supply information on evolutionary relationships that are undetectable by simple sequence searching methods, such as BLAST. There is an ever-increasing amount of sequence data being generated from high-throughput methods such as genome sequencing projects, transcriptomics and proteomics. As a consequence, it is an impossible task to functionally characterise all proteins by experiment. As well as providing important data on evolutionary mechanisms, grouping genes into protein families offers a valuable resource for integrating information on cellular and molecular function.

In this review we aim to describe the most widely used sequence- and structure-based protein family classifications and consider the benefits gained from integrating these with databases of functionally annotated genes. To date, these resources have been chiefly exploited to explore the evolutionary relationship between sequence, structure and protein function. We shall summarise some of the key works in genome annotation and comparative genomics and then discuss the potential applications of protein family resources in functional genomics and proteomics.

2. Sequence based protein family classifications

Since protein structure determination is considerably more time consuming that gene sequencing, the sequence repositories have always been several orders of magnitude larger than the structure databases. There has, in fact, been an exponential increase in the sizes of both types of data since the early 1970s but the largest sequence database, GenBank [5] still contains nearly one million non-redundant sequences (July 2004), compared to ∼25,000 structures in the Protein Databank [6].

The earliest protein family classifications exploited pairwise sequence comparison to detect evolutionary relatives. However, these methods become unreliable in the so-called 'Twilight Zone' of sequence similarity (<30% sequence identity) [14]. Fortunately, the rapid expansion of the sequence databases over the past 10 years has increased the populations of the protein families, enabling the derivation of family-based sequence profiles and motifs.

Protein motifs represent small, highly conserved stretches of contiguous sequence, which may be associated with a particular evolutionary family or biological function. Searching for these recurring 'fingerprints' is often successful where global sequence similarity becomes unreliable. In a more sophisticated way sequence profile methods, such as Hidden Markov Models (HMMs) [26] and PSI-BLAST [1], have made it possible to capture the probability of certain residue mutations and insertions/deletions occurring in the sequence relatives of a given protein family [18,30]. These have been shown to be highly discriminatory in identifying distant homologous relationships when searching sequence databases.

Despite the success of the new profile methods, very distant homologues often remain undetectable. Hence, most sequence-based protein family classifications tend to group closely related sequences. Family members share significant sequence similarity and may possess similar or identical biological functions. Many resources choose to cluster whole protein chains. However, databases such as Pfam now identify separate domains within genes, often detected using proteins structure data, and group them accordingly. Thus, one gene may comprise several domains that are members of different protein families. In reviewing the databases, we highlight the distinction between those which cluster whole protein chains and those which focus on the domain level.

Table 1 summarises the current populations of the major sequence family databases and the methodologies used to create them. An important recent development has been the establishment of the InterPro Database ([35], http://www.ebi.ac.uk/interpro) at the EBI, in the UK. This resource integrates all the major protein family classifications and provides regular mappings from these family resources onto primary sequences in the UniProt database which contains over 800,000 sequences as of July 2004. InterPro, the Integrated Resource of Protein Families, Domains and Sites, is a collaboration that aims to provide an integrated interface of protein signature databases. Databases in the collaboration include UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily, SUPERFAMILY and Gene3D.

2.1. Families of sequence domains

Pfam [4] is a highly comprehensive resource providing an optimised set of Hidden Markov Model profiles for protein domain families. Families are defined using multiple

Table 1
Protein family resources

Resource	Group	Source(s)	No. families	Method	URL
PRINTS	Zygouri	SWISSPROT, TrEMBL	1800 entries, 10,931 motifs	Iterative motif searches	http://bioinf.man.ac.uk/dbbrowser/PRINTS
Pfam	Eddy	SWISSPROT, TrEMBL	7459 families	HMM	http://www.sanger.ac.uk/Software/Pfam
SMART	Bork	Selected proteins	667 domains	HMM	http://smart.embl-heidelberg.de
ProDom	Kahn	SWISSPROT, TrEMBL	501,917 families (186,303 non-singleton)	PSI-BLAST	http://protein.toulouse.inra.fr/prodom/current/html/home.php
InterPro	Zbobnov	UniProt, PROSITE, PRINTS, Pfam, ProDom, SMART, TIGRFAMs, PIR SuperFamily, SUPERFAMILY	11,007 entries (including 2573 domains, 8166 families)	Multiple methods (HMM, PSI-BLAST, regular expression)	http://www.ebi.ac.uk/interpro
TIGRFAMs	White	SWISSPROT, TrEMBL	1976 families	HMM	http://www.tigr.org/TIGRFAMs/index.shtml
ADDA	Holm	SWISSPROT, TrEMBL, PIR, PDB, WORMPEP, ENSEMBL	34,000 families (plus 60,000 singleton)		http://ekhidna.biocenter.helsinki.fi:8080/examples/servlets/adda/index.html
CHOP	Rost	62 complete genomes	63,300 clusters (plus 118,108 singleton clusters)	PSI-BLAST	http://cubic.bioc.columbia.edu/services/CHOP
TRIBES	Ouzounis	83 complete genomes	60,934 or 82,692 depending on granularity	TribeMCL	http://maine.ebi.ac.uk:8000/services/tribes
ProtoNet	Linial	SWISSPROT, TrEMBL	User-defined	BLAST	http://www.protonet.huji.ac.il
SYSTERS	Vingron	SWISSPROT, TrEMBL, ENSEMBL (complete genomes), the Arabidopsis Information Resource, SGD and GeneDB	158,153 disjoint clusters	BLAST	http://systers.molgen.mpg.de/
iProClass	Wu	PIR, SWISSPROT, TrEMBL, Pfam, BLOCKS, PRINTS, ProSite, PDB, COG	36,000 PIR superfamilies, 100,000 families	N/A	http://pir.georgetown.edu/iproclass
SWISSPROT	Schneider	Primary database	153,871 proteins	N/A	http://us.expasy.org/sprot
COG/KOG	Natale	66 unicelluar and 7 eukaryotic complete genomes	4873 COG, 4852 KOG	Bidirectional best hit	http://www.ncbi.nlm.nih.gov/COG

sequence alignments and HMMs and cover many common protein domains and families. Pfam consists of two parts, the first is the curated part of Pfam (Pfam-A), the second is an automatically generated supplement called Pfam-B.

Similarly, simple modular architecture research tool (SMART) [31] domain families have been selected with a particular emphasis on mobile eukaryotic domains and as such are widely found among nuclear, signalling and extracellular proteins. SMART domain families are annotated with function, sub-cellular localization, phyletic distribution and tertiary structure.

COG and KOG are databases of clusters of orthologous groups of proteins, defined by groups of three or more proteins in complete genomes. KOG contains seven eukaryotic genomes whilst COG contains 66 complete unicellular genomes.

2.2. Families of whole protein chain sequences

TIGRFAMs [19] protein families are built in a similar fashion to Pfam but also contain whole protein chains. ProtoNet developed by Linial and co-workers [53], uses alternatives clustering methods to group in the UniProt database on the basis of sequence similarity. Proteins from the TrEMBL repository are later added into these initial protein clusters. The ProtoNet protocol can produce protein family clusters from three different clustering methods: harmonic, geometric and arithmetic.

The PRINTS database [3] is a collection of protein 'fingerprints': conserved sequence motifs used to characterise a protein family. These motifs are generated via multiple protein sequence alignments by identifying regions of local sequence conservation. They can subsequently be used to scan a larger sequence set (e.g. UniProt and TrEMBL [7]) to recruit new family members. The majority of families are defined by multiple motifs and must all be present for a relative to be added to the group.

The SYSTERS [27] and TRIBES [15] methods use graph-based methods and Markov clustering respectively to generate protein families of varying granularity.

A number of other resources exist that automatically cluster sequences from the completed genomes or from the large sequence repositories (e.g. GenBank or Swissprot-TrEMBL) into putative domain families. The ProDom resource [54] contains protein sequence families derived from sequences in UniProt and TrEMBL. These protein sequences are chopped into protein domains using an iterative PSI-BLAST domain boundary prediction program.

Heger and Holm [21] recently developed the ADDA algorithm to cluster sequences into domain families. ADDA takes alignments from all-against-all sequence comparison to define domains within protein sequences, which are then clustered into domain families. Recently, almost 800,000 non-redundant sequences were condensed into 100,000 domain families (33% of the families containing more than one member) covering all of the currently available sequence space. A related algorithm, CHOP [32] designed by Liu and Rost [32], assigns domain boundaries by BLAST sequence comparison and then clusters the subsequent domain-like fragments into sequence families using the CLUP clustering method. Recently, 62 completed genomes were chopped and clustered into 118,108 single and 63,300 multi-member clusters.

There is an ever-increasing number of web-accessible sequence based classifications of protein families (see Table 1). The number of families identified by those performing automated clustering of large sequence repositories varies from 65,000 to 186,000 depending on the philosophy. Ouzounis and co-workers [15] recently revealed that each newly sequenced genome leads to an increase in the total number of protein families characterised. That is, currently a certain proportion of genome sequences (between 10 and 25%) in every genome are singletons, or belong to families not present in other sequenced genomes. This may reflect limitations in the current sequence based homologue detection algorithms; or alternatively these may be genuinely novel families that have arisen following speciation. The organism-specific families may be important for expanding the functional repertoire and phenotype of the organism, perhaps by providing new biological processes or changes in gene regulation.

3. Structure based protein family classifications

Despite the advances in sequence comparison methods, remote homologues in the 'Midnight Zone' of sequence similarity (<15% identity) described by Rost, can still only be identified through protein structure comparison [63,44]. Therefore, structure-based classifications are becoming increasingly important resources for recognising these distant relatives and providing datasets for more far-reaching analyses of protein family evolution.

The earliest protein structures were solved in the 1970s and deposited in the Protein Databank (PDB, [6]). This resource, which is now based in the Research Collaboratory of Structural Biology (RCSB) Rutgers University, contains around 25,000 protein structures comprising more than 60,000 individual protein domains. Since the early 1990s, there have been sufficient structures to cluster evolutionary relatives into protein families and superfamilies. This has given rise to comprehensive hierarchical databases of protein structures, such as CATH and SCOP, which rely on a combination of manual expert classification and structural comparison methods.

In their pioneering analysis of protein structural families, Chothia and Lesk [8] first demonstrated the degree to which protein structure appears to be more conserved than sequence during evolution. This has been reaffirmed by recent analyses of larger structural classifications [39]. Fig. 1 shows the relationship between sequence similarity and structure similarity for all homologous relatives in the CATH domain structure database. Many of the very distant relatives below 20% sequence identity are paralogous relatives, arising from duplication of a domain within the genome. Once duplicated,

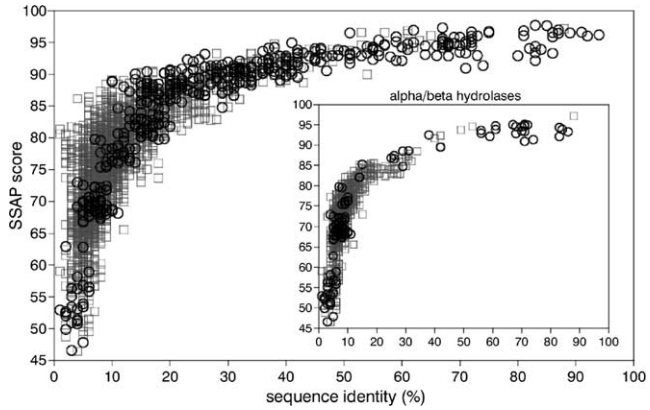

Fig. 1. Correlation between structure similarity (measured by the SSAP structure comparison algorithm, 0–100) and sequence similarity (measured by sequence identity) for all pairs of homologous domain structures in the CATH domain database.

the paralogous gene frequently evolves a new function. By recognising such relationships, the structural classifications have provided some important insights into the evolution of protein function within protein families.

3.1. Recognising domain boundaries

As well as the recognition of very remote homologues, structural data can help in determining the domain composition of a protein. Currently, approximately 40% of known structures are multi-domain proteins and this proportion is likely to rise as the techniques for structure determination advance. It is difficult to recognise individual domains using sequence data alone. However, Teichmann and co-workers [2] have recently suggested from sequence analysis of completed genomes, that at least two thirds of proteins within a genome are likely to be multi-domain proteins. This proportion could be as high as 80% in eukaryotic organisms.

Many algorithms have been written for recognising domain boundaries from structural data. These often exploit the fact that there are more contacts between residues within a domain than between different domains. Others search for hydrophobic clusters that could represent domain cores. Many domains have also been duplicated and combined with different partners during evolution. This forms the rational for methods that exploit domain recurrence to match domains in newly determined structures against libraries of classified domains (e.g. PUU method [22]). As there can be considerable structural variation in some protein families (see below), most classification resources (e.g. SCOP, CATH) manually validate domain boundaries identified by automated approaches.

3.2. Structural comparison algorithms

To facilitate the construction of structure-based family databases, some groups have developed automated methods to detect similarities between evolutionary relatives. Structure comparison and alignment algorithms were first introduced in the early 1970s and the rigid body superposition methods developed then, by Rossmann and Argos, are still used today for superposing structures and calculating a similarity measure (root mean square deviation, RMSD). This is achieved by translating and rotating the structures relative to one other until the difference between putative equivalent residues is minimised. Unfortunately, this approach runs into problems when aligning distant homologues that may contain extensive insertions and deletions of residues or shifts in the orientations of equivalent secondary structures. Therefore, more complex alignment algorithms based on dynamic programming, secondary alignment and fragment comparison have been developed.

Initially, many protein structure classifications use rapid secondary structure-based approaches to rapidly identify putative relatives before applying slower, more accurate residue based methods (e.g. GRATH—CATH database [22]; SEA—COMPASS Database [48]). As with sequence database searching, the value of many of these fast approaches is that they allow a large number of comparisons to be performed from which a rigorous statistical framework can be built for assessing the significance of any match [22]. There are far fewer secondary structures than residues and since most insertions and deletions occur in the loop regions connecting the secondary structures, this helps to eliminate the ‘noise’ created by the structural embellishments arising from divergent evolution of distant homologues. However, more computationally intensive residue-based algorithms (e.g. COMPARER [51]; SSAP [60]; STAMP [49]; DALI [23]; CE [54]) result in accurate structural alignments. Rather than simply attempting to superpose equivalent residues between protein structures, many of these methods compare the internal distances between residues within the same structure to align residues with similar sets of internal distances.

3.3. Protein structure classifications

The two major protein structure classifications, CATH and SCOP, focus on structural domains, classifying them into evolutionary superfamilies. These are further organised into a hierarchical schema, the top level of which corresponds to the protein class—the proportion of residues adopting α-helical or β-strand conformations. This gives rise to three major classes, mainly-α, mainly-β and α–β, although SCOP divides the alpha-beta class into alternating α/β and α + β, depending on the segregation of α-helices and β-strands along the polypeptide chain.

Within each class, structures are further clustered into fold groups when they possess significant structural similarity. To share the same fold, both the arrangement of secondary structures in 3D and the connectivity between them should be similar. The CATH database also recognises an intermediate level between class and fold in which structures are classified according to the orientations of the secondary structures in

3D. This architecture level describes the shape of the structure (e.g. barrel-like or layered sandwich) and can be helpful in imposing an additional level of order within each protein class. Finally, proteins adopting highly similar folds and further evidence of an evolutionary relationship (e.g. similar sequence motifs or shared functional characteristics) are grouped into the same homologous superfamily. Within these superfamilies, proteins are often further sub-clustered into families of close relatives possessing very similar functional properties.

The SCOP database was established in 1993 by Murzin et al. [36] and uses almost entirely manual validation for recognising structural similarities between proteins to generate evolutionary superfamilies. Although time consuming, this has resulted in a very high quality resource where domain boundaries are also manually assigned. In the CATH database [39], a combination of manual and automated approaches is used. Robust structure comparison methods (SSAP [60] CORA [40], GRATH [20]) have been developed to recognise structural relatives; although evolutionary relationships are only assigned following manual assessment of all available data. Several automatic methods are used for domain boundary recognition but, again, assignments are all manually validated. Table 2 shows that SCOP and CATH recognise around 800-fold groups and some 1200–1500 superfamilies in the current set of protein structures.

In contrast, the DALI domain database (DDD) established by Holm and co workers [12,13] uses a completely automated protocol. Domain boundaries are recognised using the PUU algorithm [22] and domains are assigned to fold groups and superfamilies using the robust DALI structure comparison algorithm [22]. Thresholds for clustering the structures are based on *Z*-scores, calculated by scanning new domains against all representative structures in the database. An excellent web-based search engine at the EBI in Cambridge (http://www.ebi.ac.uk/dali) has been developed for scanning new structures to identify putative relatives, which is widely used by structural biologists.

The HOMSTRAD and CAMPASS databases, constructed by Blundell and co-workers [34,55,56], are not hierarchical but focus on using SCOP, PFAM and other resources to cluster together families of evolutionary relatives, often with high sequence homology. An additional feature is the provision of validated multiple structural alignments for families and superfamilies that can be used to derive substitution matrices or to encode conserved structural features in a template—these can be used to identify further relatives. The CAMPASS database groups more distant structural homologues than HOMSTRAD by using the structural comparison algorithms COMPARER and SEA to generate multiple alignments from SCOP superfamilies.

Aside from resources that explicitly assign structures to fold groups and superfamilies, a number of 'neighbourhood' databases exist. These use automated structural comparison methods to search the PDB for structural neighbours to a query structure. The Entrez resource at the NCBI uses the

Table 2
Protein structure family resources

Database	Location and author	Coverage (July 2004)	Structure comparison method	Type	URL
CAMPASS	Cambridge University, UK, *Sowdhamini*	7580 domains in 1409 superfamilies	COMPARER [51], SEA [48]	Structure-based sequence alignments of SCOP superfamilies	http://www-cryst.bioc.cam.ac.uk/~campass
CATH/Gene3D	UCL, London, UK, *Orengo*	58,000 domains in 1459 superfamilies	SSAP [60], GRATH [20]	Automatic structural and sequence comparison methods are combined with manual validation of superfamily alignments and domain boundaries	http://www.biochem.ucl.ac.uk/bsm/cath
CE	SDSC, La Jolla, CA, USA, *Bourne*	All chains in PDB	CE [54]	Fully automatic, nearest neighbours	http://cl.sdsc.edu/ce.html
DHS	UCL, London, UK	1459 superfamilies in CATH	SSAP [60], CORA [37]	Fully automatic multiple structure alignments of close relatives in CATH superfamilies	http://www.biochem.ucl.ac.uk/bsm/dhs
ENTREZ/MMDB	NCBI, Bethesda, MD, USA, *Bryant*	All in PDB	VAST [64]	Fully automatic, nearest neighbours	http://www.ncbi.nih.gov/Structure/MMDB/mmdb.shtml
HOMSTRAD	Cambridge University, UK, *Blundell*	7500 domains in over 1400 superfamilies	COMPARER [51]	Manual classification of close protein homologues	http://www-cryst.bioc.cam.ac.uk/~homstrad
SCOP/SUPERFAMILY	LMB-MRC, Cambridge, UK, *Murzin*	54,745 domains in 1294 superfamilies	Manual	Manual classification	http://scop.mrc-lmb.cam.ac.uk/scop

VAST algorithm to identify its structural matches. The PDB has recently established a similar resource that uses the CE algorithm to detect relatives and calculates a probabilistic *E*-value for significance. Similarly, the Macromolecular Structure Database (MSD) which is the European node of the PDB uses the SSM algorithm, to find putative relatives.

4. Assigning complete genome sequences to protein families and exploiting this data in structural genomics

Although relatively few protein structures have been determined to date (∼4000 non-identical), recent genome analyses suggest that we currently have structural representatives for many of the most highly populated families in nature [30]. Several groups have used Hidden Markov Models (SUPERFAMILY [17], Gene3D [30]) and combined profile/threading based protocols [33] to assign sequences from completed genomes to structural families in the SCOP or CATH databases. Currently between 30 and 80% of sequences or partial sequences in a genome can be assigned to a structural family in one of these databases, depending on the organism and the technology used. A further 15–20% of sequences can be assigned to families in the Pfam database, suggesting that a very high proportion of genome sequences can now be classified in well-characterised and carefully curated superfamilies.

4.1. Sequence and structure based methods for detecting distant homologues

As discussed in Section 1, in order to recognise relatives in the 'Midnight Zone' of sequence similarity, the most powerful profile methods must be used. Perhaps the most widely used profile-based method is PSIBLAST [1], although Hidden Markov Models generated using the SAM-T99 technology of Karplus and co-workers [26] have recently been shown to recognise a significant proportion of distant homologues. Profile–profile methods [1] and Threading technologies are even more powerful but are prohibitively slow for performing large scale comparisons.

One of the most important developments has been the introduction of benchmarking protocols to determine reliable thresholds for accurate homologue detection. These approaches were originally pioneered by Chothia and co-workers [42] and use datasets of carefully validated, remote homologues, detected by structure comparison from both the SCOP [42,18] and CATH structure classifications [50,44]). Recent analysis, using the CATH domain database in June 2003, showed that approximately 83% of homologous pairs with less than 35% sequence identity could be recognised using the SAM-T99 [26] approach.

Fig. 2 shows that the families recognised in the genomes follow power-law like behaviour. That is, for all types of families (e.g. CATH, Pfam, uncharacterised) the vast majority of families are small and only occur once or a few

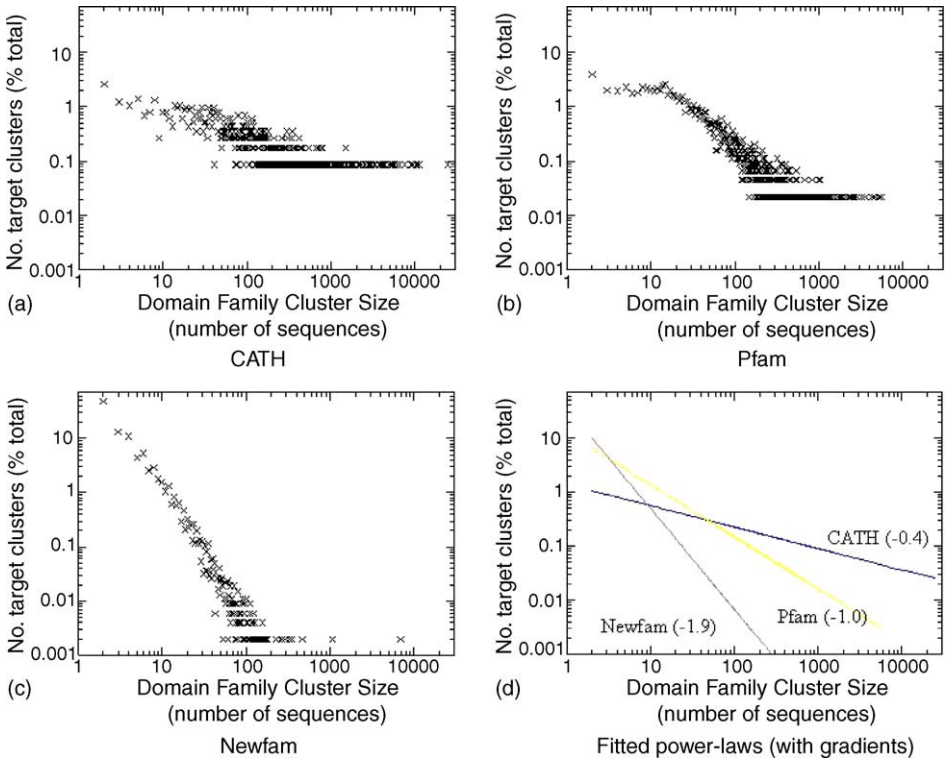

Fig. 2. Power-law like behaviour of families from (a) CATH domain database; (b) Pfam domain database; (c) uncharacterised domain families (NewFam) and (d) protein families from the Gene3D database.

times in a genome. By contrast, there are a small percentage of extremely large superfamilies that occur many times in a genome. Recent analyses of bacterial genomes using Gene3D annotations has revealed that fewer than 30 CATH superfamilies (<5% of the total number of superfamilies) are responsible for almost 50% of the domain annotations to genome sequences.

Many groups [18,30] have shown that a significant proportion of the sequences unassigned to characterised families (i.e. SCOP, CATH, Pfam) belong to very small protein families or are singletons. That is they are not found in any other species and may be contributing in some way to the specific functional repertoire of the organism. Analyses of the Gene3D database revealed that many of these sequences are small and likely to consist of a single domain [30]. Rost has demonstrated that a significant portion are predicted to have low secondary structure content, suggesting that they may only adopt a functional conformation on binding to another protein and/or that they may be peptides involved in regulation.

It is possible to use this genome annotation data and the extent to which protein superfamilies merge once structural data is acquired, to estimate the number of superfamilies and folds in nature. Estimates range from a few thousand folds to hundreds of thousands of superfamilies and folds depending on the approach used [10]. As mentioned above, Ouzounis and co-workers have shown that each sequenced genome brings new protein families with no sign of saturation in the near future. With new species constantly being discovered, for example by shotgun sequencing of environmental samples, it is difficult to estimate an upper limit for the number of protein families in nature. However, the trends observed to date suggest that the majority of these will be small and organism specific, whilst up to 70–90% of sequences within an organism will be assigned to fewer than 2000 characterised protein families (∼500 in CATH; ∼1500 in Pfam).

One important use of these genome annotation resources is to identify structurally uncharacterised superfamilies that are likely to possess a novel fold or function, so that these can be targeted for structural determination. Structural genomics initiatives are currently in progress in several countries and approximately 500 new structures have been solved by these initiatives over that last 3 years. Interestingly, although families with no structural representatives were targeted, only 15% of the structures were observed to be novel folds once the structures were solved. In the future, targeting the large, structurally uncharacterised, Pfam superfamilies will help to provide structural representatives for the majority of the genome sequences.

5. Recent insights into protein evolution from comparative genome analysis using protein structure family resources

Over the last two decades protein family resources have been used extensively to analyse evolutionary relationships within individual protein families. It is outside the scope of this review to consider all these analyses. However, the success of the international genome initiatives, which has led to the complete sequencing of nearly 200 genomes, has recently enabled some interesting large scale analyses on the distribution of protein families within and across organisms in the different kingdoms of life. Exploiting the structural data in particular, allows more ancient evolutionary relationships to be tracked and can therefore give a clearer picture of evolutionary trends. Below, we briefly consider some of the insights gleaned over the last few years by analysing the distribution of structural families in the genomes.

The huge expansion in the sequence repositories together with improvements in technologies for detecting distant evolutionary relationships (e.g. PSIBLAST, HMMS, see above) has led to the nearly tenfold expansion of the structural family databases (SCOP, CATH) with sequence relatives from the genomes (e.g. see SUPERFAMILY, Gene3D, see above and Table 2). The additional functional information this brought to these resources recently led to an evaluation of the ways in which protein functions can be modified in family relatives during evolution. More specifically, analysis of enzyme families showed that although close relatives (≥40% sequence identity for single domain proteins, ≥60% sequence identity for multi-domain proteins) were likely to share common functions, in some families considerable functional divergence could occur between remote homologues or paralogues during evolution [63].

Todd et al. [63] performed extensive analyses of the mechanisms by which function changed in ∼30 exceptionally promiscuous CATH enzyme superfamilies. Following duplication of a domain during evolution, functional divergence was frequently associated with changes in domain partnerships of these paralogous domains. Differences in oligomerisation state were also observed. In some cases, residue insertions had resulted in large structural embellishments that had modified the active site or created additional interaction sites on the protein surfaces. Whilst in some relatives, mutations of just a few residues in the active site were sufficient to significantly alter the function. Interestingly, these modifications were mostly associated with changes in the substrate specificities. By contrast, many features of the chemistry performed by the enzyme were conserved—for example, a chemical intermediate along a reaction pathway.

These observations were supported by several related analyses investigating the recurrence of structural families in the small molecule, metabolic pathways of *E. coli* [47,2]. The studies suggested that enzymes were often recruited to a new pathway during evolution, to provide a specific chemistry, leading to a patchwork model of pathway evolution (Horowitz model). Serial recruitment of homologous enzymes along the same metabolic pathway, because they provided similar active site geometries for binding substrates/products of reactions, occurred much less frequently (Jensen model).

More general analyses of domain family recurrences in the genomes, pioneered by Teichmann and co-workers [2,30], illustrated the extent to which domain duplication occurred within genomes. Some families are much more extensively duplicated than others, leading to the power-law like behaviour shown in Fig. 2. These data mirrored the earlier analyses of structural families in the early 1990s, which had also revealed bias in the population of domain families. The analysis of Teichmann's and others showed that the most commonly recurring families were often associated with important generic functions, such a providing energy or redox equivalent for a chemical reaction. Other recurrent domains were involved in information exchange and DNA binding.

Many of these analyses are dominated by sequences from bacterial genomes as until recently there were only a few completely sequenced eukaryotic genomes. In July 2004, there were still only 16 complete eukaryotic genomes available. A survey of microbial genomes can therefore provide a statistically more reliable snapshot of the evolutionary mechanisms occurring within this kingdom. Surveying some 90 bacterial genomes, Ranea et al. [45] recently showed that although there appeared to be around 200 structural superfamilies common to all bacterial genomes, a smaller subset of about 60 of these were very massively duplicated, accounting for nearly 50% of the domain annotations in the genomes. In these families, the number of relatives found in any genome increased with genome size.

Detailed analysis showed that these superfamilies were predominantly associated with the COG functional categories of metabolism and regulation. Although, some variation in the functional categories was observed [45], the trends were very clear. For metabolic superfamilies, the increase in number of relatives with genome size was found to be linear and was associated with a high duplication rate. In the superfamilies associated with regulation, a non-linear relationship was observed with a lower average duplication rate. These contrasting behaviours leads to a balance between the populations of the two superfamilies (see Fig. 3) so that at a given genome size the number of regulatory families starts to outnumber the metabolic families.

This observation supports earlier hypotheses which suggested that in bacteria, gene duplication within metabolic superfamilies, followed by functional divergence of these paralogous genes, is a mechanism for expanding the functional repertoire of an organism and can give rise to new phenotypes. However, the existence of multiple copies of a gene, albeit with slightly modified functions, can start to lead to 'noise' and necessitates increased regulation. Thus the non-linear increase in the regulatory repertoire may represent the organism's response to tempering this noise. It may also be associated with additional regulatory mechanisms for expanding the functional repertoire by controlling the expression and activation of these paralogues in alternative ways.

Nevertheless, it is intriguing to note that if the increase in regulatory genes is viewed as cost that must be endured in order to benefit from the additional value brought by expanding the metabolic repertoire, the optimal balance between these two types of families occurs in the most frequently observed genome size for non-specialist bacteria. That is, bacteria having no specialised dietary or environmental requirements.

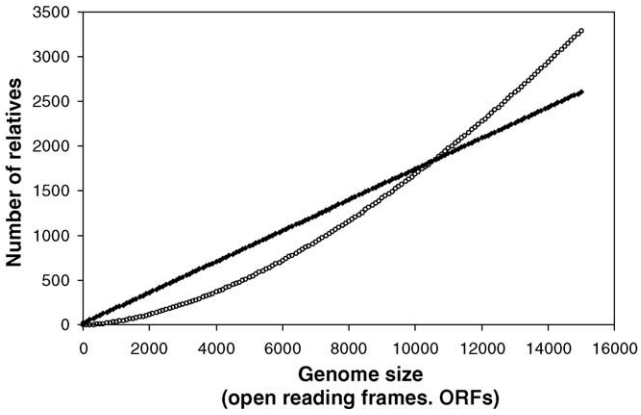

Fig. 3. The balance between expansion of metabolic families and regulatory families with increase in genome size. The linear increase in the number of domains primarily involved in metabolism are shown by the thick black line, whilst the non-linear increase in the number of domains primarily involved in regulation are shown by the dotted line.

6. Exploiting protein family resources to facilitate analysis of functional genomics data

Studies investigating the extent to which function is conserved between homologues are important for exploiting protein family resources to functionally annotate sets of genes being studied in large scale functional genomics experiments (e.g. transcriptomics and proteomics). As discussed above, the results of several groups suggests that, for enzyme families, there should be at least 30–40% sequence identity between relatives to be reasonably confident that they share a common or related function. For multi-domain proteins, higher levels of sequence similarity (50–60%) may be required [63]. Obviously it is helpful to consider the domain composition or architecture before inheriting functional information, as changes in this architecture can be clearly responsible for modulating protein function.

Protein family resources that provide information on domain partnerships are useful in this context. Recent versions of the Pfam website now provide information on domain compositions (http://www.sanger.ac.uk/Software/Pfam). The MIPS resource in Germany also clusters multi-domain proteins according to domain composition and is therefore a valuable resource for checking degree of homology and likely functional similarity between proteins. In the Gene3D resource, domain compositions are described by the CATH and Pfam annotations assigned to each protein [30].

Another important development in the use of bioinformatics resources to assist analysis of functional genomics data is the plethora of specialised functional databases established over the last decade. The enzyme (EC) and UniProt

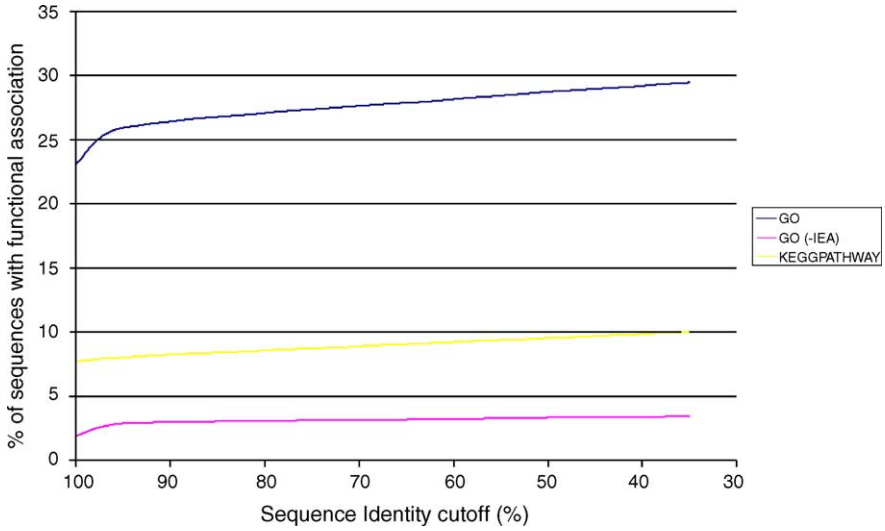

Fig. 4. The percentage of sequences in Gene3D that can be associated with a functionally-annotated gene by alignment, at a range of sequence identity cutoffs.

databases are widely used and provide useful and standardised functional descriptions. More recently, the gene ontology (GO) developed by Ashburner and co-workers [62] has played a significant role in producing a comprehensive and widely accepted scheme for describing function at different levels; molecular function, molecular process and cellular localisation. A range of other databases describing biological processes (KEGG [25], WIT [41]) and protein–protein, protein–ligand interactions has also appeared (BIND, DIP, TAP). Many of these are being integrated in the InTact database being developed by groups at the EBI (http://www.ebi.ac.uk/intact). One of the most significant developments has been the emergence of common identification code (Uniprot ID) which will unite the various sequence repositories (GenBank, EMBL, DDBJ) and facilitate mapping between all these resources.

In the light of these advances we have been developing the BIOMAP resource, at UCL. This is a data warehouse that integrates protein family resources (CATH, Gene3D) with various functional databases (e.g. GO, COG, KEGG, EC) to facilitate functional annotation of interesting genes identified by microarray analysis. Linking the protein family resources to the functionally annotated genes allows us to expand meaningful annotations that can be associated with each query gene. Fig. 4 illustrates the extent to which functional annotations can be extended by exploiting protein family relationships. By searching for relatives that have at least 35% sequence identity, it is possible to significantly increase (by more than 20%) the number of functional assignments that can be made across a protein family. Protein–protein interaction data will particularly useful for validating assignment of genes to particular pathways and functional complexes and will be incorporated in the next release of BIOMAP.

7. Conclusions

The last two decades have witnessed significant expansions in the databases storing information on the sequences and structures of proteins. This has led to the creation of many excellent protein family resources (see Tables 1 and 2) which classify these proteins according to their evolutionary relationship. Analyses of protein evolution and in particular the manner in which function is modified between paralogues have been essential in reliably exploiting these relationships to inherit functional information between experimentally characterised and uncharacterised genes. These developments have been very timely as revolutionary new technologies in biology (e.g. transcriptomics and proteomics) are enabling studies to be conducted on a genome-wide scale and generating vast datasets of proteins many of which are still experimentally uncharacterised. Applying bioinformatics and protein family/function databases to help interpret this data should help in significantly reducing the amount of experimental characterisation required and will allow us to benefit more fully from these new technologies.

References

[1] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. 25 (17) (1997) 3389.
[2] G. Apic, J. Gough, S.A. Teichmann, J. Mol. Biol. 310 (2001) 311S.
[3] T.K. Attwood, P. Bradley, D.R. Flower, A. Gaulton, N. Maudling, A.L. Mitchell, G. Moulton, A. Nordle, A.K. Paine, P. Taylor, A. Uddin, C. Zygouri, Nucleic Acids Res. 31 (2003) 400.
[4] A. Bateman, L. Coin, R. Durbin, R.D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E.L.L. Sonnhammer, D.J. Studholme, C. Yeats, S.R. Eddy, Nucleic Acids Res. 32 (2004) D138.

[5] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler, Nucleic Acids Res. 31 (2003) 23.
[6] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J.D. Westbrook, C. Zardecki, Acta Crystallogr. D Biol. Crystallogr. 58 (Pt. 6 No. 1) (2002) 899.
[7] B. Boeckmann, A. Bairoch, R. Apweiler, M. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider, Nucleic Acids Res. 31 (2003) 365.
[8] C. Chothia, A. Lesk, EMBO J. 5 (4) (1986) 823.
[10] A.F. Coulson, J.A. Moult, Proteins 46 (2002) 61.
[12] S. Dietmann, L. Holm, Nat. Struct. Biol. 11 (2001) 953.
[13] S. Dietmann, L. Holm, Nat. Struct. Biol. 8 (2001) 953.
[14] R.F. Doolittle, Methods Enzymol. 183 (1990) 99.
[15] A.J. Enright, V. Kunin, C.A. Ouzounis, Nucleic Acids Res. 31 (2003) 4632.
[17] B. Geier, G. Kastenmüller, M. Fellenberg, H.-W. Mewes, B. Morgenstern, Bioinformatics 17 (2001) 571.
[18] J. Gough, K. Karplus, R. Hughey, C. Chothia, J. Mol. Biol. 313 (4) (2001) 903.
[19] D.H. Haft, J.D. Selengut, O. White, Nucleic Acids Res. 31 (2003) 371.
[20] A. Harrison, F. Pearl, R. Mott, J. Thornton, C. Orengo, J. Mol. Biol. 323 (5) (2002) 909.
[21] A. Heger, L. Holm, J. Mol. Biol. 328 (2003) 749.
[22] L. Holm, C. Sander, Proteins 19 (1994) 256.
[23] L. Holm, C. Sander, J. Mol. Biol. 233 (1993) 123.
[25] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, Nucleic Acids Res. 32 (2004) 277.
[26] K. Karplus, B. Hu, Bioinformatics 17 (8) (2001) 713.
[27] A. Krause, J. Stoye, M. Vingron, Nucleic Acids Res. 28 (2000) 270.
[30] D. Lee, A. Grant, R. Marsden, C. Orengo, Proteins, 2004, in press.
[31] I. Letunic, L. Goodstadt, N.J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R.R. Copley, C.P. Ponting, P. Bork, Nucleic Acids Res. 30 (2002) 242.
[32] J. Liu, B. Rost, Proteins 55 (3) (2004) 678.
[33] L.J. McGuffin, S. Street, S.A. Sorensen, D.T. Jones, Bioinformatics 20 (1) (2004) 131.
[34] K. Mizuguchi, C.M. Deane, T.L. Blundell, J.P. Overington, Protein Sci. 7 (1998) 2469.
[35] N.J. Mulder, R. Apweiler, T.K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R.R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S.E. Orchard, M. Pagni, D. Peyruc, C.P. Ponting, J.D. Selengut, F. Servant, C.J.A. Sigrist, R. Vaughan, E.M. Zdobnov, InterPro Nucleic Acids Res. 31 (2003) 315.
[36] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, J. Mol. Biol. 247 (1995) 536.
[37] C.A. Orengo, Protein Sci. 8 (4) (1999) 99.
[39] C.A. Orengo, I. Sillitoe, G. Reeves, F.M. Pearl, J. Struct. Biol. 134 (2–3) (2001) 145.
[40] C.A. Orengo, Protein Sci. 8 (1999) 699.
[41] R. Overbeek, N. Larsen, G.D. Pusch, M. D'Souza, E. Selkov Jr., N. Kyrpides, M. Fonstein, N. Maltsev, E. Selkov, Nucleic Acids Res. 28 (2000) 123.
[42] J. Park, K. Karplus, C. Barrett, R. Hughey, D. Haussler, T. Hubbard, C. Chothia, J. Mol. Biol. 284 (4) (1998) 1201.
[44] F.M. Pearl, D. Lee, J.E. Bray, D.W. Buchan, A.J. Shepherd, C.A. Orengo, Protein Sci. 11 (2) (2002) 233.
[45] J.A. Ranea, D.W. Buchan, J.M. Thornton, C.A. Orengo, J. Mol. Biol. 336 (4) (2004) 871.
[47] S.C. Rison, S.A. Teichmann, J.M. Thornton, J. Mol. Biol. 318 (2002) 911.
[48] S.D. Rufino, T.L. Blundell, J. Comput. Aided. Mol. Des. 8 (1) (1994) 5.
[49] R.B. Russell, G.J. Barton, Proteins 14 (2) (1992) 309.
[50] A.A. Salamov, M. Suwa, C.A. Orengo, M.B. Swindells, Protein Eng. 12 (2) (1999) 95.
[51] A. Sali, T.L. Blundell, J. Mol. Biol. 212 (1990) 403.
[53] F. Servant, C. Bru, S. Carrère, E. Courcelle, J. Gouzy, D. Peyruc, D. Kahn, Briefings Bioinform. 3 (2002) 246.
[54] I.N. Shindyalov, P.E. Bourne, Protein Eng. 11 (9) (1998) 739.
[55] A.S. Siddiqui, G.J. Barton, Protein Sci. 4 (5) (1995) 872.
[56] R. Sowdhamini, D.F. Burke, J.-F. Huang, K. Mizuguchi, H.A. Nagarajaram, N. Srinivasan, R.E. Steward, T.L. Blundell, Structure 6 (1998) 1087.
[60] W.R. Taylor, C.A. Orengo, J. Mol. Biol. 208 (1) (1989) 1.
[62] The Gene Ontology Consortium. Nucleic Acids Res. 32 (2004) D258.
[63] A.E. Todd, C.A. Orengo, J.M. Thornton, J. Mol. Biol. 307 (4) (2001) 1113.
[64] T. Madej, J.F. Gibrat, S.H. Bryant, Proteins 23 (1995) 356.