Abstract
The study of superfamilies of protein domains using a combination of structure, sequence and function data provides insights into deep evolutionary history. In the present paper, analyses of functional diversity within such superfamilies as defined in the CATH-Gene3D resource are described. These analyses focus on structure-function relationships in very large and diverse superfamilies, and on the evolution of domain superfamily members in protein–protein complexes.

Introduction
Over the last decade, the accumulation of sequence data from hundreds of completed genomes has enabled deeper analyses of the nature of protein evolution and how this contributes to the diverse functional repertoires seen in different species and kingdoms. Protein evolution is best understood when combining structure, sequence and function information from multiple sources. Structural data, combined with sequence-based profile–profile and HMM (hidden Markov model) comparisons, often reveals very ancient relationships between proteins.

Also, proteins tend to consist of domains, i.e. subparts that can have independent evolutionary histories. Chopping proteins into their constituent domains allows the identification of even deeper evolutionary relationships between proteins, and the uncovering of homologies which are not apparent when considering proteins as a whole.

Resources that relate sequence and structure data are being developed and maintained in our group and are described briefly below. These resources have allowed us to study the evolution of proteins within superfamilies and within complexes, and some interesting examples are presented.

Structural domain superfamilies
The CATH-Gene3D database [1] provides high-quality MDA (multidomain architecture) assignments for sequences in the major protein and genome databases (e.g. UniProt [2], ENSEMBL [3]), based on the curated domain boundary and homology assignments provided by the CATH database of structural domains [4]. CATH clusters evolutionarily related domains into ‘superfamilies’ using structure, sequence and function information. Currently, 2200 domain superfamilies have been identified using this approach. For each superfamily, a set of representative seed domain sequences are then used to generate a library of profile HMMs [5].

The CATH HMM models are then searched against a set of sequence databases [2,3,6], and the results are parsed using an in-house method for assigning domain structure annotations and identifying MDAs (C. Yeats, O. Redfern and C.A. Orengo, unpublished work). Genome definitions are obtained from Integr8 for UniProt sequences [7] and from Ensembl [3]. Coverage for genomes vary from 40 to 60% of domain sequences, depending on the organism. The coverage for various sequence collections can be interactively analysed using the Gene3DViz tool (http://gene3d.biochem.ucl.ac.uk:8090/Gene3DViz/Gene3DViz.html).

Gene3D integrates these domain annotations with other types of structural prediction. These include transmembrane regions, coiled-coils and other structural elements of interest, as well as domain and family predictions from Pfam [8] and other sequence signature resources in InterPro [9]. In addition, biochemical, expression and molecular function data are integrated with the domain annotations via mappings to several commonly used resources such as GO [10], KEGG [11] or IntAct [12]. Gene3D annotations can be integrated easily with other resources via commonly used protein identifiers, including those from UniProt and Ensembl, either manually through websites or using common programmable interfaces [13].

Evolution of protein superfamilies
Whereas the vast majority of domain superfamilies classified in CATH contain only a few, generally very similar, members, a few superfamilies (<100) display great diversity in terms of sequence, structure and function, and are highly over-represented in structural and sequence databases [14–16].

These very large and diverse superfamilies have drawn increasing attention, and are the subject of many endeavours to understand better the mechanisms that bring about such structural and functional diversity between homologues [17–19]. In our group, a number of analyses have been...
undertaken to understand how structural changes among relatives can bring about changes in function. In general, it was shown that related domains can differ by the presence of large insertions, which are often located close to one another in three-dimensions, and are often involved in aspects of molecular function such as catalysis or binding partner molecules [18]. Changes in other structural properties, such as local residue variations in active sites [20] or MDAs [21], can also affect function.

We have studied a number of specific superfamilies in order to better characterize the impact of structural variations on functional changes. For each of these superfamilies, we follow a standardized procedure whereby subgroups of relatives with similar functions are first identified [these are referred to as FSGs (functional subgroups) throughout the present paper], and their structural characteristics are then compared. Superfamilies under study in our group include the vicinal oxygen chelate and haloacid dehalogenase superfamilies, which are also curated in the Structure–Function Linkage Database [22], periplasmic-binding proteins [23], α/β-hydrolases [24] and HUP domains. In the present paper, we provide highlights from our analyses of the HUP superfamily.

HUP domains are very ancient and are found in a large number of very different proteins [25]. The core of the HUP domain adopts a Rossmann fold with a central parallel β-sheet, surrounded on both sides by α-helices. This core displays much plasticity in that it accommodates large structural embellishments at several different topological insertion points and three-dimensional locations relative to the core. The location and nature of these embellishments generally depend on the function of the multidomain proteins containing the HUP domain. The structural plasticity of HUP domains is also apparent at the level of MDAs, as HUP domains are found in combination with >150 different domain types in the Gene3D database.

In total, we are able to identify nine structurally characterized FSGs of HUP domains. The largest FSG is that of class I aminoacyl-tRNA synthetases, which are responsible for the attachment of aminoacyl moieties to their cognate tRNAs for use in translation [26]. In these enzymes, the HUP domain is the seat of the main catalytic site. Other prominent FSGs in this superfamily include electron–transfer flavoproteins, which are small heterodimeric proteins involved in the transfer of electrons to other proteins [27], N-type ATP-pyrophosphatases, which have different metabolic activities but share the ability to adenylate a substrate and then attack the adenylated intermediate with a nucleophilic nitrogen from a second substrate [28], and universal stress proteins, which are poorly characterized proteins whose expression is enhanced under different stress conditions [29] (see Table 1 for a description of all FSGs identified in the HUP domain superfamily).

A comparative analysis of structures in the HUP superfamily helps us to understand how different types of structural change contribute to specifying the particular functions that characterize the different FSGs. Such an approach can, of course, be applied to other superfamilies as well. Examples discussed in the text are illustrated in Figure 1.

A first example of structural change that can cause concomitant changes in function is that of local and small-scale structural changes in the vicinity of active sites. Such an approach can, of course, be applied to other superfamilies as well. Examples discussed in the text are illustrated in Figure 1.

Table 1 | Description of FSGs within the HUP superfamily

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Representative structure (PDB code)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electron-transfer flavoproteins</td>
<td>Electron transfer</td>
<td>1O97</td>
</tr>
<tr>
<td>Class I aminoacyl-tRNA synthetases</td>
<td>Catalyse the binding of amino acids to cognate tRNAs</td>
<td>1H3F</td>
</tr>
<tr>
<td>N-type ATP-pyrophosphatases</td>
<td>Metabolic enzymes that catalyse two-step reactions where a substrate is adenylated and then attacked by a second substrate</td>
<td>1CT9</td>
</tr>
<tr>
<td>DNA-photolyases and cryptochromes</td>
<td>DNA-photolyases are activated by light and are involved in DNA damage repair</td>
<td>1DNP</td>
</tr>
<tr>
<td>Nucleotidyltransferases</td>
<td>Metabolic enzymes that catalyse various reactions involving the transfer of nucleotidyl moieties from nucleotide triphosphate to other substrates</td>
<td>1R6X</td>
</tr>
<tr>
<td>Pantothenate synthetase</td>
<td>Related to nucleotidyltransferases, but only intermediates are adenylated</td>
<td>1MOP</td>
</tr>
<tr>
<td>tRNA 2-thiouridylase</td>
<td>Catalyse tRNA modifications</td>
<td>2DER</td>
</tr>
<tr>
<td>Phosphoadenosine phosphosulphate reductase</td>
<td>Catalyse the reduction of PAPS to sulfite</td>
<td>1SUR</td>
</tr>
<tr>
<td>Universal stress protein A</td>
<td>Poorly characterized proteins whose expression is stimulated under various stress conditions</td>
<td>1JMV</td>
</tr>
</tbody>
</table>
The HUP domain core is represented in cartoons, coloured grey and displayed in a similar orientation in all panels.

(a) Carbapenam synthetase (PDB code 1Q19) is very similar in structure and function to asparagine synthetase B and β-lactam synthetase, yet local changes in the vicinity of the active site mediate different substrate specificities in these enzymes. These local changes include slight shift of residues 371–375 (coloured red) towards the centre of the active site, changes in three-dimensional configuration of residues 377–388 from a helix in asparagine synthetase B to a loop in carbapenam synthetase (coloured orange), and ordering of residues 442–450 in the holo-form of carbapenam synthetase (coloured salmon). Ligands are displayed as ball-and-stick, and an active-site Mg$^{2+}$ ion is represented as a green sphere.

An extra N-terminal domain present in this enzyme is coloured blue. (b) Changes in domain partners of HUP domains can specify changes in functions as well; for example, some Class I aminoacyl-tRNA synthetases, such as valyl-tRNA synthetase (PDB code 1GAX) shown here, have additional editing domains (coloured blue), which allow them to correct non-cognate aminoacyl-tRNA pairs. Part of the tRNA molecule is represented as ball-and-stick and coloured orange. (c) The HUP domain of electron-transfer flavoprotein β (PDB code 1EFV) possesses a structural embellishment (coloured green) that consists of a three-stranded β-sheet on the side of the HUP domain core. This embellishment mediates extensive contacts with the other subunit (coloured blue) in electron-transfer flavoprotein, and this intersubunit interaction is in turn necessary for the protein to function properly.

analyses [20,30,31], and are often linked to changes in cofactors and ligands, which are very common among homologues. Several illustrations of such changes can be found in the HUP superfamily. For example, asparagine synthetase B, carbapenam synthetase and β-lactam synthetase are enzymes that belong to the ATP-pyrophosphatase FSG. These enzymes have identical MDAs and very similar tertiary structures and catalytic mechanisms. However, Figure 1(a) illustrates how small-scale changes in the binding pocket are responsible for very different substrate specificities [32].

Changes in MDAs can also affect functional properties in multiple ways [21]. For example, some class I aminoacyl-tRNA synthetases have an extra domain inserted within the HUP domain, which allows them to verify the aminoacyl-tRNA complexes produced in the main active site and to edit (or modify) them in cases where an amino acid is erroneously attached to the wrong tRNA (see Figure 1b). In haloacid dehalogenases, changes in the so-called CAP domain, which covers the main active site of the catalytic domain, allow homologues to bind different substrates and thereby to catalyse different reactions [33].

Structural embellishments, i.e. secondary-structure insertions relative to the conserved structural domain core for the superfamily, are also often involved in specifying changes in function between homologues. For example, tRNA-2-thiouridylases possess very large embellishments close to the active site that allow these enzymes to bind their tRNA ligand productively [34]. Similarly, major structural embellishments in electron-transfer flavoproteins are responsible for mediating interactions with other protein chains (see Figure 1c).

Several examples of functional implications of structural embellishments taken from other superfamilies were documented previously [18]. It is worth noting that structural embellishments to the superfamily core often affect other structural properties such as MDAs or oligomerization states, as large insertions are often located at interfaces with such extra domains or subunits. This is, for instance, the case with valyl-tRNA synthetase, in which the interaction between the HUP domain and the above-mentioned editing domain is mediated by extensive embellishments to the HUP domain core (see Figure 1b).

Thus it appears that homologous domains with different functions are generally associated with specific structural features, and that these can often be shown to affect crucial aspects of molecular function directly.
Evolution of complexes: the reuse of protein domain superfamilies in protein complexes

In addition to diversity in enzymatic activities, different members of a domain superfamily may also be involved in a variety of different protein complexes and cellular pathways. Protein complexes are aggregates of multiple protein chains, encoded by separate genes, performing concerted functions. Well-known examples include the ribosome, the spliceosome and the proteasome. These molecular machines occur throughout cellular life and their evolution is poorly understood.

It has been noted that when a protein duplicates, both copies should share not only the same biochemical activity, but also the same set of protein–protein interactions [35]. If this is the case, duplicates should, at least over short evolutionary timescales, conserve their role in the same or similar biological processes and protein complexes.

Two clear alternatives exist. First, a duplicate protein may occur in a different regulatory context (e.g. with alternative promoters), in which case it might be expressed in a different tissue or life stage. Secondly, a single duplicated domain might find itself in a different multidomain context. If the domain of interest has interactions with other proteins or domains, then the interactions in which it was involved previously might be maintained. If, however, it has recombined with new domains, which are themselves involved in protein–protein interactions, the duplicate domain may immediately become involved in a distinct set of interactions from the ancestral domain.

Furthermore, interactions between the domain of interest and other domains with which it has recombined might allow the evolution of novel interactions. Thus duplicate domains may make evolutionary jumps between different sets of interactions rather than, over time, losing or gaining individual inherited interactions.

Several authors have successfully identified protein complexes from pairwise protein–protein interactions using clustering techniques [36]. In a recent analysis (A.J. Reid, J.A. Ranea and C.A. Orengo, unpublished work), we have used this technique to identify large datasets of protein complexes in yeast and Escherichia coli. These datasets were then used to discover how domain superfamily members are reused in complexes.

Figure 2 shows, for E. coli and yeast, the number of superfamilies compared with the number of different complexes in which each superfamily is found. There is a strong positive correlation between superfamily size and the number of complexes in which that superfamily is found. For E. coli, $r^2$ equals 0.99 and for yeast 0.97. This suggests that duplicated domains tend to change their context and move into new complexes.

We wanted to know whether different members of a superfamily are involved in similar biological processes despite their random distribution among complexes. Using biological process GO (Gene Ontology) terms, we found that 28% of superfamilies in E. coli and 22% in yeast have members that are involved in more similar biological processes than expected by chance ($P < 0.01$). Whereas homologous domains tend to become involved in different complexes after duplication, one-fifth to one-third of superfamilies appear to conserve their functional role to some extent.

However, when comparing the functional similarity of the proteins with which each superfamily member is directly
Table 2 | Non-randomly distributed superfamilies in E. coli (a) and yeast (b) MCL (Markov clustering)-GO complexes

Superfamily CATH codes are shown in parentheses. Frequency values are the number of proteins containing a member of that superfamily in that complex dataset. Functional descriptions are based on the most common GO terms from proteins containing the superfamily in that particular organism. Superfamilies are considered to belong to a given kingdom when they are found in at least 70% of completed genomes from that kingdom. Universal refers to eukaryotes, bacteria and archaea.

(a)

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Frequency</th>
<th>Function</th>
<th>Species distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAD(P)-binding Rossmann-like domain</td>
<td>73</td>
<td>Oxidoreductase activity in a wide variety of processes</td>
<td>Universal</td>
</tr>
<tr>
<td>(3.40.50.720)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(b)

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Frequency</th>
<th>Function</th>
<th>Species distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA-binding (2.30.30.100)</td>
<td>11</td>
<td>RNA binding/splicing</td>
<td>Universal</td>
</tr>
<tr>
<td>Glutamine phosphoribosylpyrophosphate, subunit 1, domain 1 (3.60.20.10)</td>
<td>18</td>
<td>Ubiquitin-mediated endopeptidase activity</td>
<td>Universal</td>
</tr>
<tr>
<td>Quinoprotein amine dehydrogenase (2.130.10.10)</td>
<td>68</td>
<td>Wide range of activities including protein synthesis</td>
<td>Universal</td>
</tr>
<tr>
<td>Protein tyrosine phosphatase superfamily (3.90.190.10)</td>
<td>12</td>
<td>Dephosphorylation in signalling pathways</td>
<td>Eukaryotic</td>
</tr>
<tr>
<td>Ribosomal protein (3.30.1370.10)</td>
<td>4</td>
<td>Binding activity in a variety of processes</td>
<td>Universal</td>
</tr>
<tr>
<td>Ubiquitin-like superfamily (3.10.20.30)</td>
<td>5</td>
<td>Tricarboxylic acid cycle</td>
<td>Universal</td>
</tr>
</tbody>
</table>

interacting, there is much less conservation (1% of E. coli and 8% of yeast superfamilies having interactors with conserved function), i.e. if protein A interacts with proteins B, C and D and protein A homologue A’ interacts with proteins E, F and G, then proteins B, C and D are not functionally similar to proteins E, F and G. This suggests that those superfamilies which conserve their function to some extent tend to diversify into distinct aspects of similar processes. This has been recognized before in yeast in the work of Baudot et al. [37]. In our work, the trend is stronger in E. coli than in yeast.

We then considered whether there are superfamilies which do not follow this trend and tend to conserve their complex membership. For each superfamily we determined how many times two proteins both containing a member of that superfamily are found together in a complex. This was compared with the number of co-complex pairs that would be expected if the proteins were distributed randomly among complexes. We found that, for most superfamilies, their members do not co-occur in complexes more than would be expected by chance. In total, 98% of E. coli superfamilies and 95% of yeast superfamilies are randomly distributed. Table 2 shows details of those superfamilies which do tend to occur in multiple copies in complexes.

What is the functional significance of multiple homologues within a complex? In E. coli, only one non-randomly distributed superfamily was identified: the NAD(P)-binding Rossmann-like domain superfamily. Those complexes containing multiple members of this superfamily tend to be large, with diverse functional roles. Therefore the role of multiple members of this superfamily in individual complexes is unclear. In yeast, there are six superfamilies found to be non-randomly distributed among complexes. These fall into three categories: RNA processing, the proteasome and signal transduction.

It appears that those members of superfamilies which cluster together tend to be involved in eukaryote-specific processes (e.g. splicing and ubiquitin-dependent protease activity). These tend, however, to be universal superfamilies, suggesting that these eukaryotic advancements have largely developed from duplication and divergence of pre-existing superfamilies.

These results suggest that, as a rule, duplicate domains are not reused within protein complexes, but instead take on different roles in the cell. Furthermore, there may be different forces acting on complex evolution in prokaryotes and eukaryotes. A restriction on duplicate proteins in prokaryotes may extend to limitations in the way in which complexes have evolved in these organisms. Alternatively, the simpler molecular machinery of the prokaryotic cell may simply not require reuse of paralogous domains.

Conclusion

In the present paper, we describe approaches used in our group to characterize functional and structural diversity within protein domain superfamilies, as defined in the in-house CATH-Gene3D resource. Once structural and functional diversity has been characterized, it is possible to come up with approaches to take advantage of these special features for predicting functions more accurately.
We have developed a number of computational approaches which identify distinctive structural features (FLORA; O.C. Redfern, B.H. DesSailly, T. Dallman and C.A. Orengo, unpublished work) or conserved residue patterns (GEMMA; D. Lee, R. Rentzsch and C.A Orengo, unpublished work) associated with particular functional subgroups in large functionally diverse superfamilies. As more superfamilies are characterized functionally using such methods and the approaches described above, CATH will provide more details on the functional diversity observed at the superfamily level.

Funding

This work was supported by the European Union Framework Programme 7 Impact Grant (B.H.D.), the European Union Experimental Network for Functional Integration (ENFIN) Newark of Excellence (J.L.), the European Union BioSapiens Network of Excellence (C.Y.) and the Biotechnology and Biological Sciences Research Council (A.J.R. and A.C.).

References


Received 27 April 2009
doi:10.1042/BST03570745

©The Authors Journal compilation ©2009 Biochemical Society