Evolution of Function in Protein Superfamilies, from a Structural Perspective

Annabel E. Todd¹, Christine A. Orengo¹ and Janet M. Thornton¹,²*

¹Biochemistry and Molecular Biology Department, University College London, Gower Street London, WC1E 6BT, UK
²Crystallography Department Birkbeck College Malet Street London, WC1E 7HX, UK

The recent growth in protein databases has revealed the functional diversity of many protein superfamilies. We have assessed the functional variation of homologous enzyme superfamilies containing two or more enzymes, as defined by the CATH protein structure classification, by way of the Enzyme Commission (EC) scheme. Combining sequence and structure information to identify relatives, the majority of superfamilies display variation in enzyme function, with 25% of superfamilies in the PDB having members of different enzyme types. We determined the extent of functional similarity at different levels of sequence identity for 486,000 homologous pairs (enzyme/enzyme and enzyme/non-enzyme), with structural and sequence relatives included. For single and multidomain proteins, variation in EC number is rare above 40% sequence identity, and above 30%, the first three digits may be predicted with an accuracy of at least 90%. For more distantly related proteins sharing less than 30% sequence identity, functional variation is significant, and below this threshold, structural data are essential for understanding the molecular basis of observed functional differences. To explore the mechanisms for generating functional diversity during evolution, we have studied in detail 31 diverse structural enzyme superfamilies for which structural data are available. A large number of variations and peculiarities are observed, at the atomic level through to gross structural rearrangements. Almost all superfamilies exhibit functional diversity generated by local sequence variation and domain shuffling. Commonly, substrate specificity is diverse across a superfamily, whilst the reaction chemistry is maintained. In many superfamilies, the position of catalytic residues may vary despite playing equivalent functional roles in related proteins. The implications of functional diversity within superfamilies for the structural genomics projects are discussed. More detailed information on these superfamilies is available at http://www.biochem.ucl.ac.uk/bsm/FAM-EC/.

Corresponding author: thornton@biochem.ucl.ac.uk

Keywords: enzyme superfamilies; three-dimensional structure; function; diversity; evolution

Introduction

Determining the biological role of all gene products is the principal objective of genome analysis.

Given that a small minority of known sequences is experimentally characterised, gene annotation relies heavily upon the accurate exploitation of evolutionary relationships; functional information is extrapolated following the identification of a sequence relative, on the basis that family members commonly exhibit some similarity in function. The recent growth in sequence and structural data, however, has revealed the remarkable functional promiscuity of many protein families. It is apparent that one fold may code for multiple functions, and conversely, one function may have more than one structural solution, having evolved independently several times during evolution.

Abbreviations used: EC, Enzyme Commission; FAD, flavin adenine dinucleotide; FMN, flavin mononucleotide; HPL, human pancreatic lipase; LpxA, UDP-N-acetylglucosamine acyltransferase; PaXAT, chloramphenicol acetyltransferase; PLP, pyridoxal-phosphate; TPP, thiamine pyrophosphate; PEP, phosphoenolpyruvate; TIM, triosephosphate isomerase.
These complexities necessitate caution in annotation transfer. The successive transfer of information between homologues based on an incorrect functional assignment would ultimately undermine the value of genome databases. There are two ways to help practically to improve genome analysis. Firstly, one can assess quantitatively the reliability of annotation transfer, and recent work has provided valuable insights into its accuracy by identifying sequence identity thresholds above which functional variation is rare (Wilson et al., 2000). Secondly, an understanding of the underlying mechanisms of evolving new functions through sequence and structural changes is vital. This is particularly relevant with the advent of structural genomics initiatives which aim to provide a structural representative for all homologous protein families. These will reveal previously undetected evolutionary relationships hidden at the sequence level, since protein structure is conserved even after all trace of sequence similarity disappears. Detailed comparisons of unknown gene products with structural relatives will reveal regions of conservation and variation, and will guide experiments by providing clues to binding, catalysis and signalling. Whilst the native structure may hint at little more than biochemical function, this functional assignment provides a valuable first step towards the experimental elucidation of cellular and physiological roles.

Structural redundancy, the reuse of the same fold in different contexts, has led to several estimates of the number of protein folds in Nature, and there are probably a few thousand at most (Chothia, 1992; Orengo et al., 1994). Given the large number of genes in the human genome, but a comparatively small number of folds, extensive combination, mixing and modulation of existing folds has occurred during evolution to generate the multitude of functions necessary to sustain life. With the first working draft of the human genome complete, and the sequencing of other multi-cellular organisms underway, a grasp of these evolutionary processes is required if we are to benefit from this wealth of data.

Ultimately, we would like to provide answers to the following questions. To what extent is functional divergence correlated with sequence divergence? What dictates the choice of ancestral proteins for the evolution of new functions? Is a broad functional repertoire limited to a few protein superfamilies which adopt particularly adaptable folds, or are all superfamilies susceptible to evolutionary changes which bring about functional variation?

To understand the global relationships between protein sequence, structure and function we are reliant on robust classifications of protein families and their functions. Several protein family data collections exist (Orengo et al., 1997; Murzin et al., 1995; Bateman et al., 2000) and these classify proteins into evolutionary families many of which are now well-populated. Of the functional classifications, the Enzyme Commission (EC) (Webb, 1992) is the best developed and most widely used. Whilst it deals only with enzymes, given the over-representation of these proteins in the Protein Data Bank (PDB) (Bernstein et al., 1977) with almost one half of entries corresponding to enzymes (Hegyi & Gerstein, 1999), this classification provides a useful starting point for addressing the above questions. Table 1 outlines the meanings of the different levels in the EC hierarchy.

Several recent analyses which have employed the EC scheme have provided novel insights into the complex relationships between protein sequence, structure and function (Martin et al., 1998; Hegyi & Gerstein, 1999; Wilson et al., 2000; Devos & Valencia, 2000). Martin et al. (1998) found little correlation between primary EC number and secondary structure class, α, β and α/β, consistent with the dependence of enzyme activity upon a few residues in the active-site. Hegyi & Gerstein (1999) assessed the versatility of protein folds with respect to enzyme function, and vice versa. They found that just a few folds, notably in the α/β structural class, have a diverse range of functions, and conversely, glycosyl hydrolysis is the most ubiquitous activity, carried out by seven different folds, covering all three fold classes.

In related work, Wilson et al. (2000) assessed quantitatively the relationship between functional similarity and sequence by considering pairs of well-characterised structural domains from the SCOP database. Using the EC scheme, and their own augmented version of the FLY database (Ashburner & Drysdale, 1994) for the classification of non-enzymes, they identified a ~40% sequence identity threshold above which precise function, as defined by the first three levels in the functional classification, is conserved. In their analysis, Devos & Valencia (2000) investigated the relationship between sequence identity and several functional descriptions, including EC assignments, SWISSPROT keywords (Bairoch & Apweiler, 2000) and binding sites, using FSSP (Holm & Sander, 1996) as a source of structural domains and pairwise alignments. Conservation of these functional descriptions decreases in this order, and all of them are less conserved than protein structure.

Zhang et al. (1999) found that for about 10% of Escherichia coli proteins with significant sequence identity to a PDB entry annotated with SITE records, there is no conservation of functional residues. This result is based on the lack of conservation of both catalytic and binding site residues as assigned in PDB SITE records, which are not consistently defined. In their analysis, Russell et al. (1998) considered only structural data and discussed the use of identifying co-located substrate binding sites in inferring functional properties. Roughly 10% of remote homologues have different binding sites implying a complete change in function.

Here we assess the functional variation of homologous enzyme superfamilies in the PDB by way
of the EC scheme, and discuss how functional changes are implemented by modulation of sequence and structure with reference to 31 functionally diverse superfamilies. We have analysed these superfamilies in detail using sequence and structural data, and relevant literature. With specific examples, we discuss the conservation and variation of catalytic residues, reaction mechanisms and substrate specificity, as well as changes in domain organisation and quaternary structure which are important routes to functional diversification.

Conservation of EC numbers of homologous enzymes: correlation with sequence identity

In our analysis we only compare functions of proteins in the PDB within the same homologous superfamily. Protein families are extracted from the CATH classification scheme (Orengo et al., 1997). Since we are dealing with structures in the PDB, note that the results are biased by the structural database content.

In CATH, the unit of classification is the structural domain, but an EC number describes the role of the protein complex as a whole and it is not always possible to identify distinct subfunctions for each constituent domain. That is, in some enzymes, the active-site can be unambiguously assigned to a specific domain, but in others, several domains play a role in the catalytic activity. In previous analyses (Martin et al., 1998; Hegyi & Gerstein, 1999; Wilson et al., 2000; Devos & Valencia, 2000), this problem was overcome by considering just single-domain proteins. However, modular construction has been an important route to new gene functions. A total of 37% of polypeptide chains in the PDB are multi-domain, but these constitute a larger fraction (62%) of homologous superfamilies in CATH. Of the single-domain proteins, 19% leave 100 residues or more uncovered upon alignment with their corresponding SWISS-PROT database sequence, suggesting that just one domain of several has been structurally determined for many of these proteins. Therefore, it is essential to consider both single and multi-domain proteins for a complete understanding of functional evolution. For our analysis of enzyme superfamilies, the function of a protein is assigned to all constituent domains.

As shown in Figure 1, in almost one half of 167 homologous structural superfamilies containing two or more enzymes, members show variation in their EC classification. In a number of these families, the EC number varies only in the fourth digit, implying a change in substrate or product specificity, or cofactor dependency, for example. Peptidases (EC 3.4.-.-) have a different classification scheme (Barrett, 1994). Note also that it is a classification of overall enzyme reactions, not enzymes, and takes no account of the details of the reaction chemistry involved (see caveats below).

<table>
<thead>
<tr>
<th>Table 1. Description of the different levels in the EC classification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First figure</strong></td>
</tr>
<tr>
<td>A. OXIDOREDUCTASES</td>
</tr>
<tr>
<td>B. TRANSFERASES</td>
</tr>
<tr>
<td>C. HYDROLASES</td>
</tr>
<tr>
<td>D. LYASES</td>
</tr>
<tr>
<td>E. ISOMERASES</td>
</tr>
<tr>
<td>F. LIGASES</td>
</tr>
</tbody>
</table>

An enzyme reaction is assigned a four-digit EC number, where the first digit denotes the class of reaction. Note that the meaning of subsequent levels depends upon the primary number, e.g. the substrate acted upon by the enzyme is described at the second level for oxidoreductases, whereas it is described at the third level for hydrolases. Different enzymes clustered together at the third level are given a unique fourth number, and these enzymes may differ in substrate/product specificity or cofactor-dependency, for example. Peptidases (EC 3.4.-.-) have a different classification scheme (Barrett, 1994). Note also that it is a classification of overall enzyme reactions, and not enzymes, and takes no account of the details of the reaction chemistry involved (see caveats below).
non-enzymes increases to 59. Whilst individual examples of changes in catalytic activity have been observed previously, such as adenylate cyclase (EC 4.6.1.1) and DNA polymerase I (EC 2.7.7.7) (Artyymiuk et al., 1997), the number of superfamilies in which it occurs is unexpected. Differences in enzyme class often belie a similarity in reaction chemistry and catalytic mechanism, as discussed in more detail below.

These results are based on limited data; many enzymes in the databases lack an EC assignment, and in addition, a number of evolutionary relationships are likely to have been undetected. As more proteins are sequenced and characterised, and previously unknown evolutionary relationships are identified, we might expect the majority of superfamilies to exhibit considerable functional diversity.

**Sequence and function**

Figure 2(a) and (b) reflects the conservation of EC number at different levels of pairwise sequence identity. Note that homologous (enzyme/enzyme and enzyme/non-enzyme) pairs only are considered in these histograms. Wilson et al. (2000) concluded that for single-domain proteins, enzyme function, as defined by the first three EC numbers, is almost completely conserved above a sequence identity threshold of 40%. Figure 2(a) is in agreement with this, and indicates also that even variation in the fourth EC digit is rare. In this analysis, sequence relatives are included also. Even at a relatively low sequence identity of 30%, enzyme function may be predicted as far as the third level in the EC hierarchy with an accuracy of almost 95%. Below this threshold, the extent of conservation falls rapidly.

Figure 2(b) includes all protein domains, belonging to both single and multi-domain proteins. Surprisingly, the effect of modular construction on functional diversification is not that dramatic, and is significant only below 40%. Even within the 30-40% sequence identity region, almost 90% of pairs share a minimum of three EC digits. Below 30%, the pairing of enzymes with non-enzymes becomes quite common. The notable reduction in the number of homologues having EC numbers conserved as far as the second level with the inclusion of multi-domain proteins may reflect the comparatively large number of oxidoreductases; these enzymes are often multi-domain proteins, having a catalytic domain fused to a cofactor-binding module, notably the Rossmann fold (e.g. medium-chain alcohol dehydrogenases). The second EC digit describes the nature of the reducing substrate (for hydrolases and isomerases it is described at the third level) so a change in substrate specificity, or even a difference in reaction direction introduces a change in the classification at this level.

The sequence identities of 75% of homologous pairs are 30% or lower. Figure 2(d) shows that the large percentage of these diverse pairs is not restricted to a limited number of superfamilies that show extensive sequence diversity. The majority of superfamilies have homologous pairs in the 10-40% sequence identity range, and three quarters of superfamilies have a mean sequence identity between all members of less than 50%. Thus, most of the superfamilies have undergone considerable “radiation” in sequence space.

The annotation of a new sequence usually involves identification of the “best” sequence hit. The distribution of the sequence identities of the closest relatives of all protein domains is shown in Figure 3. In contrast to Figure 2(a) and (b), here we consider the best pairs only, not all pairs, and the two distributions are very different. The largest fraction of best pairs have sequence identities above 90%, whilst the largest fraction of all pairs have sequence identities below 20%. This reflects the existence of sequence clusters on different branches of the evolutionary tree. Thus, the vast majority of sequences have close relatives, but most families also exhibit extreme sequence diversity and include very distantly related members. An important observation is that the sequence identity of the closest relative of only 10% of domains is under 40%, the “critical” threshold of functional variation.

Whilst the relatively high conservation in function even down to 30% sequence identity is promising for genome annotation, it is important to note that there are some well-known cases of differences in function at very high levels of sequence identity, notably the crystallins which have been recruited from enzymes and function as structural proteins in the eye lens (Wistow & Piatigorsky, 1987). Some genes acquire a new function prior to duplication (gene recruitment) (Piatigorsky & Wistow, 1991). Incomplete characterisation of gene products, and
incomplete and incorrect database annotations, may contribute to the observed infrequency of crystallin-like changes in protein function, but to what extent is unknown.

Mechanisms of Enzyme Evolution

Given the observed functional versatility of many protein superfamilies, it is useful to consider the possible routes to new functions, as outlined in Figure 4. Conceptually, the simplest route to create a new function is to make a new protein ab initio. However, there are many alternative routes and, in practice, new functions often evolve via a combination of mechanisms, notably through gene duplication and incremental mutations. An increasingly large number of genes are identified as multi-functional, where function is dependent upon biological context (Jeffery, 1999). For example, variations in expression, cellular localisation and substrate concentration can lead to modulation in function for these “moonlighting” proteins. The use of one gene for two or more functions clearly simplifies the genome, but complicates the process of genome annotation. Other routes to new functions include oligomerisation, gene fusion, alternate splicing and post-translational modifications.

Specific Examples of Structural and Functional Variations

We have focused on 31 superfamilies which show significant functional variation. In the following sections, we provide an overview of how these protein superfamilies evolved to perform different functions, illustrated by specific examples taken from these 31 superfamilies (due to space limitations, a more detailed description on each superfamily can be found at http://www.biochem.ucl.ac.uk/bsm/FAM-EC/). A webpage for each superfamily details structural members and their respective domain organisations, EC numbers and cofactors, with links to PDBsum (Laskowski et al., 1997), SWISS-PROT and ENZYME (Bairoch, 2000) databases. Sequence relatives are listed also. Each webpage provides a
short summary of the superfamily in terms of structure, substrate specificities and reaction mechanisms, and further information, such as pairwise sequence identities and SSAP (Taylor & Orengo, 1989) structural alignment scores, can be accessed from each page. Table 2 lists the members of each superfamily studied.

Tables 3 and 4 provide a summary of the structural and functional properties and variations observed within each superfamily. In 19 superfamilies, all members share a minimum of one domain in common. In 11 superfamilies, two structural domains are shared by all members. Note that seven of the enzyme superfamilies adopt the ubiquitous TIM barrel fold: ribulose-phosphate-binding TIM barrels, aldolase superfamily, phosphoenolpyruvate-binding domains, enolase superfamily, FAD-dependent oxidoreductases, TIM barrel glycosyl hydrolases and the metal-dependent hydrolases. The evolutionary origin of TIM barrel proteins has been a subject of controversy. After submission of the manuscript, authors of a large-scale sequence analysis of proteins adopting this fold proposed a common ancestry between the first five TIM barrel superfamilies listed, as well as others (Copley & Bork, 2000). In the absence of overwhelming structural and functional evidence of an evolutionary relationship, these superfamilies were not clustered together in this study. However, even within the individual superfamilies, there is some considerable divergence of function.

**Substrate specificity**

Numerous enzyme superfamilies analysed here display remarkable variations in substrate selectivity (see Table 4). Of 28 superfamilies involved in substrate-binding, in only one is the substrate absolutely conserved, and this corresponds to the phosphoenolpyruvate-binding enzymes which have the TIM barrel fold.

Enzymes in six superfamilies bind to a common substrate type, such as DNA, sugars or phosphorylated proteins. However, in at least three of these superfamilies, variations within these ligand types may be extensive. For example, members of the glycosyl hydrolase superfamily considered here act on monosaccharides as well as larger carbohydrate substrates including linear and branched-chain polysaccharides, and the carbohydrates may be linked to proteins and a variety of aryl and alkyl groups. Accordingly, their active sites differ vastly in shape and size. Endo-β-N-acetylglucosaminidase even lacks the fifth and sixth α-helices of its TIM
barrel to accommodate the protein moiety of its substrate (Van Roey et al., 1994).

Of the remaining 22 superfamilies, substrate selectivity is the least varied for the ribulose-phosphate-binding TIM barrels, with the substrates, and usually products, of their reactions having a glycerol or ribulose-phosphate group. This conservation of substrate-binding is not surprising, given that the product of one reaction is the substrate of the next.

In as many as 20 superfamilies, substrate specificity is completely diverse, in that the substrates bound vary in their size, chemical properties and/or structural scaffolds (e.g. aromatic versus linear hydrocarbons) illustrating the plasticity of protein structures with respect to ligand-binding (see examples in Figure 5). If any substrate similarity exists within these 20 enzyme superfamilies, it is limited to a small chemical moiety such as a carbonyl group or peptide bond, as identified in ten superfamilies, and typically this is the centre of chemical reactivity during catalysis.

Substrate diversity implies diverse binding sites, achieved through structural variations and exploiting the varying properties of the 20 amino acids. The helix-hairpin-helix base-excision DNA repair

\[
\begin{align*}
\text{hexapeptide repeat proteins} & \\
(a) \quad & \text{cytochrome P450s} \\
(b) \quad & \\
(c) \quad & \\
(d) & \\
\text{FAD/NAD(P)(H)-dependent disulphide oxidoreductase superfamily} \\
(j) & \\
(k) & \\
(l) & \\
(m) & \\
(n) & \\
\end{align*}
\]

Figure 5. Substrates bound by enzymes belonging to three superfamilies which display diverse substrate specificities. Hexapeptide repeat proteins: (a) chloramphenicol acetyltransferase (EC 2.3.1.28); (b) tetrahydropicolinate N-succinyltransferase (2.3.1.117); (c) UDP-N-acetylglucosamine acyltransferase (2.3.1.129); (d) carbonic anhydrase (4.2.1.1); cytochrome P450s: (e) cytochrome P450-terp (1.14.-.-); (f) cytochrome P450-cam (1.14.15.1); (g) cytochrome P450-bm3 (1.14.14.1); (h) cytochrome P450-nor (nitric oxide reductase) (1.14.-.-); (i) cytochrome P450-eryF (1.14.-.-); FAD/NAD(P)(H)-dependent disulphide oxidoreductase superfamily: (j) glutathione reductase (1.6.4.2); (k) trypanothione reductase (1.6.4.8); (l) NADH peroxidase (1.11.1.1); (m) mercuric ion reductase (1.16.1.1); (n) thioredoxin reductase (1.6.4.5).
Table 2. The enzyme superfamilies and their members

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Member</th>
<th>EC number(s)</th>
<th>PDB code</th>
<th>Pfam</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fumarase/aspartase</strong></td>
<td>Fumarate hydratase class II (fumarase)</td>
<td>4.2.1.2</td>
<td>1fuq</td>
<td>PF00206</td>
</tr>
<tr>
<td></td>
<td>Aspartate ammonia-lyase (aspartase)</td>
<td>4.3.1.1</td>
<td>1jsw</td>
<td>PF00206</td>
</tr>
<tr>
<td></td>
<td>Adenylosuccinate lyase</td>
<td>4.3.2.2</td>
<td>1c3c</td>
<td>PF00206</td>
</tr>
<tr>
<td></td>
<td>Argininosuccinate lyase</td>
<td>4.3.2.1</td>
<td>1aos</td>
<td>PF00206</td>
</tr>
<tr>
<td></td>
<td>δ crystallin II/argininosuccinate lyase</td>
<td>4.3.2.1</td>
<td>1auw</td>
<td>PF00206</td>
</tr>
<tr>
<td></td>
<td>Turkey δ crystallin</td>
<td>Non-enzyme</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Histidine ammonia-lyase</td>
<td>4.3.1.3</td>
<td>1b8f</td>
<td>PF00221</td>
</tr>
<tr>
<td><strong>Helix-hairpin-helix base-</strong></td>
<td><strong>excision DNA repair enzymes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Endonuclease III</td>
<td>4.2.99.18</td>
<td>2abk</td>
<td>PF00730</td>
</tr>
<tr>
<td></td>
<td>DNA-3-methylenadenine glycosylase</td>
<td>3.2.2.21</td>
<td>1mpg</td>
<td>PF00730</td>
</tr>
<tr>
<td></td>
<td>A/G-specific adenine glycosylase</td>
<td>3.2.2.-</td>
<td>1mun</td>
<td>PF00730</td>
</tr>
<tr>
<td></td>
<td>8-Oxoguanine DNA glycosylase</td>
<td>-</td>
<td>1ebm</td>
<td></td>
</tr>
<tr>
<td><strong>Non-heme di-iron carboxylate</strong></td>
<td><strong>proteins</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ribonucleotide reductase, β chain</td>
<td>1.17.4.1</td>
<td>1xik</td>
<td>PF00268</td>
</tr>
<tr>
<td></td>
<td>Methane monoxygenase hydroxylase component, α chain</td>
<td>1.14.13.25</td>
<td>1mty</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Methane monoxygenase hydroxylase component, β chain</td>
<td>1.14.13.25 (non-catalytic protein chain)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Δ9 stearoyl-acyl carrier protein desaturase</td>
<td>1.14.99.6</td>
<td>1afr</td>
<td>PF00487</td>
</tr>
<tr>
<td></td>
<td>Ruberythrin</td>
<td>Non-enzyme</td>
<td>1b71</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bacterioferitin</td>
<td>Non-enzyme</td>
<td>1bcf</td>
<td>PF01334</td>
</tr>
<tr>
<td></td>
<td>Ferritin heavy chain</td>
<td>Non-enzyme</td>
<td>2fha</td>
<td>PF00210</td>
</tr>
<tr>
<td></td>
<td>Ferritin light chain</td>
<td>Non-enzyme</td>
<td>1dat</td>
<td>PF00210</td>
</tr>
<tr>
<td></td>
<td>DNA protection during starvation protein</td>
<td>Non-enzyme</td>
<td>1dps</td>
<td>PF02047</td>
</tr>
<tr>
<td><strong>Cytochrome P450s</strong></td>
<td>Cytochrome P450-terp</td>
<td>1.14.--.</td>
<td>1cpt</td>
<td>PF00067</td>
</tr>
<tr>
<td></td>
<td>Cytochrome P450-cam</td>
<td>1.14.15.1</td>
<td>1phb</td>
<td>PF00067</td>
</tr>
<tr>
<td></td>
<td>Cytochrome P450-eryF</td>
<td>1.14.--.</td>
<td>1oxa</td>
<td>PF00067</td>
</tr>
<tr>
<td></td>
<td>Cytochrome P450-bn3</td>
<td>1.14.14.1 1.6.2.4 [m]</td>
<td>1bu7</td>
<td>PF00067</td>
</tr>
<tr>
<td></td>
<td>Cytochrome P450-nor (nitric oxide reductase)</td>
<td>1.14.--.</td>
<td>1rom</td>
<td>PF00067</td>
</tr>
<tr>
<td><strong>Ferredoxin-NADP reductase</strong></td>
<td>Ferredoxin-NADP reductase</td>
<td></td>
<td>2nd</td>
<td>PF00175</td>
</tr>
<tr>
<td>(FNР) modules</td>
<td>NADH reductase</td>
<td>1.6.6.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pthalate dioxygenase reductase</td>
<td>1.6.2.2</td>
<td>1ndh</td>
<td>PF00175</td>
</tr>
<tr>
<td></td>
<td>Ferredoxin--NADP reductase</td>
<td>1.18.1.2</td>
<td>1mbn</td>
<td>PF00175</td>
</tr>
<tr>
<td></td>
<td>NADPH-cytochrome P450 reductase</td>
<td>1.6.2.4</td>
<td>1amo</td>
<td>PF00175</td>
</tr>
<tr>
<td></td>
<td>Flavohemoglobin</td>
<td>-</td>
<td>1cxq</td>
<td>PF00175</td>
</tr>
<tr>
<td></td>
<td>NAD(P)H-flavin reductase</td>
<td>1.6.8.-</td>
<td>1qfj</td>
<td>PF00175</td>
</tr>
<tr>
<td><strong>Cupredoxins</strong></td>
<td>Rusticyanin</td>
<td>Non-enzyme</td>
<td>1rcy</td>
<td>PF00127</td>
</tr>
<tr>
<td></td>
<td>Stellacyanin</td>
<td>Non-enzyme</td>
<td>1jer</td>
<td>PF00127</td>
</tr>
<tr>
<td></td>
<td>Amicyanin</td>
<td>Non-enzyme</td>
<td>1aac</td>
<td>PF00127</td>
</tr>
<tr>
<td></td>
<td>Azurin</td>
<td>Non-enzyme</td>
<td>1nwp</td>
<td>PF00127</td>
</tr>
<tr>
<td></td>
<td>Pseudoazurin</td>
<td>Non-enzyme</td>
<td>1paz</td>
<td>PF00127</td>
</tr>
<tr>
<td></td>
<td>Plastocyanin</td>
<td>Non-enzyme</td>
<td>1pcl</td>
<td>PF00127</td>
</tr>
<tr>
<td></td>
<td>Cucumber basic protein</td>
<td>Non-enzyme</td>
<td>2bpe</td>
<td>PF00127</td>
</tr>
<tr>
<td></td>
<td>Copper-containing nitrite reductase</td>
<td>1.7.99.3</td>
<td>1nf</td>
<td>PF00394</td>
</tr>
<tr>
<td></td>
<td>Laccase (polyphenol oxidase)</td>
<td>1.10.3.2</td>
<td>1a65</td>
<td></td>
</tr>
<tr>
<td></td>
<td>L-ascorbate oxidase</td>
<td>1.10.3.3</td>
<td>1aoz</td>
<td>PF00394</td>
</tr>
<tr>
<td></td>
<td>Ceruloplasmin (ferroxidase)</td>
<td>1.16.3.1</td>
<td>1kwc</td>
<td>PF00394</td>
</tr>
<tr>
<td></td>
<td>Cytochrome c oxidase polypeptide II</td>
<td>1.9.3.1 (principal catalytic centre in a separate protein chain)</td>
<td>2cua</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ubiquinol oxidase polypeptide II</td>
<td>1.10.3.- (principal catalytic centre in a separate protein chain)</td>
<td>1cyw</td>
<td></td>
</tr>
<tr>
<td><strong>Discoidin-like domains</strong></td>
<td>Galactose oxidase</td>
<td>1.1.3.9</td>
<td>1gof</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sialidase</td>
<td>3.2.1.18</td>
<td>1eut</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Coagulation factor V</td>
<td>Non-enzyme</td>
<td>1czt</td>
<td>PF00754</td>
</tr>
<tr>
<td></td>
<td>Coagulation factor VIII</td>
<td>Non-enzyme</td>
<td>1d7p</td>
<td>PF00754</td>
</tr>
<tr>
<td><strong>&quot;Rieske&quot;-like iron-sulphur</strong></td>
<td><strong>domains</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ubiquinol-cytochrome c reductase, iron-sulphur subunit</td>
<td>1.10.2.2</td>
<td>1rie</td>
<td>PF00355</td>
</tr>
<tr>
<td></td>
<td>Cytochrome b6/f complex, iron-sulphur subunit</td>
<td>1.10.99.1</td>
<td>1rfs</td>
<td>PF00355</td>
</tr>
<tr>
<td></td>
<td>Naphthalene 1,2-dioxygenase, α subunit</td>
<td>1.14.12.12</td>
<td>1ndo</td>
<td>PF00355</td>
</tr>
<tr>
<td><strong>Hexapeptide repeat proteins</strong></td>
<td>Chloramphenicol acetyltransferase</td>
<td>2.3.1.28</td>
<td>1xat</td>
<td>PF00132</td>
</tr>
<tr>
<td></td>
<td>Tetrahydrodipicolinate N-succinyltransferase</td>
<td>2.3.1.117</td>
<td>3tdt</td>
<td>PF00132</td>
</tr>
<tr>
<td></td>
<td>UDP-N-acetylglucosamine acetyltransferase</td>
<td>2.3.1.129</td>
<td>1lxa</td>
<td>PF00132</td>
</tr>
<tr>
<td>Superfamily</td>
<td>Member</td>
<td>EC number(s)</td>
<td>PDB code</td>
<td>Pfam</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------------------------------------------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>--------</td>
</tr>
<tr>
<td>Ribulose-phosphate-binding TIM barrels</td>
<td>Indole-3-glycerol phosphate synthase/ N-(5-phospho-riboyl)anthranilate isomerase Indole-3-glycerol phosphate synthase Orotidin 5-monophosphate decarboxylase Tryptophan synthase, z chain Ribulose-phosphate 3-epimerase</td>
<td>4.1.1.48 5.3.1.24 [m] 4.1.1.48 4.1.1.23 4.2.1.20 5.1.3.1</td>
<td>1pii 1ggs 1dv7 4.2.1.20</td>
<td>PF00218 PF0218 PF00218</td>
</tr>
<tr>
<td>Aldolase family</td>
<td>N-acetylenuraminyl lyase Dihydrdipicolinate synthase Fructose-bisphosphate aldolase class I Transaldolase Type I dehydroquinate dehydratase Ribulose-phosphate 3-epimerase d-Aminolevulinic acid dehydratase Fructose-bisphosphate aldolase class II Phe-sensitive 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase 3-Deoxy-D-manno-octulosonate-8-phosphate synthase</td>
<td>4.1.3.3 4.1.2.13 4.1.2.10 4.2.1.10 4.1.2.13 4.1.2.13 4.1.2.13 4.1.2.13</td>
<td>1nal 1fb 1qe 1aw5 1dos 1qr7 1d9e</td>
<td>PF00701 PF00274 PF01487 PF00490 PF0116 PF00793 PF00490</td>
</tr>
<tr>
<td>TIM barrel glycosyl hydrolases</td>
<td>Narbonin Non-enzyme Concanavalin B Non-enzyme Hevamine (chitinase/lysozyme) Chitinase Endo-β-N-acetylglucosaminidase Cyclodextrin glycosyltransferase β-Amylase Chitobiase β-Glucosidase β-Glucosidase 6-Phospho-β-galactosidase Myrosinase Endo-1,4-β-glucanase Exo-1,3-β-glucanase Endo-1,4-β-xylanase Exoglucanase/endo-1,4-β-xylanase β-Galactosidase β-Galactosidase Endo-1,3-β-glucanase Endo-1,3,1,4-β-glucanase β-Mannanase 4-α-Glucanotransferase</td>
<td>3.2.1.14 3.2.1.14 3.2.1.96 2.4.1.19 3.2.1.1 3.2.1.10 2.4.1.19 3.2.1.135 3.2.1.2</td>
<td>2hvm 1eq 1vjs 1pam 1fb 1uok 2ebn 1b2 1bvy</td>
<td>PF00192 PF00192 PF00704 PF0028 PF00137 PF0028 PF00704 PF0028 PF00137</td>
</tr>
<tr>
<td>Phosphoenol-pyruvate binding domains</td>
<td>Pyruvate, phosphate dikinase Pyruvate kinase Phosphoenolpyruvate carboxylase</td>
<td>2.7.9.1 2.7.1.40 4.1.1.31</td>
<td>1dk 1a49 1fiy</td>
<td>PF00391 PF00224 PF00311</td>
</tr>
<tr>
<td>Enolase family</td>
<td>Enolase Mandelate racemase Muconate cycloisomerase I Chloromucorace cycloisomerase</td>
<td>4.2.1.11 5.1.2.2</td>
<td>10n 1muc</td>
<td>PF00113 PF01188 PF01188</td>
</tr>
<tr>
<td>FMN-dependent oxidoreductases</td>
<td>(S)-2-Hydroxy-acid oxidase (glycolate oxidase) l-Lactate dehydrogenase (flavocytochrome b) Trimethylamine dehydrogenase NADPH dehydrogenase (old yellow enzyme) Dihydroorotate dehydrogenase</td>
<td>1.1.3.15 1.1.2.3</td>
<td>1gox 1fcb 2tm 1o ya</td>
<td>PF01070 PF01070 PF00724 PF00724</td>
</tr>
<tr>
<td>Metal-dependent hydrolases</td>
<td>Adenosine deaminase Phosphotriesterase Urease, z subunit</td>
<td>3.5.4.4 3.1.8.1</td>
<td>1a4m 1psc 3.5.1.5</td>
<td>PF00962 PF00216 PF00449</td>
</tr>
<tr>
<td>Type I PLP-dependent aspartate aminotransferase superfamily</td>
<td>Aspartate aminotransferase Tyrosine aminotransferase Aromatic-amino-acid aminotransferase L-Aminocyclopropane-1-carboxylate synthase Ornithine decarboxylase Tryptophanase Tyrosine phenol-lyase Cystathionine γ-synthase</td>
<td>2.6.1.1 2.6.1.5 2.6.1.57</td>
<td>1ars 1bw0 2ay1 4.4.1.14 4.1.1.17 1ax4 4.1.1.4 4.2.9.9</td>
<td>PF00155 PF00155 PF00155 PF00155 PF00155 PF00155 PF00155 PF00155</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Member</th>
<th>EC number(s)</th>
<th>PDB code</th>
<th>Pfam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cystathione β-lyase</td>
<td>4.4.1.8</td>
<td>1cl1</td>
<td>PF01053</td>
</tr>
<tr>
<td></td>
<td>Serine hydroxymethyltransferase</td>
<td>2.1.2.1</td>
<td>1bj4</td>
<td>PF00464</td>
</tr>
<tr>
<td></td>
<td>8-Amino-7-oxononanoate synthase</td>
<td>2.3.1.47</td>
<td>1bs0</td>
<td>PF00222</td>
</tr>
<tr>
<td></td>
<td>Glutamate-1-semialdehyde 2,1-aminomutase</td>
<td>5.4.3.8</td>
<td>2gsa</td>
<td>PF02020</td>
</tr>
<tr>
<td></td>
<td>2,2-Dialkylglycine decarboxylase</td>
<td>4.1.1.64</td>
<td>1d7u</td>
<td>PF02020</td>
</tr>
<tr>
<td></td>
<td>Ornithine aminotransferase</td>
<td>2.6.1.13</td>
<td>2oat</td>
<td>PF02020</td>
</tr>
<tr>
<td></td>
<td>4-Aminobutyrate aminotransferanse</td>
<td>2.6.1.19</td>
<td>1gtx</td>
<td>PF02020</td>
</tr>
<tr>
<td></td>
<td>Adenosylmethionine-8-amino-7-oxononanoate aminotransferase</td>
<td>2.6.1.62</td>
<td>1dty</td>
<td>PF02020</td>
</tr>
<tr>
<td></td>
<td>Phosphoserine aminotransferanse</td>
<td>2.6.1.52</td>
<td>1bh4</td>
<td>PF00266</td>
</tr>
<tr>
<td></td>
<td>3-Amino-5-hydroxybenzoic acid synthase</td>
<td>4.2.1</td>
<td>1bh9</td>
<td></td>
</tr>
<tr>
<td>ATP-dependent carboxylate-</td>
<td>Acetyl-CoA carboxylase, biotin carboxylase subunit</td>
<td>6.3.4.14</td>
<td>1bnc</td>
<td>PF00289</td>
</tr>
<tr>
<td>amine/thiol ligases</td>
<td>d-alanine-d-alanine ligase</td>
<td>6.3.2.4</td>
<td>1iow</td>
<td>PF01820</td>
</tr>
<tr>
<td></td>
<td>Glycinamide ribonucleotide synthetase</td>
<td>6.3.4.13</td>
<td>1gso</td>
<td>PF01071</td>
</tr>
<tr>
<td></td>
<td>Phosphoribosylaminimidazole carboxylase</td>
<td>4.1.1.21</td>
<td>1ber</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glutathione synthetase</td>
<td>6.3.2.3</td>
<td>1gsh</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Succinyl-CoA synthetase, β chain</td>
<td>6.2.1.5</td>
<td>2scu</td>
<td>PF00549</td>
</tr>
<tr>
<td></td>
<td>Carbamoyl-phosphate synthase, large chain</td>
<td>6.3.5.5</td>
<td>1c00</td>
<td>PF00289</td>
</tr>
<tr>
<td></td>
<td>Synapsin la</td>
<td>-</td>
<td>1auv</td>
<td>PF02078</td>
</tr>
<tr>
<td></td>
<td>Pyruvate, phosphate dikinase</td>
<td>2.7.9.1</td>
<td>1dik</td>
<td>PF01326</td>
</tr>
<tr>
<td>Thioredoxin superfamily</td>
<td>Thioredoxin</td>
<td>Non-enzyme</td>
<td>2trx</td>
<td>PF00085</td>
</tr>
<tr>
<td></td>
<td>Glutaredoxin</td>
<td>Non-enzyme</td>
<td>1kte</td>
<td>PF01323</td>
</tr>
<tr>
<td></td>
<td>Thiol-disulphide interchange protein DsbA</td>
<td>-</td>
<td>1fvk</td>
<td>PF00085</td>
</tr>
<tr>
<td></td>
<td>Protein disulphide isomerase</td>
<td>5.3.4.1</td>
<td>1mev</td>
<td>PF00085</td>
</tr>
<tr>
<td></td>
<td>Glutathione S-transferase</td>
<td>2.5.1.18</td>
<td>1gse</td>
<td>PF00043</td>
</tr>
<tr>
<td></td>
<td>Prostaglandin D-synthase</td>
<td>5.3.9.29</td>
<td>1pd2</td>
<td>PF00043</td>
</tr>
<tr>
<td></td>
<td>Glutathione peroxidase</td>
<td>1.11.1.9</td>
<td>1gp1</td>
<td>PF00255</td>
</tr>
<tr>
<td></td>
<td>Thioredoxin peroxidase 2</td>
<td>-</td>
<td>1q21</td>
<td>PF00578</td>
</tr>
<tr>
<td></td>
<td>Peroxidase Horf6</td>
<td>1.11.1.7</td>
<td>1prx</td>
<td>PF00578</td>
</tr>
<tr>
<td></td>
<td>Phosducin</td>
<td>Non-enzyme</td>
<td>2trc</td>
<td>PF02114</td>
</tr>
<tr>
<td></td>
<td>Calsequestrin</td>
<td>Non-enzyme</td>
<td>1a8y</td>
<td>PF01216</td>
</tr>
<tr>
<td>Phosphoglycerate mutase-like</td>
<td>Phosphoglycerate mutase</td>
<td>5.4.2.1</td>
<td>1qhf</td>
<td>PF03000</td>
</tr>
<tr>
<td></td>
<td>6-Phosphofructo-2-kinase/fructose-2,6-biphosphatase</td>
<td>2.7.1.105</td>
<td>1bif</td>
<td>PF03000</td>
</tr>
<tr>
<td></td>
<td>Prostatic acid phosphatase</td>
<td>3.1.3.2</td>
<td>1rpa</td>
<td>PF00328</td>
</tr>
<tr>
<td></td>
<td>3-Phytase</td>
<td>3.1.3.8</td>
<td>1ihp</td>
<td>PF00328</td>
</tr>
<tr>
<td>α/β-Hydrolases</td>
<td>Carboxylesterase</td>
<td>3.1.1.1</td>
<td>1auo</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Triacylglycerol lipase</td>
<td>3.1.1.3</td>
<td>1cul</td>
<td>PF00561</td>
</tr>
<tr>
<td></td>
<td>Acetylcholine esterase</td>
<td>3.1.1const</td>
<td>2ack</td>
<td>PF00135</td>
</tr>
<tr>
<td></td>
<td>Bile-salt-activated cholesterol esterase</td>
<td>3.1.1.3 const</td>
<td>2bce</td>
<td>PF00135</td>
</tr>
<tr>
<td></td>
<td>Para-nitrobenzyl esterase</td>
<td>3.1.1const</td>
<td>1qg1</td>
<td>PF00135</td>
</tr>
<tr>
<td></td>
<td>Brefeldin A esterase</td>
<td>-</td>
<td>1lkm</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Prolyl oligopeptidase</td>
<td>3.4.21.26</td>
<td>1qfm</td>
<td>PF03026</td>
</tr>
<tr>
<td></td>
<td>Serine carboxypeptidase i (cathepsin A)</td>
<td>3.4.16.5</td>
<td>1ivy</td>
<td>PF00450</td>
</tr>
<tr>
<td></td>
<td>Serine carboxypeptidase II</td>
<td>3.4.16.6</td>
<td>1wht</td>
<td>PF00450</td>
</tr>
<tr>
<td></td>
<td>Proline aminopeptidase</td>
<td>3.4.11.5</td>
<td>1azw</td>
<td>PF00561</td>
</tr>
<tr>
<td></td>
<td>2-Hydroxy-6-oxo-6-phenylhexa-2,4-dienoate hydrolase</td>
<td>3.7.1.8</td>
<td>1c4x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Epoxide hydrolase</td>
<td>3.3.2.3</td>
<td>1cr6</td>
<td>PF00561</td>
</tr>
<tr>
<td></td>
<td>Halalkane dehalogenase</td>
<td>3.8.1.5</td>
<td>1b6g</td>
<td>PF00561</td>
</tr>
<tr>
<td></td>
<td>(S)-acetone-cyanohydrin lyase</td>
<td>4.1.2.39</td>
<td>1qj1</td>
<td>PF00561</td>
</tr>
<tr>
<td></td>
<td>Non-heme chloroperoxidase</td>
<td>1.11.1.10</td>
<td>1a8s</td>
<td>PF00561</td>
</tr>
<tr>
<td></td>
<td>Myristoyl-ACP-specific thioesterase</td>
<td>2.3.1.1</td>
<td>1ht1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dienes lactone hydrolase</td>
<td>3.1.1.45</td>
<td>1d1h</td>
<td>PF01738</td>
</tr>
<tr>
<td></td>
<td>Acetylxylan esterase</td>
<td>3.1.1.6</td>
<td>1bs9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cutinase (serine esterase)</td>
<td>3.1.1.1</td>
<td>1cex</td>
<td>PF01083</td>
</tr>
<tr>
<td>TPP-dependent enzymes</td>
<td>Pyruvate oxidase</td>
<td>1.2.3.3</td>
<td>1pox</td>
<td>PF02025</td>
</tr>
<tr>
<td></td>
<td>Pyruvate dehydrogenase</td>
<td>4.1.1.1</td>
<td>1pvd</td>
<td>PF02025</td>
</tr>
<tr>
<td></td>
<td>Benzoylformate dehydrogenase</td>
<td>4.1.1.16</td>
<td>1bzf</td>
<td>PF02025</td>
</tr>
<tr>
<td></td>
<td>Transketolase</td>
<td>2.2.1.1</td>
<td>1trk</td>
<td>PF00456</td>
</tr>
<tr>
<td></td>
<td>Pyruvate-dehydrogenase reductase</td>
<td>1.2.7.1</td>
<td>1bop</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2-Oxoisovalerate dehydrogenase, α subunit</td>
<td>1.2.4.4</td>
<td>1qso</td>
<td>PF00676</td>
</tr>
<tr>
<td></td>
<td>2-Oxoisovalerate dehydrogenase, β subunit</td>
<td>1.2.4.4</td>
<td>1qso</td>
<td></td>
</tr>
<tr>
<td>5'-Nucleases</td>
<td>DNA polymerase I</td>
<td>2.7.7.7</td>
<td>1bgx</td>
<td>PF01367</td>
</tr>
<tr>
<td></td>
<td>Ribonuclease H</td>
<td>3.1.26.4</td>
<td>1frt</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5'-exonuclease</td>
<td>3.1.11.3</td>
<td>1exn</td>
<td>PF01367</td>
</tr>
<tr>
<td></td>
<td>Flap endonuclease-1</td>
<td>-</td>
<td>1a76</td>
<td></td>
</tr>
</tbody>
</table>
enzyme superfamily provides one of the best examples of this exploitation, and illustrates the extent to which the nature of active-sites can vary between homologues. Members act on a broad range of DNA lesions, including ultra violet photoadducts, mismatches and oxidised and alkylated bases, but each enzyme is normally specific for a particular type of lesion. Endonuclease III has a polar active-site pocket filled with water molecules for the recognition of a wide range of damaged pyrimidines, that of 3-methyladenine DNA glycosylase II is rich in electron-donating aromatic residues for binding electron-deficient alkylated bases, and active-site residues in A/G-specific adenine glycosylase hydrogen-bond specifically to adenine (Mol et al., 1999).

Substrates are often bound to surface loops, notably in enzymes with the TIM barrel or Ω/β-hydrolase folds; this facilitates rapid evolutionary adaptation since the loops can vary whilst the structural integrity of the protein fold is maintained (Perona & Craik, 1997). This makes the

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Member</th>
<th>EC number(s)</th>
<th>PDB code</th>
<th>Pfam</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAD/NAD(P)H dependent disulphide oxidoreductase superfamily</td>
<td>Glutathione reductase</td>
<td>1.6.4.2</td>
<td>3grs</td>
<td>PF00070</td>
</tr>
<tr>
<td></td>
<td>Low molecular weight thioredoxin reductase</td>
<td>1.6.4.5</td>
<td>1de</td>
<td>PF00070</td>
</tr>
<tr>
<td></td>
<td>Trypanothione reductase</td>
<td>1.6.4.8</td>
<td>1fc</td>
<td>PF00070</td>
</tr>
<tr>
<td></td>
<td>Dihydrolipoamide dehydrogenase</td>
<td>1.8.1.4</td>
<td>3ad</td>
<td>PF00070</td>
</tr>
<tr>
<td></td>
<td>Flavocytochrome c sulphide dehydrogenase</td>
<td>1.8.2.-</td>
<td>1ec</td>
<td>PF00070</td>
</tr>
<tr>
<td></td>
<td>NADH peroxidase</td>
<td>1.11.1.1</td>
<td>1px</td>
<td>PF00070</td>
</tr>
<tr>
<td></td>
<td>Mercier ion reductase*</td>
<td>1.16.1.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Adrenodoxin reductase</td>
<td>1.18.1.2</td>
<td>1jc</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Trimethylamine dehydrogenase</td>
<td>1.5.99.7</td>
<td>2tmol</td>
<td></td>
</tr>
<tr>
<td>Zn peptidases</td>
<td>Carboxypeptidase A</td>
<td>3.4.17.1</td>
<td>2tc</td>
<td>PF00246</td>
</tr>
<tr>
<td></td>
<td>Carboxypeptidase B</td>
<td>3.4.17.2</td>
<td>1ns</td>
<td>PF00246</td>
</tr>
<tr>
<td></td>
<td>Carboxypeptidase A2</td>
<td>3.4.17.15</td>
<td>1aye</td>
<td>PF00246</td>
</tr>
<tr>
<td></td>
<td>Carboxypeptidase T</td>
<td>3.4.17.18</td>
<td>1obr</td>
<td>PF00246</td>
</tr>
<tr>
<td></td>
<td>Bacterial leucyl aminopeptidase</td>
<td>3.4.11.10</td>
<td>1amp</td>
<td>PF01546</td>
</tr>
<tr>
<td></td>
<td>Carboxypeptidase G2</td>
<td>3.4.17.11</td>
<td>1cg2</td>
<td>PF01546</td>
</tr>
<tr>
<td></td>
<td>Cytosol aminopeptidase</td>
<td>3.4.11.1 3.4.11.5 [s]</td>
<td>1lam</td>
<td>PF00883</td>
</tr>
<tr>
<td></td>
<td>Aminopeptidase</td>
<td>3.4.11.-</td>
<td>1xjo</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Transferrin receptor protein</td>
<td>Non-enzyme</td>
<td>1cx8</td>
<td></td>
</tr>
<tr>
<td>DD-peptidase/β-lactamase superfamily</td>
<td>α-alanyl-β-alanine carboxypeptidase/ transpeptidase</td>
<td>3.4.16.4</td>
<td>3pte</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-Lactamase</td>
<td>3.5.2.6</td>
<td>1dj</td>
<td>PF00144</td>
</tr>
<tr>
<td></td>
<td>Penicillin-binding protein</td>
<td>-</td>
<td>1pm</td>
<td>PF09065</td>
</tr>
<tr>
<td>Phosphohistidine domains of PEP-utilising enzymes</td>
<td>Pyruvate, phosphate dikinase</td>
<td>2.7.9.1</td>
<td>1dk</td>
<td>PF00391</td>
</tr>
<tr>
<td></td>
<td>Phosphoenolpyruvate-protein phosphotransferase</td>
<td>2.7.3.9</td>
<td>2ezb</td>
<td>PF00391</td>
</tr>
<tr>
<td>Medium-chain alcohol dehydrogenases</td>
<td>Quinone oxidoreductase</td>
<td>1.6.5.5</td>
<td>1qor</td>
<td>PF00107</td>
</tr>
<tr>
<td></td>
<td>Alcohol dehydrogenase</td>
<td>1.1.1.1</td>
<td>1deh</td>
<td>PF00107</td>
</tr>
<tr>
<td></td>
<td>Alcohol dehydrogenase class III</td>
<td>1.1.1.1 1.1.2.1.1 [s]</td>
<td>1teh</td>
<td>PF00107</td>
</tr>
<tr>
<td></td>
<td>NADP-dependent alcohol dehydrogenase</td>
<td>1.1.1.2</td>
<td>1kev</td>
<td>PF00107</td>
</tr>
<tr>
<td></td>
<td>Glucose dehydrogenase*</td>
<td>1.1.1.47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein tyrosine phosphatase superfamily</td>
<td>Protein-tyrosine phosphatase</td>
<td>3.1.3.48</td>
<td>1bzh</td>
<td>PF00102</td>
</tr>
<tr>
<td></td>
<td>Dual specificity protein phosphatase</td>
<td>3.1.3.16 3.1.3.48 [s]</td>
<td>1hvr</td>
<td>PF00782</td>
</tr>
<tr>
<td>Crotonase-like</td>
<td>4-Chlorobenzoyl-CoA dehalogenase</td>
<td>3.8.1.6</td>
<td>1ny</td>
<td>PF00378</td>
</tr>
<tr>
<td></td>
<td>Enoyl-CoA hydratase</td>
<td>4.2.1.17</td>
<td>2dub</td>
<td>PF00378</td>
</tr>
<tr>
<td></td>
<td>Δ3,5-A2,4-dienoyl-CoA isomerase</td>
<td>5.3.3.-</td>
<td>1dci</td>
<td>PF00378</td>
</tr>
<tr>
<td></td>
<td>ATP-dependent Clp protease, proteolytic subunit</td>
<td>3.4.21.92</td>
<td>1tyf</td>
<td>PF00574</td>
</tr>
<tr>
<td>Creatinase/methionine aminopeptidase superfamily</td>
<td>Creatinase</td>
<td>3.5.3.3</td>
<td>1cm</td>
<td>PF00557</td>
</tr>
<tr>
<td></td>
<td>Methionine aminopeptidase</td>
<td>3.4.11.18</td>
<td>1xgs</td>
<td>PF00557</td>
</tr>
<tr>
<td></td>
<td>Aminopeptidase P</td>
<td>3.4.11.9</td>
<td>1a29</td>
<td>PF00557</td>
</tr>
</tbody>
</table>

[m] denotes “multienzymes” (enzymes with two or more catalytic functions contributed by distinct domains and/or separate subunits) and [s] denotes “single enzymes” (enzymes catalysing two or more reactions using the same catalytic site). Note that some enzymes have different types, isoforms and classes, e.g. β-lactamase classes A and C both belong to the DD-peptidase/β-lactamase superfamily, and prokaryotic type I and eukaryotic type II methionine aminopeptidases are single-domain and two-domain proteins, respectively. To restrict the length of the Table, with few exceptions only a single entry for each distinct function is given, with a representative PDB entry. Where two or more homologous protein chains belonging to the same enzyme complex have different functional roles, the individual chains are listed. Readers are referred to www.biochem.ucl.ac.uk/bsm/FAM-EC/for a more comprehensive list of superfamily members. Pfam accession numbers (Bateman et al., 2000) are provided for those PDB code with a SWISS-PROT entry, and these illustrate the sequence variability of the superfamily.

* Those proteins of known structure which do not have a PDB entry.

** During analysis it became apparent that discoidin-like domains are not catalytic, but they have been included to illustrate some of the complexities of functional annotation, and of the analysis of structure/function relationships.

* The Rieske-like domains essentially function as electron transfer agents; the catalytic site of naphthalene 1,2-dioxygenase, a subunit is in a domain attached to the C terminus of the Rieske-like domain.

DNA glycosylase II is rich in electron-donating aromatic residues for binding electron-deficient alkylated bases, and active-site residues in A/G-specific adenine glycosylase hydrogen-bond specifically to adenine (Mol et al., 1999). Substrates are often bound to surface loops, notably in enzymes with the TIM barrel or Ω/β-hydrolase folds; this facilitates rapid evolutionary adaptation since the loops can vary whilst the structural integrity of the protein fold is maintained (Perona & Craik, 1997). This makes the
wide variability of the substrates, which are bound to the core of cytochrome P450s all the more remarkable; collectively they act on fatty acids, steroids, prostaglandins and many different drugs. The P450s show large differences in the orientations of their α-helices to accommodate these diverse substrates (Poulos, 1995). Presumably they have been under intense evolutionary pressure to evolve, since many P450s play a crucial role in detoxification.

Only enzymes and their ligands are considered here. Non-enzymes may bind substrates which bear no resemblance to those of their catalytically active homologues, e.g., a single-domain

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Domain</th>
<th>Domain</th>
<th>Domain</th>
<th>Subunit</th>
<th>Motif</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fumarase/aspartase</td>
<td>83-195</td>
<td>2-3</td>
<td>2-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Helix-hairpin-helix base-excision</td>
<td>51-113</td>
<td>2-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA repair enzymes</td>
<td>146-512</td>
<td>1-2</td>
<td>*2-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-borne di-iron carboxylate proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cytochrome P450s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ferredoxin-NADP reductase</td>
<td>2-5</td>
<td>1-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(FN) modules</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cupredoxins</td>
<td>96-214</td>
<td>S</td>
<td>1-6</td>
<td>*2-24</td>
<td>*</td>
</tr>
<tr>
<td>Discoidin-like domains</td>
<td>*</td>
<td>3-9</td>
<td>*</td>
<td>1-2</td>
<td>*</td>
</tr>
<tr>
<td>Rieske-like iron-sulphur domains</td>
<td>2</td>
<td>*</td>
<td>6-22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hexapeptide repeat proteins</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ribulose-phosphate-binding TIM barrels</td>
<td>S</td>
<td>1-2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aldolase superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TIM barrel glycolyl hydrolases</td>
<td>*</td>
<td>1-5</td>
<td>*</td>
<td>1-4</td>
<td>*</td>
</tr>
<tr>
<td>Phosphoenol-pyruvate binding domains</td>
<td>274-883</td>
<td>1-5</td>
<td>2-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enolase superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FMN-dependent oxidoreductases</td>
<td>*</td>
<td>1-3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metal-dependent hydrolases</td>
<td>1-2</td>
<td>*</td>
<td>1-9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Type I PLP-dependent aspartate aminotransferase superfamily</td>
<td>2-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATP-dependent carboxylate-amine/thiol ligases</td>
<td>121-288, 69-155</td>
<td>S</td>
<td>3-8</td>
<td>*</td>
<td>1-4</td>
</tr>
<tr>
<td>Thioredoxin superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphoglycerate mutase-like α/β-hydrolases</td>
<td>75-184</td>
<td>S</td>
<td>1-4</td>
<td>*</td>
<td>1-N</td>
</tr>
<tr>
<td>2-431</td>
<td>1-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>197-547</td>
<td>1-2</td>
<td>*</td>
<td>1-8</td>
<td>*</td>
<td></td>
</tr>
<tr>
<td>TPP-dependent enzymes 5'-Nucleases</td>
<td>178-395</td>
<td>S</td>
<td>1-5</td>
<td>*</td>
<td>2-4</td>
</tr>
<tr>
<td>2-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FAD/NAD(P)H-dependent disulphide oxidoreductase superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zn peptidases</td>
<td>1-2</td>
<td>1-5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>α/β-Peptidase/β-lactamase superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphohistidine domains of PEP-utilising enzymes</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium-chain alcohol dehydrogenases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein tyrosine phosphatase superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crotonase-like</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creatinase/methionine aminopeptidase superfamily</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The Table indicates those superfamilies in which (i) the sizes of at least one domain common to all members vary by at least twofold, with the size range(s) in residues as shown; (ii) at least one member contains two or more homologous domains along a single polypeptide chain indicating a duplication event (S: duplicated domains have evolved specialised functional roles); (iii) domain organisation varies between members, with the range of the number of domains along a single polypeptide chain as shown; (iv) the relative location of two or more domains along the polypeptide chain differs (v) subunit assembly varies, with the range of the number of subunits within a complex as shown, where N denotes a large number (an asterisk (*) indicates that one or more members are hetero-oligomers, thus the variation results not only from a differing number of identical chains within the assembly); note that data regarding subunit assembly were extracted from the literature for species to which the PDB codes listed in Table 2 correspond; identifying the oligomerisation state for all species was beyond the scope of this analysis; (vi) sequence and/or structural data indicate motif duplication. For some proteins the structural domain organisation of the entire chain is unknown, and information regarding modular content, on the basis of sequence comparisons, is extracted from the literature and Pfam (Bateman et al., 2000), where available. Similarly, for some proteins subunit assembly is unknown, or it is not clear from the literature, and so the Table is incomplete in part (see www.biochem.ucl.ac.uk/bsm/FAM-EC/ for more details).
protein involved in DNA protection during starvation displays significant sequence and structural similarity to various non-heme di-iron carboxylate enzymes (Grant et al., 1998), which collectively bind fatty acids, methane and other hydrocarbons.

**Reaction chemistry**

Here we make a distinction between reaction chemistry and catalytic mechanism. Chemistry refers to the overall strategy of changing substrate into product, and the nature of the intermediates involved, whereas catalytic mechanism describes the roles played by specific residues in the active-site.

Far more common than the conservation of substrate binding is the conservation of reaction chemistry within the 31 enzyme superfamilies studied. There are 27 superfamilies for which details of catalysis are known, and in four, the reaction chemistry is conserved (see Table 4). For example, catalysis by members of the protein tyrosine phosphatase superfamily involves nucleophilic attack by a conserved cysteine on the substrate and the thiol-phosphate intermediate formed is subsequently hydrolysed. In a further 18 superfamilies, the chemistry is “semi-conserved” (see below).

**Common chemistry in context of diverse reactions**

In these semi-conserved superfamilies, enzyme members utilise a common chemical strategy in the contexts of different overall transformations. Observations of this nature have been observed in other enzyme superfamilies (e.g. see Lee et al., 1998; Aravind et al., 1998; Schofield & Zhang, 1999). Typically, it is the initial catalytic step that is conserved, whilst the reaction paths that follow vary, sometimes extensively, as illustrated by the examples given in Table 5. These superfamilies illustrate not only the versatility of protein folds with respect to function, but also the versatility of the chemistry involved, in that a single chemical step can be re-used in a number of contexts to provide completely different outcomes. The chemistry employed by members of the PLP-dependent amiotransferase superfamily is particularly versatile; the amino acid substrates may undergo covalency changes at the α, β or γ carbon atoms after formation of the external aldimine intermediate (Alexander et al., 1994), leading to a vast array of reactions catalysed by these enzymes, as seen in Figure 6.

In three superfamilies the similarity in reaction chemistry between enzyme members is more limited and may be described as “poorly conserved”. The crotonase-like superfamily includes structural and sequence members catalysing dehydratase, dehalogenase, isomerase, decarboxylase, and peptidase activities. Typical members act on a coenzyme A thioester, so the identification of ClpP protease as an evolutionary relative was unexpected (Murzin, 1998). Stabilisation of an oxyanion intermediate by a conserved oxyanion hole, in which the intermediate is hydrogen-bonded to two backbone amide groups, is the only functional similarity conserved across all members, and in contrast to those “semi-conserved” superfamilies listed in Table 5, the reaction paths to this intermediate vary widely (Babbitt & Gerlt, 1997). They include proton abstraction, peptide hydrolysis and nucleophilic aromatic addition, and involve different catalytic residues within the active-site. Similar observations have been made in the vicinal oxygen chelate superfamily (Babbitt & Gerlt, 1997); catalysis involves metal-assisted stabilisation of developing negative charge on a vicinal oxygen, but beyond this similarity, the nature of the intermediates and the chemistry involved are diverse. The thioredoxin superfamily provides yet another example. The common factor of the redox-active/catalytic members is their sulphur redox chemistry, which is employed to carry out transferase, peroxidase and isomerase reactions; during catalysis, they are likely to stabilise a cysteine thiolate (or selenolate ion), but the cysteine is an intramolecular active-site residue for some members, but from an external substrate, such as glutathione, in others. Not one active-site residue is conserved.
### Table 4. Conservation and variation of substrate specificity, reaction chemistry and catalytic residues

<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Substrate specificity</th>
<th>Co-factor</th>
<th>Similarity in reaction chemistry</th>
<th>Conservation of chemistry</th>
<th>Conservation of principal catalytic residues</th>
<th>Variability of functionally equivalent residues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fumarase/aspartase</td>
<td>Similar; moiety-carboxylate group</td>
<td>-</td>
<td>β-Elimination reactions involving cleavage of C-N or C-O bonds to the carboxylate group; β-proton abstraction</td>
<td>Semi-conserved</td>
<td>No</td>
<td>*</td>
</tr>
<tr>
<td>Helix-hairpin-helix base-excision DNA repair enzymes</td>
<td>Type-DNA substrates; diverse</td>
<td>-</td>
<td>Nucleophilic activation involving conserved aspartate</td>
<td>Semi-conserved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Non-heme di-iron carboxylate proteins</td>
<td>Diverse</td>
<td>Two Fe atoms</td>
<td>O₂/H₂O₂ activation and di-iron oxidation</td>
<td>Semi-conserved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Cytochrome P450s</td>
<td>Diverse</td>
<td>heme-thiolate</td>
<td>Activated iron-oxygen species [with exception of nitric oxide reductase, but its reaction mechanism follows the same principle]</td>
<td>Semi-conserved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Ferredoxin-NADP reductase (FNR) modules</td>
<td>N/A</td>
<td>FAD or FMN</td>
<td>Hydride transfer between flavin and NAD(P)(H); direction varies</td>
<td>Conserved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Cupredoxins</td>
<td>Diverse</td>
<td>Cu</td>
<td>Intramolecular electron transfer; Cu-dependent oxidoreductase activity</td>
<td>Semi-conserved</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Discoidin-like domains</td>
<td>Type-cell-surface carbohydrates</td>
<td>N/A</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>“Rieske”-like iron-sulphur domains</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>-</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Hexapeptide repeat proteins</td>
<td>Diverse</td>
<td>Unconserved metal-binding site</td>
<td>Varied</td>
<td>Unconserved</td>
<td>No</td>
<td>*</td>
</tr>
<tr>
<td>Ribulose-phosphate-binding TIM barrels</td>
<td>Similar; moiety-ribulose phosphate</td>
<td>-</td>
<td>Varied</td>
<td>Unconserved</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Aldolase superfamily</td>
<td>Diverse; moiety-carboxylate group</td>
<td>Unconserved metal-binding sites</td>
<td>Carbonyl chemistry</td>
<td>Poorly-conserved</td>
<td>No</td>
<td>*</td>
</tr>
<tr>
<td>TIM barrel glycosyl hydrolases</td>
<td>Type-carbohydrates</td>
<td>-</td>
<td>General acid catalysis-glutamate protonates scissile glycosidic oxygen</td>
<td>Semi-conserved</td>
<td>No</td>
<td>*</td>
</tr>
<tr>
<td>Phosphoenol-pyruvate binding domains</td>
<td>Conserved</td>
<td>Divalent metal</td>
<td>Detailed mechanisms unknown</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Enolase superfamily</td>
<td>Diverse; moiety-carboxylate group</td>
<td>Divalent metal</td>
<td>Metal-assisted abstraction of proton to carboxylic acid</td>
<td>Semi-conserved</td>
<td>Semi</td>
<td>*</td>
</tr>
<tr>
<td>FMN-dependent oxidoreductases</td>
<td>Diverse</td>
<td>FMN</td>
<td>FMN-dependent oxidoreductase activity</td>
<td>Semi-conserved</td>
<td>No</td>
<td>*</td>
</tr>
<tr>
<td>Metal-dependent hydrolases</td>
<td>Diverse</td>
<td>One or two divalent metal atoms</td>
<td>Metal-assisted activation of water for nucleophilic attack on substrate</td>
<td>Semi-conserved</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Evolution of Function in Protein Superfamilies
<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Type-amino acids; diverse</th>
<th>PLP</th>
<th>Covalent binding of PLP to conserved lysine and formation of aldimine intermediate with substrate</th>
<th>Semi-conserved Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP-dependent carboxylate-amine/thiol ligases</td>
<td>Diverse; moiety - acceptor, carboxylate group</td>
<td>-</td>
<td>Formation of an acylphosphate intermediate; C-N or C-S bond formation</td>
<td>Semi-conserved Yes</td>
</tr>
<tr>
<td>Thioredoxin superfamily</td>
<td>Diverse</td>
<td>-</td>
<td>Sulphur redox chemistry; stabilisation of cysteine thiolate (or selenolate)</td>
<td>Poorly-conserved No</td>
</tr>
<tr>
<td>Phosphoglycerate mutase-like</td>
<td>Diverse; moiety- phosphate group(s)</td>
<td>-</td>
<td>Conserved histidine forms covalent phosphohistidine intermediate</td>
<td>Semi-conserved Yes</td>
</tr>
<tr>
<td>α/β-Hydrolases</td>
<td>Diverse</td>
<td>-</td>
<td>Substrate undergoes nucleophilic attack by the nucleophile in a nucleophile-His-acid catalytic triad</td>
<td>Semi-conserved Semi</td>
</tr>
<tr>
<td>TPP-dependent enzymes</td>
<td>Diverse; moiety- carbonyl group</td>
<td>TPP</td>
<td>Conserved glutamate activates TPP for attack on the substrate carbonyl, leading to C-C bond cleavage and formation of a TPP-aldehyde intermediate</td>
<td>Semi-conserved Yes</td>
</tr>
<tr>
<td>5'-Nucleases</td>
<td>Type-DNA substrates</td>
<td>Two divalent metal atoms</td>
<td>Detailed mechanisms unknown</td>
<td>Semi-conserved No</td>
</tr>
<tr>
<td>FAD/NAD(P)(H) dependent disulphide oxidoreductase superfamily</td>
<td>Diverse</td>
<td>FAD</td>
<td>FAD/NAD(P)(H)-dependent oxidoreductase activity</td>
<td>Semi-conserved Yes</td>
</tr>
<tr>
<td>Zn peptidases</td>
<td>Diverse; moiety- peptide bond</td>
<td>One or two Zn metal atoms</td>
<td>Nucleophilic attack of Zn-activated water on substrate carbonyl; peptide bond hydrolysis</td>
<td>Conserved Semi</td>
</tr>
<tr>
<td>DD-peptidase/β-lactamase superfamily</td>
<td>Diverse; moiety- peptide bond</td>
<td>-</td>
<td>Nucleophilic attack of a conserved serine on substrate carbonyl, leading to formation of an acyl-enzyme intermediate</td>
<td>Conserved Yes</td>
</tr>
<tr>
<td>Phosphohistidine domains of PEP-utilising enzymes</td>
<td>N/A</td>
<td>Divalent metal</td>
<td>Phosphotransfer involving phosphohistidine intermediate</td>
<td>Semi-conserved Yes</td>
</tr>
<tr>
<td>Medium-chain alcohol dehydrogenases</td>
<td>Diverse</td>
<td>Unconserved Zn-binding site</td>
<td>NAD(P)(H)-dependent alcohol dehydrogenation or aldehyde/ketone reduction</td>
<td>Semi-conserved No</td>
</tr>
<tr>
<td>Protein tyrosine phosphatase superfamily</td>
<td>Type-phosphorylated proteins</td>
<td>-</td>
<td>Conserved cysteine forms thiol-phosphate intermediate which is subsequently hydrolysed</td>
<td>Conserved Yes</td>
</tr>
<tr>
<td>Crotonase-like</td>
<td>Diverse; moiety- carbonyl group</td>
<td>-</td>
<td>Stabilisation of oxyanion intermediate</td>
<td>Poorly-conserved No</td>
</tr>
<tr>
<td>Creatinase/methionine aminopeptidase superfamily</td>
<td>Diverse; moiety- C-N bond</td>
<td>Unconserved binuclear metal site</td>
<td>C-N bond hydrolysis via tetrahedral intermediate</td>
<td>Semi-conserved No</td>
</tr>
</tbody>
</table>

The nature of conservation of substrate specificity and reaction chemistry usually refer to the most distantly related members of the superfamily. More closely related enzymes may exhibit more extensive similarities in function. Structural conservation of the principal catalytic residues across a superfamily is sometimes difficult to qualify, and this column provides only a rough guide (see www.biochem.ucl.ac.uk/bsm/FAM-EC/ for more details).
<table>
<thead>
<tr>
<th>Superfamily</th>
<th>Common reaction step</th>
<th>Enzyme</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-heme di-iron carboxylate proteins</td>
<td>$O_2/H_2O_2$ activation and di-iron oxidation</td>
<td>Methane monoxygenase</td>
<td>Hydroxylation of methane</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\Delta(9)$ Stearoyl-acyl carrier protein desaturase</td>
<td>Oxidative desaturation of a fatty acid</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ribonucleotide reductase</td>
<td>Formation of an active-site tyrosine radical; this is involved in a long-range electron transfer with the substrate bound to a separate subunit</td>
</tr>
<tr>
<td>Cytochrome P450s</td>
<td>Activated iron-oxygen species (with exception of nitric oxide reductase, but its reaction mechanism follows the same principle (Halkier, 1996))</td>
<td>Cytochrome P450-cam</td>
<td>Hydroxylation (monoxygenase activity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nitric oxide reductase</td>
<td>Nitric oxide reduction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Berbamunine synthase*</td>
<td>Oxidative dimerisation (oxidase activity)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Allene oxide synthase*</td>
<td>Epoxide formation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Thromboxane-A synthase*</td>
<td>Isomerisation (intramolecular oxidoreduction)</td>
</tr>
<tr>
<td>Enolase superfamily</td>
<td>Metal-assisted abstraction of proton $\alpha$ to carboxylic acid</td>
<td>Enolase</td>
<td>$\beta$-Elimination of water</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Methylaspartate ammonia-lyase*</td>
<td>$\beta$-Elimination of ammonia</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Mandelate racemase</td>
<td>Racemisation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D-Glucose dehydratase</td>
<td>Cycloisomeration</td>
</tr>
<tr>
<td></td>
<td></td>
<td>D-Glucose dehydratase</td>
<td>Epimerisation</td>
</tr>
<tr>
<td>Type I PLP-dependent aspartate</td>
<td>Covalent binding of PLP to conserved lysine and formation of aldime intermediate with substrate</td>
<td>Aspartate aminotransferase</td>
<td>Transamination</td>
</tr>
<tr>
<td>aminotransferase superfamily</td>
<td></td>
<td>Ornithine decarboxylase</td>
<td>$\alpha$-Decarboxylation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Serine hydroxymethyltransferase</td>
<td>Retro-aldo cleavage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tryptophanase</td>
<td>$\beta$-Elimination</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Methionine $\gamma$-lyase*</td>
<td>$\gamma$-Elimination</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cystathione $\gamma$-synthase</td>
<td>$\gamma$-Replacement</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Glutamate-1-semialdehyde 2,1-amino mutase</td>
<td>Isomerisation (intramolecular exchange)</td>
</tr>
<tr>
<td>TPP-dependent enzymes</td>
<td>Conserved glutamate activates TPP for attack on the substrate carbonyl, leading to C-C bond cleavage and formation of a TPP-aldehyde intermediate</td>
<td>Pyruvate oxidase</td>
<td>Oxidative decarboxylation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Pyruvate decarboxylase*</td>
<td>Decarboxylation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Transketolase</td>
<td>Ketol group transfer</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Benzaldehyde lyase*</td>
<td>Reverse aldol condensation</td>
</tr>
<tr>
<td>$\alpha$/ $\beta$-Hydrolases</td>
<td>Conserved histidine activates the catalytic nucleophile which forms a covalent intermediate with the substrate</td>
<td>Lipase</td>
<td>Carboxylic ester hydrolysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Haloalkane dehalogenase</td>
<td>Halide hydrolysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Epoxide hydrolyase</td>
<td>Ether hydrolysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Serine carboxypeptidase</td>
<td>Peptide hydrolysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(S)-Acetone-cyano hydrin lyase</td>
<td>Decomposition of hydroxynitrile</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Non-heme chloroperoxidase</td>
<td>Halogenation of organic compounds</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Myristoyl-ACP-specific thioesterase</td>
<td>Acyl transfer</td>
</tr>
</tbody>
</table>

The Table summarizes the functional variation of seven superfamilies in which the reaction chemistry is described as “semi-conserved”. Members within a superfamily catalyse a common reaction step in the context of diverse chemical transformations. An asterisk (*) indicates enzymes of unknown structure. Note that these sequence relatives are included here to illustrate the diversity of reactions catalysed by these superfamilies, but they are not otherwise included for analysis in the summary Tables presented in this paper.
across all members, an observation that has been made for other enzyme homologues (Murzin, 1998).

Reactions catalysed by members of the aldolase superfamily involve carbonyl chemistry. The details of the reactions, the nature of the principal reaction intermediates as well as the catalytic residues are variable, and so the chemistry of this superfamily is described as poorly conserved. This superfamily is discussed in more detail below.

Variation in chemistry

Within two superfamilies, the ribulose-phosphate-binding barrels and the hexapeptide repeat proteins, the reaction chemistry is completely unconserved in at least one pair of enzymes. In this second group of proteins, there is absolutely no correlation in catalytic function between two subfamilies. The enzymes comprise a series of hexapeptide motifs which encodes a solenoid-like left-handed β-helix domain, the number of “coils” varying between five and ten. Typical members of enzymes include also an archaeal carbonic anhydrase (EC 4.2.1.1) which catalyses reversible Zn-dependent hydration of carbon dioxide. Evidence for a common ancestry is provided by the specific sequence and structural features of this superfamily; all function as trimers with each active-site located at the interface of two subunits, the crossovers between strands in the solenoid-like structure are left-handed which is observed very rarely, and each strand corresponds to a poorly conserved hexapeptide motif. The disparate functions may have evolved by an ancient gene duplication event of an ancestral enzyme, or alternatively through the independent duplication and fusion of a variable number of hexapeptide motifs.

No two enzymes of the ribulose-phosphate-binding TIM barrel superfamily shows any similarity in reaction chemistry. Transformations catalysed are diverse, but they involve similar substrates. Accordingly, of all of the structural loops in the binding-site region, those contributing residues directly involved in catalysis are the least structurally conserved (Wilmanns et al., 1992).

Conservation of catalytic residues but diversity in function

The α/β hydrolase superfamily is one of the most structurally and functionally divergent of superfamilies known but catalysis invariably involves a catalytic triad comprising a nucleophile, an acidic residue and a conserved histidine (Ollis et al., 1992; Nardini & Dijkstra, 1999; Heikinheimo et al., 1999). Each active site contains also an oxyanion hole formed usually by the backbone amides of two amino acid residues, required for the stabilisation of the oxyanion intermediate formed in the reaction. The catalytic triad typically comprises a Ser, His and Asp, although variations are observed, and these same three residues are used by many functionally distinct enzymes acting on diverse substrates, including various peptidases (EC 3.4.-.-), lipases (EC 3.1.1.3), a thioesterase (EC 2.3.1.-) and haloperoxidases (EC 1.11.1.-) (see Table 5).

In enzymes of the non-heme di-iron carboxylate protein superfamily, the catalytic residues are strictly conserved; they are characterised as having a four-helix bundle core, comprising a duplicated two-helix iron-binding motif (Nordlund et al., 1990). Catalysis involves oxygen activation and di-iron oxidation, a chemical strategy employed to carry out various enzyme reactions (see Table 5). The specific geometric properties of each active-site dictate the fate of the peroxy intermediates. In some non-enzyme members of this superfamily this di-iron site is lacking, whilst in others it is conserved, such as the iron-storage protein ferritin. This uses its ferroxidase activity to allow storage of iron as an insoluble oxide in the central cavity of its oligomeric structure (Banyard et al., 1978; Ford et al., 1984).

Non-equivalence of catalytic residues despite similarity in function

One might expect that any similarity in reaction chemistry displayed by homologous enzymes is mediated by common functional groups conserved through evolution, and so at least some aspects of the mechanisms of these enzymes would be identical (Hasson et al., 1998a,b). Thus, a surprising find is the number of superfamilies in which there is poor positional conservation of residues which play equivalent catalytic roles in related proteins. Variability of this nature has been commented upon elsewhere, with reference to the enolase superfamily; the lysine residues which have equivalent enolate-stabilisation roles in enolase and mandelate racemase are on different strands (Hasson et al., 1998b). In this analysis, catalytic residue variation is observed in at least 12 superfamilies (see Table 4). In several cases, whilst the functionally equivalent residues are located at non-homologous positions in the structural scaffold, the residues are identical. A number of cases involve proteins with apparently identical biochemical

Catalytic machinery

Two alternative situations with regard to the conservation or divergence of function and active-site architectures are observed. The same active-site framework may be used to catalyse a host of diverse activities, and conversely, different catalytic apparatus may exist in related proteins with very similar functions.
functions, as defined by their EC numbers, but note that such enzymes can differ in their substrate preferences.

Both chloramphenicol acetyltransferase (PaXAT) (EC 2.3.1.28) and UDP-N-acetylglucosamine acyltransferase (LpxA) (EC 2.3.1.129) of the hexapeptide repeat protein superfamily function as acetyltransferases and contain an essential histidine residue putatively involved in deprotonation of a hydroxyl group in their respective substrates. However, these residues are located at different points within the protein fold; in LpxA, the histidine is in a hexapeptide repeat in the core of the domain (Wyckoff & Raetz, 1999), whereas in PaXAT, it is located in a loop which projects out from the solenoid structure (Beaman et al., 1998) (see Figure 7).

Members of the fumarase/aspartase superfamily contain three partially conserved sequence regions, and these motifs, contributed by three separate subunits, come together to form the active-site in the homotetrameric enzyme complexes (Simpson et al., 1994). Fumarase (EC 4.2.1.2) and aspartase (EC 4.3.1.1) catalyse reversible \( \beta \)-elimination reactions with the release of fumarate, and these reactions differ only in the nature of the leaving group (water versus ammonia). Despite their functional similarity and a high sequence identity of 38% (Woods et al., 1986), catalysis is mediated by different residues in the two proteins. Some residues essential for catalysis in fumarase are not conserved and lack the necessary functionality in aspartase (Jayasekera et al., 1997), and the precise details of the aspartase mechanism are still unknown.

Glycosyl hydrolases (EC 3.2.1.-) have been identified as the most widespread group of enzymes (Hegyi & Gerstein, 1999) and they have been classified into 70 families on the basis of amino acid sequence similarities (Henrissat, 1991). Members of several families adopt the ubiquitous TIM barrel fold and they have been clustered together into the same homologous superfamily in both CATH and SCOP (Murzin et al., 1995). In the literature, however, a common ancestry for these proteins has been doubted given the variable location of their functional groups; all enzyme members use a catalytic glutamate to donate a proton to the scissile glycosidic oxygen atom, and in the crystal structures determined thus far, this residue is situated at the C terminus of, or on the loops that follow, \( \beta \)-strands four or five in the central \( \beta \)-barrel. The Asp or Glu nucleophile required for the formation of a glycosyl-enzyme intermediate in “retaining” glycosyl hydrolases is located on or after strands four, six or seven. Given the observed positional variability of catalytic residues in other enzyme superfamilies, a shared ancestry cannot be ruled out. The retaining enzyme hevamine lacks the nucleophile present in its close relatives, not because it is located elsewhere in the fold, but it is missing altogether. Instead, catalysis by hevamine proceeds via a charged intermediate which is stabilised by the N-acetyl group in the substrate (Terwisscha van Scheltinga et al., 1996). This phenomenon is known as substrate-assisted catalysis, and renders the “missing” amino acid unimportant.

In most flavoenzymes involved in dehydrogenation reactions, the N5 atom of the flavin is within hydrogen bond distance from a hydrogen bond donor, regardless of the enzyme folding topology (Fraaije & Mattevi, 2000). Conservation of this interaction reflects its importance in catalysis, and the interaction exists in the FMN-dependent oxidoreductases of the TIM barrel fold, with the exception of glycolate oxidase. In three members of this superfamily, the donor is a backbone amide located in a loop that follows strand one, but in dihydroorotate dehydrogenase A from Lactococcus lactis, the donor is a nitrogen atom in a lysine side-chain located after strand two.

The “migration” of catalytic residues is likely to have occurred in the \( \alpha/\beta \) hydrolase superfamily (see Figure 8). A subfamily of lipases and esterases have Ser, His and Asp/Glu in their active sites, with the Asp or Glu acid located in the “usual” position following the seventh strand. In a more

![Figure 7](image-url) MOLSCRIPT (Kraulis, 1991) diagrams of the homologous enzymes (a) chloramphenicol acetyltransferase (PaXAT), and (b) UDP-N-acetylglucosamine acyltransferase (LpxA). The catalytic histidine residues putatively involved in deprotonation of the substrate hydroxyl are shown in ball-and-stick and circled in blue.
distant homologue, human pancreatic lipase (HPL), the Ser and His are conserved, yet the Asp is on a loop after strand six. The ancestral acidic Asp residue on strand seven remains as an evolutionary relic, and points away from the catalytic His, unable to make hydrogen-bonding contact (Schrag et al., 1992). Lipoprotein lipases also have the acid on strand six, but have lost that on strand seven, thus HPL, as well as other mammalian pancreatic lipases, represents an intermediate in the movement of the acid from one strand to the other (Schrag et al., 1992). Interestingly, a double-mutant fungal lipase in which the acidic residue is shifted from strand seven to six is catalytically active (Schrag et al., 1994).

The tyrosine residue essential to the activity of a large number of eukaryotic glutathione transferases is conserved in the Escherichia coli enzyme yet it is not necessary for catalysis. Instead, cysteine and histidine residues in the active-site are more suitably poised to play the equivalent functional role (Nishida et al., 1998). The tyrosine is lacking in the theta class of enzymes, and this group has a third mechanism, utilising a serine residue located at yet another structurally distinct position in the active-site pocket.

As noted recently, variations of this nature exist also in a family of homeodomain-like recombinases not analysed in this study (Grishin, 2000). Proteins of this family use a tyrosine nucleophile to form a DNA-enzyme intermediate, but this residue is located on a different secondary structural element in each of four members, including three enzymes with the same function, namely topoisomerases.

This variability is unexpected. It may reflect evolutionary divergence and optimisation of the catalytic efficiency of these enzymes. Similar reactions which involve different substrates are likely to vary in the precise details of their reaction chemistry, such as in their transition-state stabilisation requirements. As new functional groups fortuitously evolve in the active-site they may take over the roles of “old” residues if they are better suited to the job. That homologous active sites can offer these alternative positions is promising for protein design.

Secondly, genetic rearrangements can account for the apparent migration of catalytic residues, as observed in the aldolase superfamily. Fructose-bisphosphate aldolase class I and its mechanistically similar homologues, transaldolase, N-acetylneuraminate lyase, dihydrodipicolinate synthase, type I dehydroquinate dehydratase and δ-aminolevulinic acid dehydratase have an active-site lysine which forms a Schiff base intermediate with the substrate. Typically, this residue is located on strand six of the TIM barrel fold. In transaldolase, however, it is located on strand four and optimal sequence and structural alignment with aldolase class I requires a circular permutation of the beta-strands in the barrel (Jia et al., 1996). The strong similarity of these two enzymes in their substrate specificities and reaction chemistry, combined with the conservation of active-site residues, supports the evolution of transaldolase from an ancestral class I aldolase; such a relationship would require movement of the first two beta-strands to the C terminus after other members of this superfamily diverged (Jia et al., 1996). Circular permutation events have been observed in other folds (Russell & Ponting, 1998; Rojas et al., 1999).

**Functional convergence following evolutionary divergence.** Alternatively, these enzyme functions may have evolved independently, and nature has arrived at two or more distinct solutions to the
same catalytic conundrum within homologous structures. Whilst the recurrence of the same enzyme activity within different structural scaffolds, typified by subtilisin and chymotrypsin (Wallace et al., 1996), is well-known (Galperin et al., 1998), the independent evolution of similar functions within the same homologous superfamily is not so easy to identify. In two examples discussed below, evidence points to independent evolution of the same function within each of the two superfamilies. Lactate dehydrogenase from Trichomonas vaginalis presents another example of this phenomenon (Wu et al., 1999).

The Zn peptidase superfamily comprises two distinct subfamilies, the Zn carboxypeptidase (ZnCP) and the Zn aminopeptidase (ZnAP) families. Sequence homology between the two groups is undetectable by PSI-BLAST, and only with structural data can one infer an evolutionary relationship (Makarova & Grishin, 1999). Enzymes of both families have a co-located Zn binding-site but they differ in other aspects; members of the ZnAP family bind a second Zn atom in the active-site, and whilst both ZnCP and ZnAP enzymes use a glutamate residue as a catalytic base for the activation of water, in ZnCP it is in the C-terminal region and in ZnAP it is in the middle of the polypeptide chain. Nevertheless, both families include classical proteases, as well as desuccinylases and deacylases which act on N-modified amino acid residues. These functional specificities appear to have evolved in parallel after the two families diverged from an ancestral non-specific peptide hydrolase (Makarova & Grishin, 1999).

An extreme case of active-site variation is observed in the FAD/NAD(P)(H)-dependent disulphide oxidoreductase superfamily (see Figure 9). All members contain a FAD-binding domain, and a NAD(P)(H)-binding region. Typical members, including glutathione reductase (EC 1.6.4.2) and mammalian thioredoxin reductases (EC 1.6.4.5), have a third, interface domain at the C terminus, and have a catalytic redox-active disulphide located in the FAD domain. In contrast, plant and bacterial low-molecular mass thioredoxin reductases lack the interface domain, the disulphide is located in the NAD(P)(H)-binding module (Kuriyan et al., 1991) and it is in a catalytically incompetent position on the re face of the isoalloxazine ring of FAD, so a large conformational change must take place prior to catalysis. Thus, the same function has probably evolved twice independently within a superfamily of enzymes (Kuriyan et al., 1991). The atypical low-molecular weight thioredoxin reductases may have evolved either through independent fusion of the same two FAD and NAD(P)(H)-binding modules, or by loss of the interface domain of an ancestral protein (Petsko, 1991). Interestingly, flavocytochrome c sulphide dehydrogenase (EC 1.8.2.-) has yet another active-site arrangement. Whilst the redox-active cysteine residues are separated by just several amino acid residues along the primary sequence in other members, in this enzyme they are located on two distinct domains and lie almost 200 residues apart (Chen et al., 1994). Thus, three members of this superfamily use the same sulphur redox chemistry in one of three distinct active sites.

The origin of the active-site variation observed in some superfamilies is unclear. Determination of more sequences and structures in a superfamily may allow the identification of evolutionary “stepping stones” between members, and these can give clues to the origin of any positional variation of active-site residues.

**Structural equivalence of catalytic atoms.** The position of specific atoms involved in catalysis may be conserved across a family, whilst the residues to which they belong lie at different points in the protein scaffold (Hasson et al., 1998b). This has previously been observed with the catalytic bases in the enolase superfamily, members of which adopt the TIM barrel fold (Hasson et al., 1998b), and in the thioredoxin superfamily regarding the atom in each member which interacts with the substrate cysteine (Martin, 1995). Flavocytochrome b$_2$ and dihydroorotate dehydrogenase present another case; superimposition of their structures reveals just a 2 A˚ separation between the atoms which hydrogen bond with the flavin N5 locus, although the residues which contain them differ in both identity and position in the protein fold.
Loss/gain of metal sites

The loss or gain of catalytic metal-binding sites is observed in eight superfamilies. There are two situations. The absence of a catalytic metal-binding site in a protein may coincide with a lack of enzyme activity. Alternatively, metal content may differ between catalytically active proteins, implying a variation in the catalytic mechanism, a phenomenon identified in seven superfamilies.

For example, methionine aminopeptidase and aminopeptidase P both have a binuclear metal centre in their active sites, but this is absent in the more distant relative, creatinase. Their high structural similarity, combined with their shared ability to catalyse the hydrolytic cleavage of a C-N bond (Murzin, 1993), albeit on very different substrates and by different mechanistic strategies, and the conserved interaction of an invariant histidine with the nitrogen atom of the substrate scissile bond (Lowther et al., 1999), provide evidence for a distant evolutionary relationship between these enzymes. Similarly, quinone oxido reductase lacks both the catalytic and structural Zn sites of its homologues, the functionally analogous dehydrogenases (Thorn et al., 1995).

The aldolase superfamily contains both class I and class II fructose-bisphosphate aldolases (EC 4.1.2.13). Whilst the class I aldolase forms a Schiff base intermediate with the substrate via an active-site lysine residue, as discussed above, in the class II reactions, a covalent enzyme-substrate adduct is not formed and instead catalysis involves a divalent metal cofactor, normally Zn. The nucleophile adding to the acceptor substrate is a Schiff base enamine for aldolase class I, but an enolate anion for aldolase class II. The evolutionary origin of these two classes was unclear until the recent sequence analysis of an aldolase from E. coli. This forms a Schiff base, yet shares higher sequence similarity with the metal-dependent enzymes (Galperin et al., 2000). Thus, two entirely different mechanisms have evolved within the same superfamily to catalyse the same reaction. This superfamily exhibits further variability in metal content. δ-Aminolevulinic acid dehydratase forms a Schiff base and requires Zn for activity, but its metal site is in a position different from that of class II aldolase (Erskine et al., 1999). The metal sites of aldolase class II and the class II-like Phenylalanine ammonia-lyase 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase (EC 4.1.2.15), on the other hand, are co-located. 3-deoxy-D-manno-octulosonate-8-phosphate synthase (EC 4.1.2.16) from E. coli catalyses a reaction almost identical to that of 3-deoxy-D-arabino-heptulosonate-7-phosphate synthase, but its activity does not depend on a metal ion nor a Schiff base-forming lysine.

The cupredoxins present the best case of variation in metal content (Ryden & Hunt, 1993; Murphy et al., 1997). Four types of Cu sites exist in this superfamily, where all sites are involved in electron transfer, and the number of Cu ligands varies from zero to eight with any one domain providing ligands for up to four Cu atoms. Interestingly, the level of structural similarity does not correlate with Cu content and thus electron transfer function (Murphy et al., 1997), implying the acquisition and loss of Cu sites through evolution.

Enzyme and non-enzyme functions: correlation with catalytic residues

For the majority of non-enzyme homologues belonging to the structural superfamilies studied in this analysis, the lack of enzyme function is correlated with a lack of the catalytic residues identified in close relatives. However, apparently “missing” functional groups do not always correlate with a non-enzymatic function, as indicated in the previous two sections.

Conversely, the presence of functional residues preserved in enzyme homologues does not necessarily imply that a protein is catalytically active, and then only structural details of a non-enzyme can provide clues to its inactivity. Narbonin provides a good example. As far as is known, this protein is catalytically inactive despite the presence of a glutamate, essential to the activity of its close chitinase homologues. A salt-bridge as well as other structural variations which render the active-site cleft inaccessible to chitin are in part responsible for the lack of catalytic activity (Terwisscha van Scheltinga et al., 1995).

Changes at the domain and quaternary levels

Domain enlargement

Whilst most protein folds vary in size between homologues to a certain extent, the sizes of 12 domains (belonging to 11 superfamilies) vary by at least twofold. Size variations may involve the addition/loss of subdomains, variability in loop length, and/or changes to the structural core, such as beta-sheet extension.

An extreme example is provided by the phosphoenolpyruvate-binding domains which have a TIM barrel structure ranging from 274 to 883 residues in size. Phosphoenolpyruvate carboxylase is the unusual member in this superfamily, having a total of 40 alpha-helices, including the eight which pack round the beta-barrel. Most of the additional helices, introduced at different points throughout the primary sequence, are situated at the C-terminal end of the barrel, and in this cluster of helices lies the binding site for aspartate, an allosteric inhibitor (Kai et al., 1999). Extensive domain embellishment has provided the enzyme with a regulatory site, and probably also an interface for stabilising its tetrameric structure.

Sizes vary by 3.5-fold in the non-heme di-iron carboxylase protein superfamily. Members have an alpha bundle structure, but in non-enzymes this comprises four helices, and in enzymes, at least eight. Of the α/β hydrolases, cutinase (EC 3.1.1.-)
is the smallest enzyme (197 residues), with five strands in the main beta-sheet (Longhi et al., 1997), in contrast to bovine bile-salt activated cholesterol esterase (EC 3.1.1.13) (547 residues) which has 11 strands, and loop structures up to 79 residues in length (Chen et al., 1998). A structurally conserved core and catalytic triad confirms a common ancestry for these enzymes.

Phosphoglycerate mutase and its homologues have an alpha-helical subdomain formed by two loops of their six-stranded three-layer sandwich structures. Loop lengths in 3-phytase are more than double those in fructose-2,6-biphosphatase, so that the subdomains differ vastly in size and architecture, and this confers different substrate specificities on these enzymes.

Other examples of size variation include the smaller alpha domain of the helix-hairpin-helix base-excision DNA repair enzymes, with two proteins having an extra [4Fe-4S] cluster-binding loop (Thayer et al., 1995; Guan et al., 1998), and the beta-sheet extension of the C-terminal ATP-grasp domain of biotin carboxylase to mediate dimer contacts (Waldrop et al., 1994).

Nature has probably embellished more simple ancestral folds in the evolution of protein complexity and functional specialisation, through amino acid and intron insertions. However, the reverse scenario does occur. For example, the loss of two α-helices in endo-β-N-acetylglucosaminidase of the TIM barrel glycosyl hydrolases (see above), and the interface domain of the more typical homodimeric members of the FAD/NAD(P)(H)-dependent disulphide oxidoreductase superfamily has probably been truncated in the evolution of flavocytochrome c sulphide dehydrogenase to accommodate the cytochrome c subunit (Van Driessche et al., 1996).

**Domain recruitment**

In 27 of the 31 superfamilies, the domain organisation varies between members. Additional modules fused to the catalytic domain of an enzyme may play a role in regulation, oligomerisation, cofactor dependency, subcellular targeting or, commonly, substrate specificity.

An accessory domain may modulate the substrate selectivity of a protein by providing a specific binding site, or, by playing a purely structural role, may shape the active site for the recognition of a substrate of a different shape or size. For example, prokaryotic methionine aminopeptidase is a monomeric single-domain protein which cleaves large polypeptides. In contrast, creatinase is a two-domain protein and exists as a homodimer; the additional domain of the second subunit caps the active site allowing the binding of the small molecule creatine (Hoefken et al., 1988) (see Figure 10).

In the family of glycosyl hydrolases adopting the TIM barrel fold, both methods of modifying substrate specificity are observed, and interestingly, the number of additional domains does not necessarily correlate with the nature of the carbohydrate substrate (see Figure 11). Endo-1,4-β-glucanase C and cyclodextrin glycosyltransferase have cleft-like active sites, and bind the polysaccharides cellulose and starch, respectively. The former is a single domain protein, whilst the latter comprises four domains, the C-terminal domain providing a starch-binding site. In contrast, β-glucosidase and the five-domain protein β-galactosidase have pocket-like active sites, and consequently bind monosaccharides and disaccharides, respectively. The extra domains of β-galactosidase (including two fibronectin III-like domains) do not play the typical functional roles they exhibit in other con-
texts, but appear to play a purely structural role, providing loops which give the active site its pocket-like shape for the binding of small sugar molecules (Juers et al., 1999). Cyanogenic β-glucosidase achieves its selectivity for monosaccharides through the specific structural properties of its single, catalytic domain.

Gene fusion does not necessarily modify the biochemical function of a protein and domain combination can just reflect organisation of the genome. In three families, modular variation results only from the fusion of a second enzyme onto the same polypeptide chain of one member. For example, in cytochrome P450BM-3 the redox partner, NADPH-cytochrome P450 reductase, is fused to the P450 heme domain (Narhi & Fulco, 1987). Other members of the superfamily exist as single-domain proteins and the reductase module is located on a separate subunit.

In two of the four superfamilies which show no variation in domain organisation, members are multi-domain, implying that gene fusion has probably played a role in the evolution of these proteins. Indeed, in 11 superfamilies, members share a minimum of two domains in common and the individual domains of these superfamilies may be found in isolation or in combination with other modules as more protein structures are determined. For example, one of the two domains in the ATP-dependent carboxylate-amine/thiol ligase superfamily has recently been observed in phosphoribosylaminomimidazolesuccinocarboxamide (SAICAR) synthase (Levdikov et al., 1998).

In the vast majority of the 27 superfamilies which exhibit variations in modular construction, domain recruitment, in combination with point mutations which affect ligand-binding or catalysis, account for the observed functional diversity. With reference to Figure 1, having considered both single and multi-domain proteins in this data analysis, however, there may be some apparently functionally diverse superfamilies in which domain members have similar, if not identical, biochemical roles, and the observed functional variation is attributed to differences in modular and/or subunit assembly. Such superfamilies illustrate the use of the same functional domain in quite diverse contexts, and highlight the importance of considering modular construction in gene annotation.

One example is provided by the “Rieske”-like iron-sulphur domains which bind a [2Fe-2S] cluster via four conserved ligands, and these domains mediate electron-transfer between donor and acceptor molecules. These domains have been identified in the oligomeric cytochrome bc complexes which play a central role in the electron transfer chains of mitochondria, chloroplasts and bacteria, within the oxygenase subunits of a large family of multi-component aromatic ring-hydroxylating dioxygenases, as well as in other oxidoreductases such as nitrate reductase. The iron-sulphur domains of these proteins are embedded in enzyme complexes which differ vastly in domain organisation and subunit assembly, accounting for the functional diversity of this superfamily.

### Domain duplication with functional specialisation

In 11 superfamilies, domain duplication has occurred, such that the same module is repeated two or more times along the same polypeptide chain. In seven, functional specialisation has followed the duplication event, and distinct roles may be ascribed to each duplicated module (see Table 3).

One of the best examples is provided by the ferroxidase ceruplasmin. This enzyme has six domains all of which adopt the cupredoxin fold. The second, fourth and sixth domains bind a type I Cu, a trinuclear Cu site is located at the interface between domains one and six, and domains three and five lack Cu altogether.

### Domain rearrangement

In four superfamilies, the relative location of two or more domains along the polypeptide chain varies between members.

The thiamin pyrophosphate (TPP)-dependent enzymes use the cofactor to catalyse cleavage of carbon-carbon bonds adjacent to carbonyl groups. All enzyme complexes of this superfamily share two domains in common, which together bind TPP, and these two domains are homologous to each other (see Figure 12). They have evolved specific functions, one binding the pyrimidine ring
of TPP (pyr), and the other binding the diphosphate (pp). However, their order in the primary sequence varies (Muller et al., 1993), and in one member, 2-oxoisovalerate dehydrogenase, they are located on separate subunits (Aëvarsson et al., 1999). Even when these domains occur on the same chain, the TPP-binding unit within an enzyme complex comprises two subunits, with the TPP sandwiched between the pyr and pp domains.

Enzyme complexes are made up of separate chains (Muller et al., 1993). The TPP-binding unit is highly conserved (Muller et al., 1993), so independent fusion of the pyr and pp domains onto the same chain may have occurred after the unit evolved.

Eukaryotic and prokaryotic glutathione synthetases (GS) of the ATP-dependent carboxylate-amine/thiol ligase superfamily present another example of domain rearrangement. They share three domains in common, two of which together form the ATP-grasp fold (Murzin, 1996) which defines this superfamily. In human GS, a structural outlier of this superfamily, the large C-terminal domain of the ATP-grasp fold is split in two with one half attached to the N terminus owing to a circular permutation event (Polekhina et al., 1999); thus two segments of sequence located at its N and C termini form a structural domain which is equivalent to the third contiguous domain of prokaryotic GS. An evolutionary relationship between the two types of GS had gone undetected until the structure of human GS was determined.

**Subunit assembly**

In at least 23 superfamilies, subunit assembly varies between members. For 12 of these, variation results only from a differing number of identical chains in the protein complex (see Table 3). Usually, a change in the oligomerisation state of this nature appears to confer no functional difference, unless an additional subunit contributes residues to the active site and modifies it in some way. An example of this type of modification is provided by methionine aminopeptidase and creatinase discussed above.

In 11 superfamilies, one or more members functions in combination with different subunits, that is, as a hetero-oligomer. Additional subunits may be important for substrate-binding or electron transfer, for example, or they may even play a direct role in catalysis. The non-homologous proteins tryptophan synthase alpha-chain and the large subunit of carbamoyl phosphate synthase are responsible for catalysing just one step in their respective reactions, with their partnering subunits necessary for complete catalysis. Tunnels to connect the distal active sites in the protein complexes allow the diffusion of intermediates and prevent their contact with solvent (Hyde et al., 1988; Thoden et al., 1997).

The “moonlighting” thioredoxin-like protein disulphide isomerase presents an example where its function depends upon its oligomerisation state; it is the beta subunit of prolyl-4-hydroxylase and is also one subunit of the triglyceride transfer protein complex (Pihlajaniemi et al., 1987; Wetterau et al., 1990). This protein belongs to one of several superfamilies which are particularly variable in terms of domain content and subunit assembly. Protein disulphide isomerase comprises four thioredoxin-like modules, only two of which are catalytic; the homologue thioredoxin functions as a single-domain monomer; the two-domain non-enzyme phosducin forms a complex with the beta and gamma components of the GTP-binding protein transducin; and high-capacity Ca$^{2+}$ binding by calsequestrin, which comprises three thioredoxin-like motifs (Wang et al., 1998), requires aggregation into a polymeric state.

**Discussion**

**Caveats**

Our work depends upon both the reliability of annotation in SWISS-PROT, and the accuracy of crystal structure and mutagenesis studies from which catalytic residues and reaction mechanisms are inferred. In addition, the data presented are based upon proteins for which structures have

---

**Figure 12.** Domain organisation and TPP-binding units of TPP-dependent enzymes. Each shape corresponds to a structural domain, and domains of the same shape are homologous. The light and dark grey squares represent the TPP-binding domains involved in interacting with the pyrimidine and diphosphate moiety, respectively, and highlight the domain rearrangement observed within this superfamily. Dotted lines represent domain-domain interfaces involved in TPP-binding. (a) Pyruvate oxidase, pyruvate decarboxylase and benzoylformate decarboxylase; TPP-binding unit is a homodimer; (b) transketolase; TPP-binding unit is a homodimer; (c) pyruvate ferredoxin reductase; TPP-binding unit is a monomer; (d) 2-oxoisovalerate dehydrogenase; TPP-binding unit is a heterodimer of alpha and beta-chains.
been determined, and their sequence relatives from GenBank identified by PSI-BLAST. Both databases inherently contain biases which affect analyses of this nature, the PDB towards smaller proteins and those which are easy to crystallise, and GenBank towards microbial sequences. Structural genomics initiatives and the sequencing of more genomes will minimise these biases in the future.

The EC classification system has some well-known limitations. Firstly, the reaction direction is chosen arbitrarily, such that phosphoenolpyruvate carboxylase (EC 4.1.1.31), for example, is classified as a lyase yet it catalyses irreversible carbon-carbon bond formation. Secondly, EC numbers describe the overall reaction catalysed, and provide no details of the chemistry involved, thus two enzymes which share a common chemistry, and even mechanistic strategy may have completely different EC classifications. For example, pyruvate oxidase and pyruvate decarboxylase both catalyse TPP-dependent decarboxylation of pyruvate, but in the former, it is oxidative, so the two enzymes belong to different primary EC classes, their EC numbers being 1.2.3.3 and 4.1.1.1, respectively. In addition, the overall transformation of substrate into product may involve a number of individual reaction steps. Dehydroquinate synthase is correctly classified as a phosphorus-oxygen lyase (EC 4.6.1.3), but it is actively involved in the catalysis of five reaction steps, alcohol oxidation, phosphate beta-elimination, carbonyl reduction, ring-opening and intramolecular aldol condensation, all in one active-site (Carpenter et al., 1998).

The functional variation observed in many, if not all, of the 31 superfamilies studied may be underestimated. In the future, new sequences and structures may considerably expand their functional repertoire. We may find that any similarity in function, such as reaction chemistry or substrate-binding, reported here is limited to a subgroup of related proteins, and a more subtle functional property or structural feature may instead define the superfamily as a whole. Addition of Clp protease to the crotonase-like superfamily illustrates this point. Formerly a superfamily in which all members catalysed one of a variety of reactions at the acyl group of a CoA thioester substrate, it is now apparent that similarity in function is limited to the ability to stabilise the negative charge on an oxygen atom in an oxynion intermediate.

Overview of functional diversification

With reference to 31 homologous structural superfamilies, we have sought to gather the data and thereby understand how specialised functions have evolved, through changes at the atomic level to gross structural rearrangements. We have presented the structural and functional attributes which are conserved within a superfamily and those that differ, and what bearing, if any, these similarities and changes have on protein function.

This work has implications for interpreting the plethora of data generated in the genome sequencing and planned structural genomics projects from which we must unravel gene functions.

These 31 structural superfamilies, in combination with other modules, support almost 200 protein functions. With the inclusion of sequence data this number more than doubles. Accommodating diverse functions on a common scaffold has involved many types of variations within one superfamily. For example, the di-iron centre of the non-heme di-iron carboxylate enzymes has evolved through motif duplication, and this superfamily shows also loss of a catalytic site, extensive domain enlargement, and considerable variations in domain organisation and subunit assembly. With these changes has emerged a superfamily involved in DNA-synthesis, iron-storage, fatty acid biosynthesis, methane fixation and the protection of DNA from oxidative damage. Other superfamilies of this nature include the cupredoxins, the thioredoxin-like proteins, the TPP-dependent enzymes and the ATP-dependent carboxylate-amine/thiol liga superfamily. The number of superfamilies with changes in modular construction and subunit assembly illustrates the importance of such rearrangements in creating new functions, confirmed by the existence of numerous active sites at domain and subunit interfaces.

On the other hand, some superfamilies appear to have achieved their functional diversity through incremental mutations in the active site alone. The crotonase-like enzymes, and the TIM barrel enolase and aldolase superfamilies catalyse diverse reactions on a wide variety of substrates, nevertheless domain organisation is constant, and changes in oligomerisation state result only from a variable number of identical chains in the enzyme complexes.

Substrate specificity and reaction chemistry

Substrate specificity and catalytic activity define the absolute biochemical function of an enzyme. During evolution, one property may vary whilst the other is maintained, and the role they play in enzyme recruitment has been a subject of discussion.

Jensen (1976) proposed that enzyme recruitment exploited the substrate ambiguity of ancestral proteins. Broad substrate specificity provided the ancestral cell with a “biochemical leakiness” (Jensen, 1976), and thus biological diversity came from a limited number of genes; enzymes with improved substrate selectivity and catalytic efficiency have evolved through gene duplication and specialisation of active-site architectures. More recently, these ideas have been extended by O’Brien & Herschlag (1999). The results of our analysis certainly highlight the evolution of multiple specialised functions within an enzyme superfamily in support of this hypothesis. Without phylogenetic studies, however, it is impossible to
say definitively whether enzyme substrate specificities have narrowed during the course of evolution. Whilst this seems to be the more common scenario as more complex, multi-cellular systems have evolved, many enzymes do exhibit broad substrate selectivities (Jensen, 1976), and this property could also evolve through adaptation of a specific ancestral active site for the recognition of multiple substrates.

In more recent work (Petsko et al., 1993; Babbitt & Gerlt, 1997; Gerlt & Babbitt, 1998), it has been proposed that chemistry, as opposed to substrate specificity, has dictated the choice of ancestral proteins for the evolution of new enzyme activities. Indeed, the majority of superfamilies studied here display little conservation of substrate specificity, whereas conservation of reaction chemistry is far more common. In this dataset, the ribulose-phosphate-binding barrels present the only example in which substrate-binding is well conserved yet the chemistry is varied, implying that the former has dictated the course of evolution of this superfamily. At the other end of the spectrum, five superfamilies, such as the metal-dependent hydrolases, utilise a common reaction chemistry and catalytic framework to act on an array of substrates which do not have even a small moiety in common.

Whilst conservation of chemistry does appear to be the dominant theme, there may be isolated cases within a few superfamilies where substrate specificity has been the important factor in the evolution of particular members. For example, of the four structural members of the crotonase-like superfamily all but Clp protease binds a coenzyme A thioester (Babbitt & Gerlt, 1997). Indeed, it can be difficult to treat chemistry and substrate specificity as two distinct properties, and distinguish which of the two drives evolution, since a common chemistry often implies a common substrate moiety, which may undergo change during the reaction. For example, members of the enolase and type I PLP-dependent aspartate aminotransferase superfamilies bind a carboxylate and amino acid group, respectively. Similarly, chemistry and the nature of the cofactor are inherently linked; duplication of cofactor-binding domains has been extensive throughout evolution, exemplified by the ubiquitous nucleotide-binding Rossmann fold.

There are several examples of pairs of homologous enzymes, including the tryptophan biosynthesis enzymes of the ribulose-phosphate-binding barrel superfamily, which catalyse adjacent steps within a metabolic pathway, implying that substrate specificity is the critical influence in enzyme recruitment. However, a recent analysis of evolutionary relationships in small molecule metabolic pathways in Escherichia coli showed that conservation of the main substrate-binding site is far more rare than the conservation of chemistry in support of the “chemistry first strategy” (S. A. Teichmann & S. C. G. R. Rison, personal communication).

**Catalytic residues**

Over one-third of superfamilies exhibit catalytic residue migration, despite the limited data. Indeed, with the growth in the sequence databases, we may find that this variability is no longer an exception, and that many active sites are “malleable”. Alternatively, it may be limited to a few protein folds which are particularly amenable to the topological shift of active-site residues, in that they can offer a number of points in their scaffolds from which catalytic groups can be recruited. A greater degree of flexibility of an active site clearly facilitates functional diversification. This may account, at least in part, for the functional versatility of the TIM barrel.

**Conclusions**

Typically, functional analyses are done for a single family in isolation, but together, they provide a better insight into protein evolution by allowing the identification of preferred mechanisms of functional diversification. In this work, data have been laboriously extracted from the literature and combined with our own analyses, to provide an overview of the extent and the mechanisms by which proteins evolve new functions. Functional variation occurs mostly in more distantly related proteins (<40%) and the structural data have been essential for understanding the molecular basis of observed functional differences.

It is important to note that we can only underestimate the functional variation observed. Usually, in the “midnight zone” of sequence and structural similarity, an evolutionary relationship between proteins is inferred only if they share some similarity in function, such as a conserved substrate-binding site, active-site architecture or protein-protein interface. New genes and structures can provide “bridges” between superfamilies which previously showed little evidence of homology, and we are likely to observe even more extensive and unexpected variations in function within many superfamilies in the future.

With the onset of structural genomics projects, the interpretation of protein function using structural data will become increasingly important. The observations discussed here hint that functional characterisation by structure determination will not be straightforward. Certainly in a number of superfamilies, the prediction of enzyme activity of an uncharacterised protein on the basis of the presence or absence of catalytic residues identified in its relatives may be difficult. We have seen that the same catalytic framework may be used for a variety of reactions, and little can be inferred other than some aspects of the reaction mechanism. On the other hand, the absence of one or more catalytic residues does not necessarily imply a lack of enzyme activity. Indeed, various attempts at structure-based functional assignment have been met with mixed success (Ren et al., 1998; Yang et al.,...
Several important questions remain. Given a family of protein structures, to what extent can we predict biochemical function and ligand-binding? Is it possible to extract rules for the design of novel functions and can we accurately annotate genome sequence data? The data we have collected and analysed in this work will provide the experimental basis to explore these questions.

Materials and Methods

To perform the analysis of enzyme structure and function, information was extracted from the CATH structural classification scheme (Orengo et al., 1997), PDBsum (Laskowski et al., 1997), SWISS-PROT (Bairoch & Apweiler, 2000) and ENZYME (Bairoch, 2000) databases, and the literature. Due to the scope of this analysis and space restrictions, it has not been possible to reference many of the articles which provide the data referred to in this paper. References are accessible through SWISS-PROT, PDB and the literature. Due to the inherent problem of identifying the relationship sometimes requires manual intervention, reference to the literature, and PDBsum, SWISS-PROT and ENZYME databases; those sharing some similarity in function, a co-located functional site, or an unusual structural feature are classified into the same superfamily in CATH. For this analysis, we used non-identical domains, filtered at 95% sequence identity, to extract EC numbers. This was necessary due to limited annotation in the PDB. PDB to SWISS-PROT links provided in a manually curated list at the European Bioinformatics Institute were supplemented by way of a two-way sequence scan using the FASTA sequence comparison algorithm (Pearson & Lipman, 1988). EC numbers were identified by way of the SWISS-PROT accession code from the ENZYME and SWISS-PROT databases.

CATH is a classification of protein structural domains. Proteins are classified into homologous superfamilies by a semi-automatic procedure. Those sharing high sequence and/or structural similarity are merged automatically into the same superfamily. For more distantly related proteins in the twilight zone of sequence and structural similarity, confirmation of an evolutionary relationship sometimes requires manual intervention, with reference to the literature, and PDBsum, SWISS-PROT and ENZYME databases; those sharing some similarity in function, a co-located functional site, or an unusual structural feature are classified into the same superfamily in CATH. For this analysis, we used non-identical domains, filtered at 95% sequence identity, from the CATH release dated 5 April 2000.

For the identification of sequence relatives, the amino acid sequence of each structural domain was scanned against a non-redundant subset of GenBank from the NCBI using the profile-based PSI-BLAST program (Alschul et al., 1997). Each PSI-BLAST scan was terminated after 20 iterations, or on convergence. Matched GenBank fragments having a final e-value of 0.0005 or smaller and an overlap of 80% with the query sequence were added to the CATH homologous superfamily, with consensus domain boundaries evaluated using in-house software. Due to the inherent problem of identifying the sequence relatives of discontinuous domains by this method, only those homologous superfamilies in which all domain members are contiguous were considered for analysis. This corresponds to 732 out of a total number of 903 superfamilies in CATH, and 2358 non-identical CATH95 representatives. With the inclusion of sequence data, identical or near-identical domains were again filtered out at 95% sequence identity, providing a set of 65,303 non-identical PF95 (Protein Family) representatives. EC numbers were extracted from ENZYME and SWISS-PROT for those GenBank entries contained within SWISS-PROT, else they were extracted from GenBank itself.

For the identification of non-enzymes in SWISS-PROT, any entry matching one or more of the following criteria was ignored (a) contains the word “hypothesical” in the description (DE) or keyword (KW) lines (b) has “FUNCTION: NOT KNOWN” in the comment (CC) lines, (c) has no keywords nor “FUNCTION” description in the CC lines. Of the remaining entries, those which (a) do not have an EC number assignment and (b) do not contain a word ending in “ase” (”disease” and “permease” excluded) in the DE or KW lines (with the exception of proteins described as inhibitors) were identified as non-enzymes.

In the identification of single-domain proteins, those PF95 domains without a SWISS-PROT link were discarded, and any marked as a fragment in SWISS-PROT were discarded also. In the remaining domains, those structural protein (classified as single-domain in CATH) and sequence relatives which left 100 residues or fewer uncovered upon alignment with their SWISS-PROT sequence, were taken as single-domain. A total of 7299 out of 65,303 PF95 representatives were identified as single domain.

To assess the correlation between functional similarity (as defined by EC number) and sequence identity, homologous superfamilies containing two or more enzymes were considered. Some enzymes have two or more EC numbers assigned to them. These may be “multienzymes”, with the catalytic functions contributed by distinct domains and/or separate subunits, or, more rarely, they may be “single enzymes”, catalysing different reactions using the same catalytic site. For ease of comparison of EC numbers, enzymes having two or more EC numbers assigned to them, or those with incomplete EC numbers (e.g. EC 4.2.1.-) were ignored, corresponding to 16% of the 11,961 PF95 representatives having EC assignments. A pairwise sequence alignment of each unique PF95 enzyme/enzyme or enzyme/non-enzyme pair within a superfamily was carried out using the method of Needleman & Wunsch (1971) to evaluate sequence identities. Non-enzyme/non-enzyme pairs were ignored. There were 486,084 homologous pairs in total contained in 369 superfamilies, and 81,312 single-domain pairs contained in 127 superfamilies.

For the detailed superfamily analysis, 31 superfamilies were considered. This involved extensive reading of the literature to extract information regarding structural details, active-site residues and the catalytic mechanism for each member. Pairwise and multiple structural alignments were generated using SSAP (Taylor & Orengo, 1989) and CORA (Orengo, 1999), respectively, to determine structurally equivalent regions and to assess conservation of catalytic residues. Superimposition of structures was done using in-house software. Recent RCSB entries (deposited January 2000 or earlier) identified in the literature or by PSI-BLAST as members of any one of these superfamilies, but not yet classified in CATH, were added to the dataset for analysis. More detailed structural and functional information on these superfamilies may be found at www.biochem.ucl.ac.uk/
bsm/FAM-EC/, including pairwise sequence identities and SSAP structural alignment scores.

Acknowledgements

A.E.T. acknowledges support from the BBSRC and Oxford Molecular Ltd, and C.A.O. for the support of the MRC. The software used to identify the PDB to SWISS-PROT links was written by Alex Michie and Andrew Martin at University College London. We thank David Lee and Frances Pearl for their work in identifying sequence relatives for all domains in CATH. This is a publication from the Bloomsbury Centre for Structural Biology, funded by the BBSRC.

References


Needleman, S. B. & Wunsch, C. D. (1971). A general method applicable to the search for similarities in...
the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-453.


*Edited by A. R. Fersht*

(Received 4 September 2000; received in revised form 2 February 2001; accepted 2 February 2001)