doi:10.1006/jmbi.2001.4513 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 307, 1113–1143

JMB

Evolution of Function in Protein Superfamilies, from a Structural Perspective

Annabel E. Todd[1], Christine A. Orengo[1] and Janet M. Thornton[1,2*]

[1]*Biochemistry and Molecular Biology Department, University College London, Gower Street London, WC1E 6BT, UK*

[2]*Crystallography Department Birkbeck College Malet Street London, WC1E 7HX, UK*

**Corresponding author*

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 multi-domain 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 supefamilies 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/.

© 2001 Academic Press

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.

E-mail address of the corresponding author: thornton@biochem.ucl.ac.uk

0022-2836/01/041113–31 $35.00/0

© 2001 Academic Press

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, SWISS-PROT 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

Table 1. Description of the different levels in the EC classification

First figure	Second figure	Third figure
A. *OXIDOREDUCTASES* Substrate is oxidised-regarded as the hydrogen or electron donor	Describes substrate acted on by enzyme	Type of acceptor
B. *TRANSFERASES* Transfer of a group from one substrate to another	Describes group transferred	Further information on the group transferred
C. *HYDROLASES* Hydrolytic cleavage of bond	Describes type of bond	Nature of substrate
D. *LYASES* Cleavage of bonds by elimination	Type of bond	Further information on the group eliminated
E. *ISOMERASES*	Type of reorganisation	Type of substrate
F. *LIGASES* Enzyme catalysing the joining of two molecules in concert with hydrolysis of ATP	Describes type of bond formed	Describes type of compound formed

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).

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. However, in as many as 22 superfamilies, the EC number is not conserved to any level, and members have different enzyme activities. A total of 13 of these superfamilies contain members which do not function as enzymes at all.

With inclusion of the PSI-BLAST (Altschul *et al.*, 1997) sequence relatives of the structural members of these superfamilies, the fraction of superfamilies showing variation in enzyme function increases to almost 70 %. A total of 43 families display absolutely no conservation in EC number, and the number of superfamilies containing both enzymes and

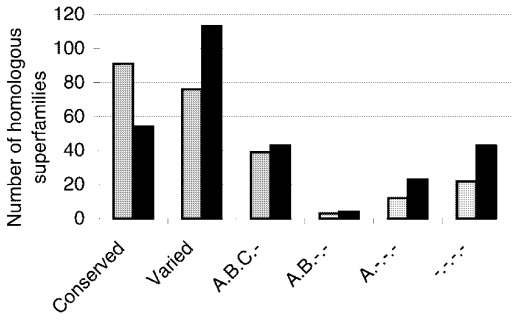

Figure 1. Conservation of enzyme function as defined by EC number (*x*-axis). Shown in grey is the conservation of EC number between CATH95 representatives in 167 CATH homologous superfamilies containing two or more enzymes. In black is the conservation of EC number between all PF95 representatives within these superfamilies (i.e. sequence relatives included also). Enzymes having multiple or incomplete EC numbers were ignored.

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) (Artymiuk *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

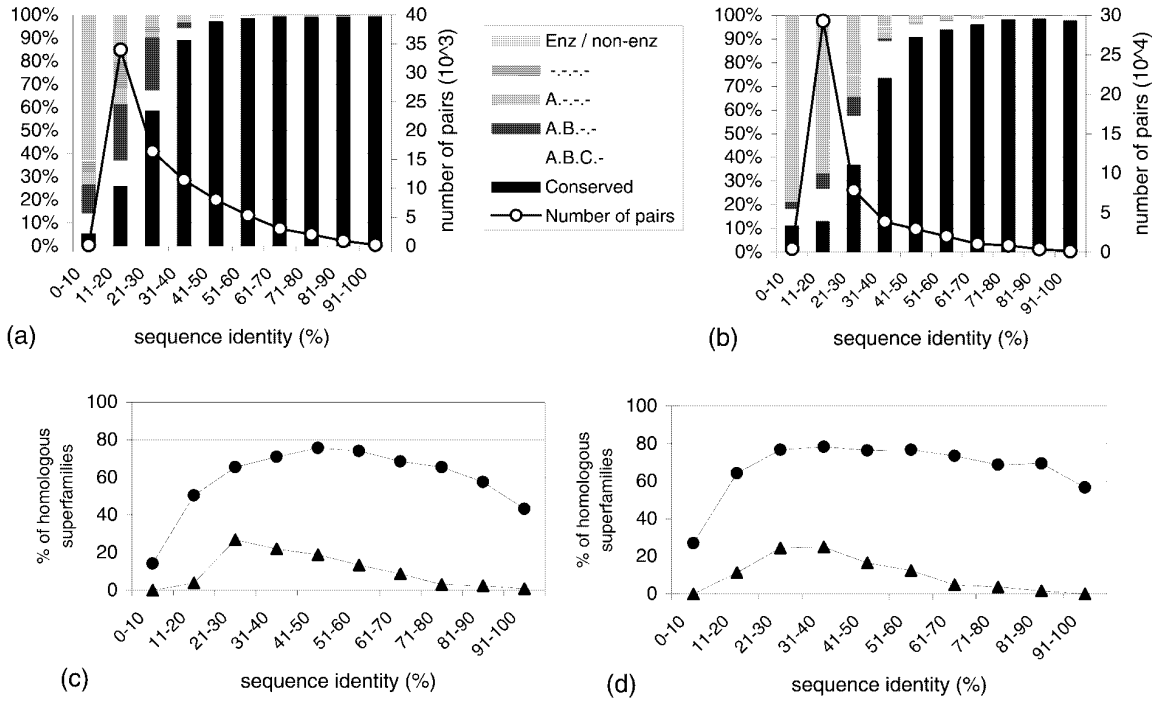

Figure 2. (a) and (b) Conservation of EC number versus sequence identity for homologous enzyme/enzyme and enzyme/non-enzyme pairs. Non-enzyme/non-enzyme pairs are ignored. The analysis is limited to those superfamilies containing two or more enzymes. (a) Single-domain PF95 enzymes and non-enzymes only (contained within 127 homologous superfamilies); (b) single and multi-domain PF95 enzymes and non-enzymes (contained within 369 homologous superfamilies). Each level of functional similarity is represented as a fractional percentage of the total number of unique homologous pairs (indicated by line graph) in a given range of sequence identities. Functional variations occur largely below 40% sequence identity. (c) and (d) Sequence diversity of the homologous superfamilies. Circles and triangles indicate the percentage of homologous superfamilies containing one or more pairs and having a mean sequence identity, respectively, in a given sequence identity range. (c) Single-domain PF95 superfamilies (corresponds to pairs in (a)); (d) single and multi-domain PF95 superfamilies (corresponds to pairs in (b)). These plots show that the large number of homologous pairs sharing less than 40% sequence identity is not contributed by a few highly populated, divergent superfamilies, but instead most superfamilies exhibit extensive sequence diversity.

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

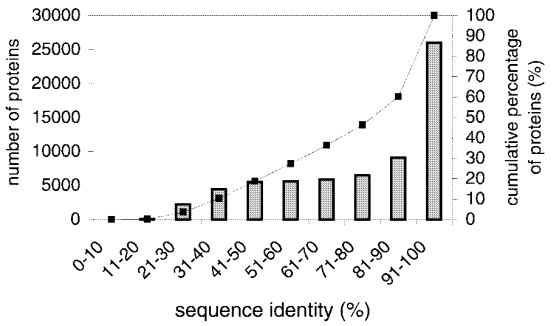

Figure 3. Distribution of the sequence identities of the closest relatives of all protein domains. The closest relative, in terms of sequence identity, was identified for all non-identical PF95 representatives with the exception of eight PDB domains which lack a homologue. Near-identical and identical domains identified in the non-redundant dataset of GenBank from NCBI were included as closest relatives but near-identical and identical PDB domains were ignored (limited to CATH95 representatives). The histogram illustrates the number of closest relatives which fall within each level of sequence identity and the line graph is a cumulative percentage of closest relatives with increasing sequence identity. A similar distribution is observed when single-domain proteins only are considered. The sequence identity of the ''best'' hit of only 10% of domains is less than 40%, the threshold below which functional variation is significant. Note that only those matched sequence fragments identified by PSI-BLAST using the CATH95 domains as probes are included, and the plot provides no indication of the number of ''singletons'' in the database, that is the fraction of proteins which lack a detectable homologue.

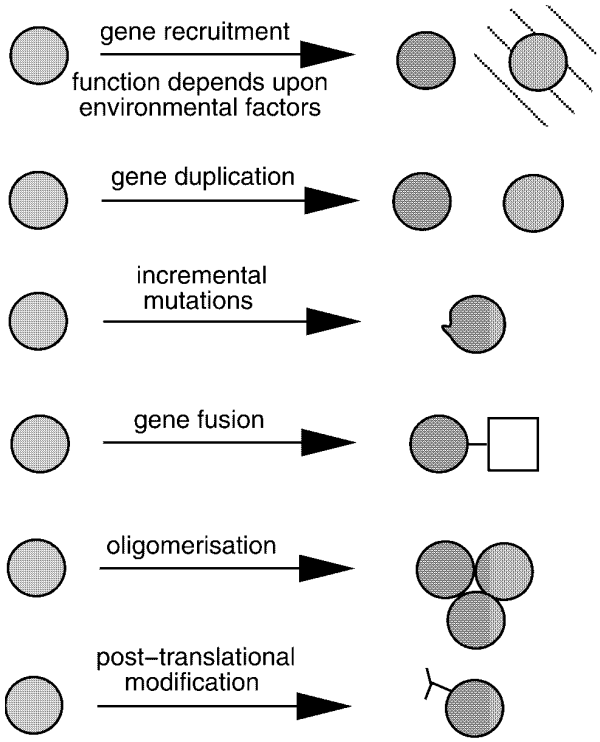

Figure 4. Schematic diagram to illustrate the mechanisms by which new functions are created. In practice, new functions often evolve via a combination of mechanisms.

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, FMN-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 three of five structural members catalyse sequential steps in tryptophan biosynthesis, and 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-chain 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

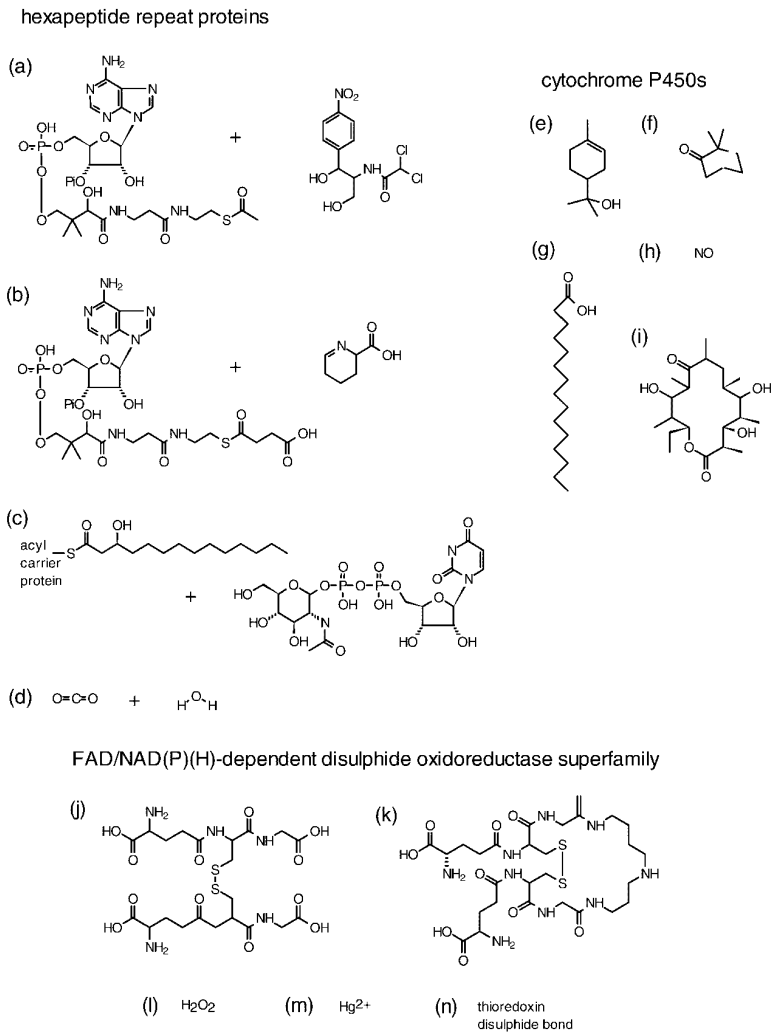

Figure 5. Substrates bound by enzymes belonging to three superfamilies which display diverse substrate specificities. Hexapeptide repeat proteins: (a) chloramphenicol acetyltransferase (EC 2.3.1.28); (b) tetrahydropicolinate *N*-succinyltransferase (2.3.1.117); (c) UDP-*N*-acetylglucosamine acyltransferase (2.3.1.129); (d) carbonic anhydrase (4.2.1.1); cytochrome P450s: (e) cytochrome *P*450-terp (1.14.-.-); (f) cytochrome *P*450-cam (1.14.15.1); (g) cytochrome *P*450-bm3 (1.14.14.1); (h) cytochrome *P*450-nor (nitric oxide reductase) (1.14.-.-); (i) cytochrome *P*450-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

Superfamily	Member	EC number(s)	PDB code	Pfam
Fumarase/aspartase	Fumarate hydratase class II (fumarase)	4.2.1.2	1fuq	PF00206
	Aspartate ammonia-lyase (aspartase)	4.3.1.1	1jsw	PF00206
	Adenylosuccinate lyase	4.3.2.2	1c3c	PF00206
	Argininosuccinate lyase	4.3.2.1	1aos	PF00206
	δ crystallin II/argininosuccinate lyase	4.3.2.1	1auw	PF00206
	Turkey δ crystallin[a]	Non-enzyme		
	Histidine ammonia-lyase	4.3.1.3	1b8f	PF00221
Helix-hairpin-helix base-excision DNA repair enzymes	Endonuclease III	4.2.99.18	2abk	PF00730
	DNA-3-methyladenine glycosylase	3.2.2.21	1mpg	PF00730
	A/G-specific adenine glycosylase	3.2.2.-	1mun	PF00730
	8-Oxoguanine DNA glycosylase	-	1ebm	
Non-heme di-iron carboxylate proteins	Ribonucleotide reductase, β chain	1.17.4.1	1xik	PF00268
	Methane monooxygenase hydroxylase component, α chain	1.14.13.25	1mty	
	Methane monooxygenase hydroxylase component, β chain	1.14.13.25 (non-catalytic protein chain)	1mty	
	Δ9 stearoyl-acyl carrier protein desaturase	1.14.99.6	1afr	PF00487
	Rubrerythrin	Non-enzyme	1b71	
	Bacterioferritin	Non-enzyme	1bcf	PF01334
	Ferritin heavy chain	Non-enzyme	2fha	PF00210
	Ferritin light chain	Non-enzyme	1dat	PF00210
	DNA protection during starvation protein	Non-enzyme	1dps	PF02047
Cytochrome P450s	Cytochrome *P*450-terp	1.14.-.-	1cpt	PF00067
	Cytochrome *P*450-cam	1.14.15.1	1phb	PF00067
	Cytochrome *P*450-eryF	1.14.-.-	1oxa	PF00067
	Cytochrome *P*450-bm3	1.14.14.1 1.6.2.4 [m]	1bu7	PF00067
	Cytochrome *P*450-nor (nitric oxide reductase)	1.14.-.-	1rom	PF00067
Ferredoxin-NADP reductase (FNR) modules	Nitrate reductase	1.6.6.1	2cnd	PF00175
	NADH-cytochrome b_5 reductase	1.6.2.2	1ndh	PF00175
	Phthalate dioxygenase reductase	1.-.-.-	2pia	PF00175
	Ferredoxin–NADP reductase	1.18.1.2	1fnb	PF00175
	NADPH-cytochrome P450 reductase	1.6.2.4	1amo	PF00175
	Flavohemoglobin	-	1cqx	PF00175
	NAD(P)H-flavin reductase	1.6.8.-	1qfj	PF00175
Cupredoxins	Rusticyanin	Non-enzyme	1rcy	PF00127
	Stellacyanin	Non-enzyme	1jer	PF00127
	Amicyanin	Non-enzyme	1aac	PF00127
	Azurin	Non-enzyme	1nwp	PF00127
	Pseudoazurin	Non-enzyme	1paz	PF00127
	Plastocyanin	Non-enzyme	1plc	PF00127
	Cucumber basic protein	Non-enzyme	2cbp	PF00127
	Copper-containing nitrite reductase	1.7.99.3	1nif	PF00394
	Laccase (polyphenol oxidase)	1.10.3.2	1a65	
	L-ascorbate oxidase	1.10.3.3	1aoz	PF00394
	Ceruplasmin (ferroxidase)	1.16.3.1	1kcw	PF00394
	Cytochrome *c* oxidase polypeptide II	1.9.3.1 (principal catalytic centre in a separate protein chain)	2cua	
	Ubiquinol oxidase polypeptide II	1.10.3.- (principal catalytic centre in a separate protein chain)	1cyw	
Discoidin-like domains[b]	Galactose oxidase	1.1.3.9	1gof	
	Sialidase	3.2.1.18	1eut	
	Coagulation factor V	Non-enzyme	1czt	PF00754
	Coagulation factor VIII	Non-enzyme	1d7p	PF00754
''Rieske''-like iron-sulphur domains[c]	Ubiquinol-cytochrome *c* reductase, iron-sulphur subunit	1.10.2.2	1rie	PF00355
	Cytochrome b_6f complex, iron-sulphur subunit	1.10.99.1	1rfs	PF00355
	Naphthalene 1,2-dioxygenase, α subunit	1.14.12.12	1ndo	PF00355
Hexapeptide repeat proteins	Chloramphenicol acetyltransferase	2.3.1.28	1xat	PF00132
	Tetrahydrodipicolinate N-succinyltransferase	2.3.1.117	3tdt	PF00132
	UDP-N-acetylglucosamine acyltransferase	2.3.1.129	1lxa	PF00132

Corteva Exhibit 2027 - Page 8 of 31
Bayer v. Corteva - IPR2023-01038

Superfamily	Member	EC number(s)	PDB code	Pfam
	γ-Carbonic anhydrase	4.2.1.1	1qre	PF00132
Ribulose-phosphate-binding TIM barrels	Indole-3-glycerol phosphate synthase/ N-(5′-phospho-ribosyl)anthranilate isomerase	4.1.1.48 5.3.1.24 [m]	1pii	PF00218 PF00697
	Indole-3-glycerol phosphate synthase	4.1.1.48	1igs	PF00218
	Orotidine 5′-monophosphate decarboxylase	4.1.1.23	1dv7	
	Tryptophan synthase, α chain	4.2.1.20	1ubs	PF00290
	Ribulose-phosphate 3-epimerase	5.1.3.1	1rpx	PF00834
Aldolase superfamily	*N*-acetylneuraminate lyase	4.1.3.3	1nal	PF00701
	Dihydrdipicolinate synthase	4.2.1.52	1dhp	PF00701
	Fructose-bisphosphate aldolase class I	4.1.2.13	1fba	PF00274
	Transaldolase	2.2.1.2	1onr	PF00923
	Type I dehydroquinate dehydratase	4.2.1.10	1qfe	PF01487
	δ-Aminolevulinic acid dehydratase	4.2.1.24	1aw5	PF00490
	Fructose-bisphosphate aldolase class II	4.1.2.13	1dos	PF01116
	Phe-sensitive 3-deoxy-D-*arabino*-heptulosonate-7-phosphate synthase	4.1.2.15	1qr7	PF00793
	3-Deoxy-D-*manno*-octulosonate-8-phosphate synthase	4.1.2.16	1d9e	
TIM barrel glycosyl hydrolases	Narbonin	Non-enzyme	1nar	
	Concanavalin B	Non-enzyme	1cnv	PF00192
	Hevamine (chitinase/lysozyme)	3.2.1.14 3.2.1.17 [s]	2hvm	PF00192
	Chitinase	3.2.1.14	1edq	PF00704
	Endo-β-*N*-acetylglucosaminidase	3.2.1.96	2ebn	
	Cyclodextrin glycosyltransferase	2.4.1.19	1pam	PF00128
	α-Amylase	3.2.1.1	1vjs	PF00128
	Oligo-1,6-glucosidase	3.2.1.10	1uok	PF00128
	Glucan 1,4-α-maltotetrahydrolase	3.2.1.60	2amg	PF00128
	Isoamylase	3.2.1.68	1bf2	PF00128
	α-Amylase II (neopullulanase)	3.2.1.135	1bvz	PF00128
	β-Amylase	3.2.1.2	1byb	PF01373
	Chitobiase	3.2.1.52	1qba	PF00728
	β-Glucosidase	3.2.1.21	1cbg	PF00232
	β-Glycosidase	3.2.1.23	1gow	PF00232
	6-Phospho-β-galactosidase	3.2.1.85	1pbg	PF00232
	Myrosinase	3.2.3.1	2myr	PF00232
	Endo-1,4-β-glucanase	3.2.1.4	1ceo	PF00150
	Exo-1,3-β-glucanase	3.2.1.58	1cz1	PF00150
	Endo-1,4-β-xylanase	3.2.1.8	1xyz	PF00331
	Exoglucanase/endo-1,4-β-xylanase	3.2.1.8 3.2.1.91 [s]	1exp	PF00331
	β-Galactosidase	3.2.1.23	1bgl	PF00703
	β-Glucuronidase	3.2.1.31	1bhg	PF00703
	Endo-1,3-β-glucanase	3.2.1.39	1ghs	PF00332
	Endo-1,3-1,4-β-glucanase	3.2.1.73	1aq0	PF00332
	β-Mannanase	3.2.1.78	1bqc	
	4-α-Glucanotransferase	2.4.1.25	1cwy	
Phosphoenol-pyruvate binding domains	Pyruvate, phosphate dikinase	2.7.9.1	1dik	PF00391
	Pyruvate kinase	2.7.1.40	1a49	PF00224
	Phosphoenolpyruvate carboxylase	4.1.1.31	1fiy	PF00311
Enolase superfamily	Enolase	4.2.1.11	1one	PF00113
	Mandelate racemase	5.1.2.2	2mnr	PF01188
	Muconate cycloisomerase I	5.5.1.1	1muc	PF01188
	Chloromuconate cycloisomerase	5.5.1.7	1chr	PF01188
	D-Glucarate dehydratase	4.2.1.40	1bqg	
FMN-dependent oxidoreductases	(S)-2-Hydroxy-acid oxidase (glycolate oxidase)	1.1.3.15	1gox	PF01070
	L-Lactate dehydrogenase (flavocytochrome b_2)	1.1.2.3	1fcb	PF01070
	Trimethylamine dehydrogenase	1.5.99.7	2tmd	PF00724
	NADPH dehydrogenase (old yellow enzyme)	1.6.99.1	1oya	PF00724
	Dihydroorotate dehydrogenase	1.3.3.1	1dor	PF01180
Metal-dependent hydrolases	Adenosine deaminase	3.5.4.4	1a4m	PF00962
	Phosphotriesterase	3.1.8.1	1psc	PF02126
	Urease, α subunit	3.5.1.5	2kau	PF00449
Type I PLP-dependent aspartate aminotransferase superfamily	Aspartate aminotransferase	2.6.1.1	1ars	PF00155
	Tyrosine aminotransferase	2.6.1.5	1bw0	PF00155
	Aromatic-amino-acid aminotransferase	2.6.1.57	2ay1	PF00155
	1-Aminocyclopropane-1-carboxylate synthase	4.4.1.14	1b8g	PF00155
	Ornithine decarboxylase	4.1.1.17	1ord	PF01276
	Tryptophanase	4.1.99.1	1ax4	PF01212
	Tyrosine phenol-lyase	4.1.99.2	1tpl	PF01212
	Cystathione γ-synthase	4.2.99.9	1cs1	PF01053

Superfamily	Member	EC number(s)	PDB code	Pfam
	Cystathione β-lyase	4.4.1.8	1cl1	PF01053
	Serine hydroxymethyltransferase	2.1.2.1	1bj4	PF00464
	8-Amino-7-oxononanoate synthase	2.3.1.47	1bs0	PF00222
	Glutamate-1-semialdehyde 2,1-aminomutase	5.4.3.8	2gsa	PF00202
	2,2-Dialkylglycine decarboxylase	4.1.1.64	1d7u	PF00202
	Ornithine aminotransferase	2.6.1.13	2oat	PF00202
	4-Aminobutyrate aminotransferase	2.6.1.19	1gtx	PF00202
	Adenosylmethionine-8-amino-7-oxononanoate aminotransferase	2.6.1.62	1dty	PF00202
	Phosphoserine aminotransferase	2.6.1.52	1bt4	PF00266
	3-Amino-5-hydroxybenzoic acid synthase	4.2.1	1b9h	
ATP-dependent carboxylate-amine/thiol ligases	Acetyl-CoA carboxylase, biotin carboxylase subunit	6.3.4.14	1bnc	PF00289
	D-alanine-D-alanine ligase	6.3.2.4	1iow	PF01820
	Glycinamide ribonucleotide synthetase	6.3.4.13	1gso	PF01071
	Phosphoribosylaminoimidazole carboxylase	4.1.1.21	1b6r	
	Glutathione synthetase	6.3.2.3	1gsh	
	Succinyl-CoA synthetase, β chain	6.2.1.5	2scu	PF00549
	Carbamoyl-phosphate synthase, large chain	6.3.5.5	1c30	PF00289
	Synapsin Ia	-	1auv	PF02078
	Pyruvate, phosphate dikinase	2.7.9.1	1dik	PF01326
Thioredoxin superfamily	Thioredoxin	Non-enzyme	2trx	PF00085
	Glutaredoxin	Non-enzyme	1kte	PF00462
	Thiol-disulphide interchange protein DsbA	-	1fvk	PF01323
	Protein disulphide isomerase	5.3.4.1	1mek	PF00085
	Glutathione *S*-transferase	2.5.1.18	1gse	PF00043
	Prostaglandin D-synthase	5.3.99.2	1pd2	PF00043
	Glutathione peroxidase	1.11.1.9	1gp1	PF00255
	Thioredoxin peroxidase 2	-	1qq2	PF00578
	Peroxidase Horf6	1.11.1.7	1prx	PF00578
	Phosducin	Non-enzyme	2trc	PF02114
	Calsequestrin	Non-enzyme	1a8y	PF01216
Phosphoglycerate mutase-like	Phosphoglycerate mutase	5.4.2.1	1qhf	PF00300
	6-Phosphofructo-2-kinase/fructose-2,6-biphosphatase	2.7.1.105 3.1.3.46 [m]	1bif	PF00300
	Prostatic acid phosphatase	3.1.3.2	1rpa	PF00328
	3-Phytase	3.1.3.8	1ihp	PF00328
α/β-Hydrolases	Carboxylesterase	3.1.1.1	1auo	
	Triacylglycerol lipase	3.1.1.3	1cul	PF00561
	Acetylcholine esterase	3.1.1.7	2ack	PF00135
	Bile-salt-activated cholesterol esterase	3.1.1.3 3.1.1.13 [s]	2bce	PF00135
	Para-nitrobenzyl esterase	3.1.1.-	1qe3	PF00135
	Brefeldin A esterase	-	1jkm	
	Prolyl oligopeptidase	3.4.21.26	1qfm	PF00326
	Serine carboxypeptidase I (cathepsin A)	3.4.16.5	1ivy	PF00450
	Serine carboxypeptidase II	3.4.16.6	1wht	PF00450
	Proline aminopeptidase	3.4.11.5	1azw	PF00561
	2-Hydroxy-6-oxo-6-phenylhexa-2,4-dienoate hydrolase	3.7.1.8	1c4x	
	Epoxide hydrolase	3.3.2.3	1cr6	PF00561
	Haloalkane dehalogenase	3.8.1.5	1b6g	PF00561
	(*S*)-acetone-cyanohydrin lyase	4.1.2.39	1qj4	PF00561
	Non-heme chloroperoxidase	1.11.1.10	1a8s	PF00561
	Myristoyl-ACP-specific thioesterase	2.3.1.-	1tht	
	Dienelactone hydrolase	3.1.1.45	1din	PF01738
	Acetylxylan esterase	3.1.1.6	1bs9	
	Cutinase (serine esterase)	3.1.1.-	1cex	PF01083
TPP-dependent enzymes	Pyruvate oxidase	1.2.3.3	1pox	PF00205
	Pyruvate decarboxylase	4.1.1.1	1pvd	PF00205
	Benzoylformate decarboxylase	4.1.1.7	1bfd	PF00205
	Transketolase	2.2.1.1	1trk	PF00456
	Pyruvate-ferredoxin reductase	1.2.7.1	1b0p	
	2-Oxoisovalerate dehydrogenase, α subunit	1.2.4.4	1qs0	PF00676
	2-Oxoisovalerate dehydrogenase, β subunit	1.2.4.4	1qs0	
5′-Nucleases	DNA polymerase I	2.7.7.7	1bgx	PF01367
	Ribonuclease H	3.1.26.4	1tfr	
	5′-exonuclease	3.1.11.3	1exn	PF01367
	Flap endonuclease-1	-	1a76	

Superfamily	Member	EC number(s)	PDB code	Pfam
FAD/NAD(P)(H) dependent disulphide oxidoreductase superfamily	Glutathione reductase	1.6.4.2	3grs	PF00070
	Low molecular weight thioredoxin reductase	1.6.4.5	1tde	PF00070
	Trypanothione reductase	1.6.4.8	1fec	PF00070
	Dihydrolipoamide dehydrogenase	1.8.1.4	3lad	PF00070
	Flavocytochrome *c*:sulphide dehydrogenase	1.8.2.-	1fcd	
	NADH peroxidase	1.11.1.1	1npx	PF00070
	Mercuric ion reductase[a]	1.16.1.1		
	Adrenodoxin reductase	1.18.1.2	1cjc	
	Trimethylamine dehydrogenase	1.5.99.7	2tmol	
Zn peptidases	Carboxypeptidase A	3.4.17.1	2ctc	PF00246
	Carboxypeptidase B	3.4.17.2	1nsa	PF00246
	Carboxypeptidase A2	3.4.17.15	1aye	PF00246
	Carboxypeptidase T	3.4.17.18	1obr	PF00246
	Bacterial leucyl aminopeptidase	3.4.11.10	1amp	PF01546
	Carboxypeptidase G2	3.4.17.11	1cg2	PF01546
	Cytosol aminopeptidase	3.4.11.1 3.4.11.5 [s]	1lam	PF00883
	Aminopeptidase	3.4.11.-	1xjo	
	Transferrin receptor protein	Non-enzyme	1cx8	
DD-peptidase/β-lactamase superfamily	D-alanyl-D-alanine carboxypeptidase/ transpeptidase	3.4.16.4	3pte	
	β-Lactamase	3.5.2.6	1dja	PF00144
	Penicillin-binding protein	-	1pmd	PF00905
Phosphohistidine domains of PEP-utilising enzymes	Pyruvate, phosphate dikinase	2.7.9.1	1dik	PF00391
	Phosphoenolpyruvate-protein phosphotransferase	2.7.3.9	2ezb	PF00391
Medium-chain alcohol dehydrogenases	Quinone oxidoreductase	1.6.5.5	1qor	PF00107
	Alcohol dehydrogenase	1.1.1.1	1deh	PF00107
	Alcohol dehydrogenase class III	1.1.1.1 1.2.1.1 [s]	1teh	PF00107
	NADP-dependent alcohol dehydrogenase	1.1.1.2	1kev	PF00107
	Glucose dehydrogenase[a]	1.1.1.47		
Protein tyrosine phosphatase superfamily	Protein-tyrosine phosphatase	3.1.3.48	1bzh	PF00102
	Dual specificity protein phosphatase	3.1.3.16 3.1.3.48 [s]	1vhr	PF00782
Crotonase-like	4-Chlorobenzoyl-CoA dehalogenase	3.8.1.6	1nzy	
	Enoyl-CoA hydratase	4.2.1.17	2dub	PF00378
	Δ3,5-Δ2,4-dienoyl-CoA isomerase	5.3.3.-	1dci	PF00378
	ATP-dependent Clp protease, proteolytic subunit	3.4.21.92	1tyf	PF00574
Creatinase/methionine aminopeptidase superfamily	Creatinase	3.5.3.3	1chm	PF00557
	Methionine aminopeptidase	3.4.11.18	1xgs	PF00557
	Aminopeptidase P	3.4.11.9	1a29	PF00557

[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, isozymes 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.

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

[b] 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.

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

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 α/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 3. Summary of variations at the domain and quaternary levels

Superfamily	Domain enlargement	Domain duplication	Domain recruitment/ loss	Domain rearrangement	Subunit assembly	Motif duplication
Fumarase/aspartase	83-195		2-3			
Helix-hairpin-helix base-excision DNA repair enzymes	51-113		2-3			
Non-heme di-iron carboxylate proteins	146-512		1-2		*2-24	*
Cytochrome *P*450s			1-3+			
Ferredoxin-NADP reductase (FNR) modules			2-5		1-2	
Cupredoxins	96-214	S	1-6		* 2-?	
Discoidin-like domains		*	3-9	*	* 1-2	*
Rieske'-like iron-sulphur domains			2		* 6-22	
Hexapeptide repeat proteins			1-2			*
Ribulose-phosphate-binding TIM barrels		S	1-2		* 1-6	
Aldolase superfamily					2-8	
TIM barrel glycosyl hydrolases		*	1-5	*	1-4	*
Phosphoenol-pyruvate binding domains	274-883		1-5		2-4	
Enolase superfamily					2-8	
FMN-dependent oxidoreductases		*	1-3		1-4	
Metal-dependent hydrolases			1-2		* 1-9	
Type I PLP-dependent aspartate aminotransferase superfamily			2-4		2-12	
ATP-dependent carboxylate-amine/thiol ligases	121-288, 69-155	S	3-8	*	* 1-4	
Thioredoxin superfamily	75-184	S	1-4		* 1-N	
Phosphoglycerate mutase-like	222-431		1-2		2-4	
α/β-hydrolases	197-547		1-2		* 1-8	*
TPP-dependent enzymes	178-395	S	1-5	*	* 2-4	
5′-Nucleases			2-6			
FAD/NAD(P)(H) dependent disulphide oxidoreductase superfamily		S	2-5		* 1-4	
Zn peptidases			1-4		1-6	
DD-Peptidase/β-lactamase superfamily		*	1-5			
Phosphohistidine domains of PEP-utilising enzymes			3-5			
Medium-chain alcohol dehydrogenases					2-4	
Protein tyrosine phosphatase superfamily	144-297	S	1-14			
Crotonase-like					3-14	
Creatinase/methionine aminopeptidase superfamily			1-2		1-4	*

The Table indicates those superfamilies in which (i) the sizes of at least one domain common to all members vary by at least two-fold, 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).

wide variability of the substrates, which are bound to the core of cytochrome *P*450s 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

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 aminotransferase 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

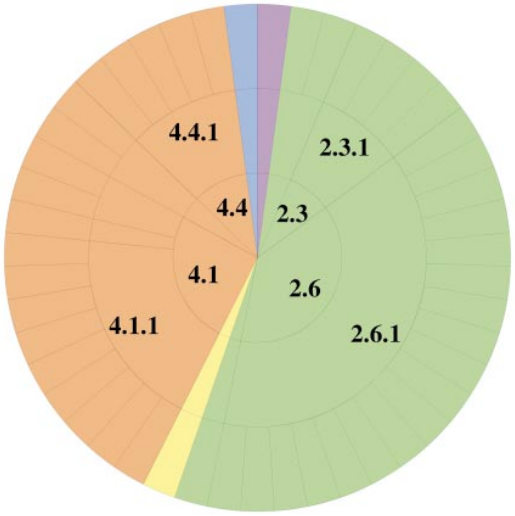

Figure 6. Diversity of enzyme functions catalysed by members of the PLP-dependent type I aspartate aminotransferase superfamily. Functions of both structural and sequence members are included, and incomplete EC numbers have been ignored. This representation comprises a set of concentric pie charts. Sectors are coloured according to the enzyme class (pink, oxidoreductases; green, transferases; yellow, hydrolases; orange, lyases; blue, isomerases). The circles, from inner to outer, represent the second, third and fourth levels in the EC hierarchy. The angle subtended by any segment reflects the proportion of enzyme functions it contains. At least 47 enzyme functions are catalysed by this superfamily. Some of the more populated sectors are labelled with the EC levels they denote.

Table 4. Conservation and variation of substrate specificity, reaction chemistry and catalytic residues

Superfamily	Substrate specificity	Co-factor	Similarity in reaction chemistry	Conservation of chemistry	Conservation of principal catalytic residues	Variability of functionally equivalent residues
Fumarase/aspartase	Similar; moiety-carboxylate group	-	β-Elimination reactions involving cleavage of C-N or C-O bonds α to the carboxylate group; β-proton abstraction	Semi-conserved	No	*
Helix-hairpin-helix base-excision DNA repair enzymes	Type-DNA substrates; diverse	-	Nucleophilic activation involving conserved aspartate	Semi-conserved	Yes	
Non-heme di-iron carboxylate proteins	Diverse	Two Fe atoms	O_2/H_2O_2 activation and di-iron oxidation	Semi-conserved	Yes	
Cytochrome *P*450s	Diverse	heme-thiolate	Activated iron-oxygen species [with exception of nitric oxide reductase, but its reaction mechanism follows the same principle]	Semi-conserved	Yes	
Ferredoxin-NADP reductase (FNR) modules	N/A	FAD or FMN	Hydride transfer between flavin and NAD(P)(H); direction varies	Conserved	Yes	
Cupredoxins	Diverse	Cu	Intramolecular electron transfer; Cu-dependent oxidoreductase activity	Semi-conserved	Yes	
Discoidin-like domains	Type-cell-surface carbohydrates	N/A	N/A	-	-	
''Rieske''-like iron-sulphur domains	N/A	N/A	N/A	-	-	
Hexapeptide repeat proteins	Diverse	Unconserved metal-binding site	Varied	Unconserved	No	*
Ribulose-phosphate-binding TIM barrels	Similar; moiety-ribulose phosphate	-	Varied	Unconserved	No	
Aldolase superfamily	Diverse; moiety-carbonyl group	Unconserved divalent metal-binding sites	Carbonyl chemistry	Poorly-conserved	No	*
TIM barrel glycosyl hydrolases	Type-carbohydrates	-	General acid catalysis-glutamate protonates scissile glycosidic oxygen	Semi-conserved	No	*
Phosphoenol-pyruvate binding domains	Conserved	Divalent metal	detailed mechanisms unknown	-		
Enolase superfamily	Diverse; moiety-carboxylate group	Divalent metal	metal-assisted abstraction of proton α to carboxylic acid	Semi-conserved	Semi	*
FMN-dependent oxidoreductases	Diverse	FMN	FMN-dependent oxidoreductase activity	Semi-conserved	No	*
Metal-dependent hydrolases	Diverse	One or two divalent metal atoms	Metal-assisted activation of water for nucleophilic attack on substrate	Semi-conserved	Yes	

Corteva Exhibit 2027 - Page 14 of 31
Bayer v. Corteva - IPR2023-01038

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

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 5. Common chemistry in the context of diverse enzyme reactions

Superfamily	Common reaction step	Enzyme	Activity
Non-heme di-iron carboxylate proteins	O_2/H_2O_2 activation and di-iron oxidation	Methane monooxygenase Δ(9) Stearoyl-acyl carrier protein desaturase Ribonucleotide reductase	Hydroxylation of methane Oxidative desaturation of a fatty acid Formation of an active-site tyrosine radical; this is involved in a long-range electron transfer with the substrate bound to a separate subunit
Cytochrome *P*450s	Activated iron-oxygen species (with exception of nitric oxide reductase, but its reaction mechanism follows the same principle (Halkier, 1996))	Cytochrome *P*450-cam Nitric oxide reductase Berbamunine synthase* Allene oxide synthase* Thromboxane-A synthase*	Hydroxylation (monooxygenase activity) Nitric oxide reduction Oxidative dimerisation (oxidase activity) Epoxide formation Isomerisation (intramolecular oxidoreduction)
Enolase superfamily	Metal-assisted abstraction of proton α to carboxylic acid	Enolase Methylaspartate ammonia-lyase* Mandelate racemase Muconate cycloisomerase D-Glucarate dehydratase	β-Elimination of water β-Elimination of ammonia Racemisation Cycloisomerisation Epimerisation
Type I PLP-dependent aspartate aminotransferase superfamily	Covalent binding of PLP to conserved lysine and formation of aldimine intermediate with substrate	Aspartate aminotransferase Ornithine decarboxylase Serine hydroxymethyltransferase Tryptophanase Methionine γ-lyase* Cystathione γ-synthase Glutamate-1-semialdehyde 2,1-aminomutase	Transamination α-Decarboxylation Retro-aldol cleavage β-Elimination γ-Elimination γ-Replacement Isomerisation (intramolecular exchange)
TPP-dependent enzymes	Conserved glutamate activates TPP for attack on the substrate carbonyl, leading to C-C bond cleavage and formation of a TPP-aldehyde intermediate	Pyruvate oxidase Pyruvate decarboxylase Transketolase Benzaldehyde lyase*	Oxidative decarboxylation Decarboxylation Ketol group transfer Reverse aldol condensation
α/β-Hydrolases	Conserved histidine activates the catalytic nucleophile which forms a covalent intermediate with the substrate	Lipase Haloalkane dehalogenase Epoxide hydrolyase Serine carboxypeptidase (*S*)-Acetone-cyanohydrin lyase Non-heme chloroperoxidase Myristoyl-ACP-specific thioesterase	Carboxylic ester hydrolysis Halide hydrolysis Ether hydrolysis Peptide hydrolysis Decomposition of hydroxynitrile Halogenation of organic compounds Acyl transfer

The Table summaries 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.

Corteva Exhibit 2027 - Page 16 of 31
Bayer v. Corteva - IPR2023-01038

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 are cofactor-independent acyltransferases (EC 2.3.1.-), and bind acyl-CoA or acylated-acyl protein carrier as a donor substrate. However, this family of enzymes includes 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).

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.

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 peroxo 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

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 β-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, β-strands four or five in the central β-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 α/β 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

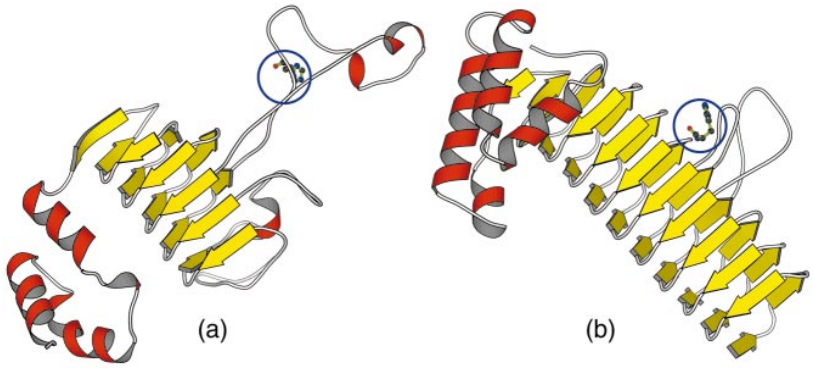

Figure 7. MOLSCRIPT (Kraulis, 1991) diagrams of the homologous enzymes (a) chloramphenicol acetyltransferase (PaXAT), and (b) UDP-*N*-acetylglucosamine acyltransferase (LpxA). The catalytic histidine residues putatively involved in deprotonation of the substrate hydroxyl are shown in ball-and-stick and 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).

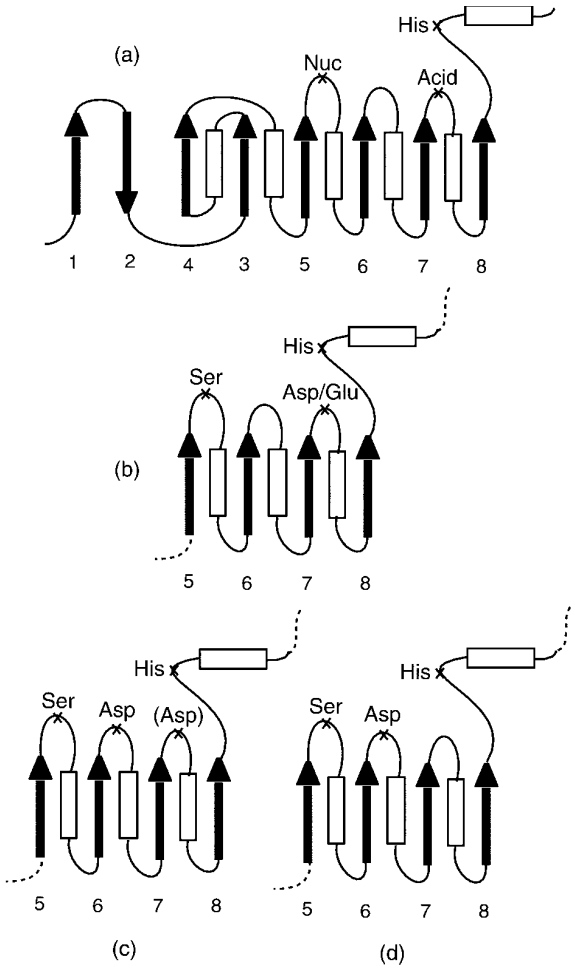

Figure 8. Topology diagrams to illustrate migration of the catalytic acid in members of the α/β hydrolase superfamily (Schrag *et al.*, 1992). Strands are represented by arrows and helices by rectangles. Residues belonging to the catalytic triad are indicated (Nuc represents the catalytic nucleophile) but note the Asp in brackets corresponds to an evolutionary relic not involved in catalysis. (a) The canonical α/β hydrolase fold (Ollis *et al.*, 1992) and strands 5-8 of (b) acetylcholinesterases, fungal lipases and bovine salt-activated cholesterol esterase (Heikinheimo *et al.*, 1999); (c) the ''evolutionary intermediate'' human pancreatic lipase, and (d) lipoprotein lipases, by homology (Schrag *et al.*, 1992).

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.

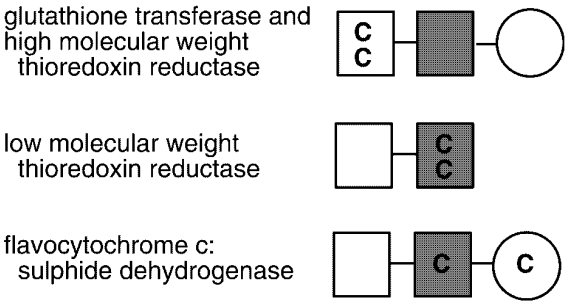

Figure 9. Schematic diagram to illustrate the domain organisation and active site location of members of the FAD/NAD(P)(H)-dependent disulphide oxidoreductase superfamily. Each shape corresponds to a structural domain, and domains of the same shape are homologous. The light grey square represents the FAD-binding domain, the dark grey square represents the NAD(P)(H)-binding domain and the circle represents the interface domain. The NAD(P)(H)-binding domain interrupts the first domain in all proteins, but this is not indicated in the diagram for clarity. Each cysteine in the catalytic disulphide is represented by C to illustrate its domain location. High molecular mass thioredoxin reductase has unknown structure, but sequence comparisons have shown that it has an active site location and domain organisation identical to those of glutathione transferase.

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 Å 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.

Corteva Exhibit 2027 - Page 20 of 31
Bayer v. Corteva - IPR2023-01038

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 oxidoreductase lacks both the catalytic and structural Zn sites of its homologues, the functionally analagous 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 Phe-sensitive 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 (Rydén & 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 carboxylate 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 3.1.1.3 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 αβα 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 (Hoeffken *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-

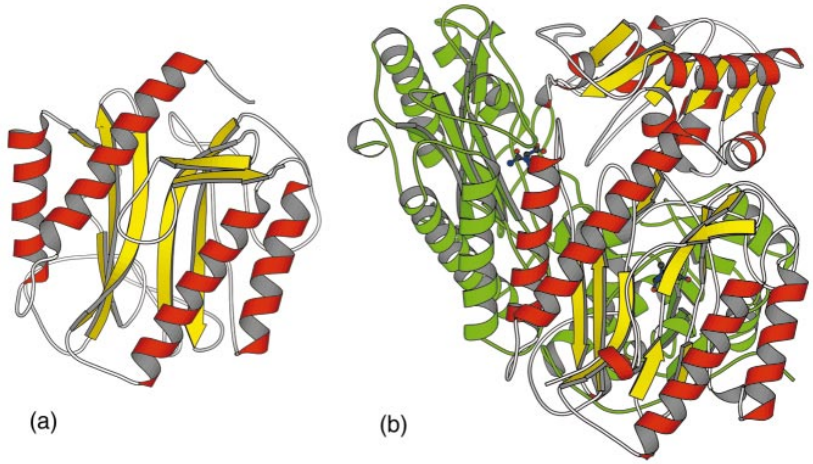

Figure 10. MOLSCRIPT diagrams of (a) prokaryotic methionine aminopeptidase and (b) the creatinase homodimer; the additional subunit is coloured green, and its extra N-terminal domain shapes the active site for the recognition of its smaller substrate, creatine (Hoeffken *et al.*, 1988); the inhibitor carbamoyl sarcosine is shown in ball-and-stick.

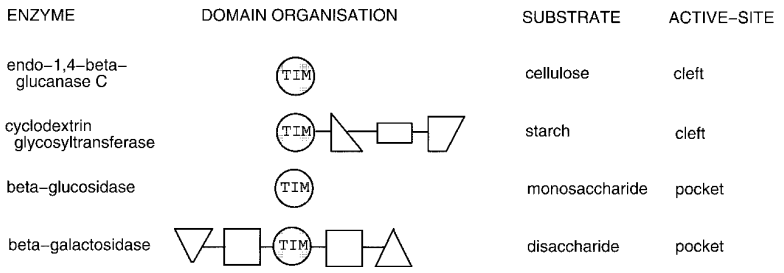

Figure 11. Domain organisation, substrate preferences and active-site shapes of four members of the TIM barrel glycosyl hydrolase superfamily. Each shape corresponds to a structural domain, and domains of the same shape are homologous. The catalytic domains are shaded.

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 *P*450BM-3 the redox partner, NADPH-cytochrome *P*450 reductase, is fused to the *P*450 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 phosphoribosylaminoimidazolesuccinocarboxamide (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 (AEvarsson *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 of separate chains (Muller *et al.*, 1993). Pyruvate ferrodoxin reductase is an exception; TPP is bound by two domains on the same subunit (Chabriere *et al.*, 1999). Despite these variations, 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.

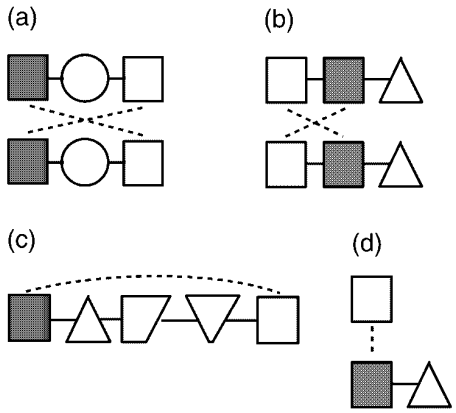

Figure 12. Domain organisation and TPP-binding units of TPP-dependent enzymes. Each shape corresponds to a structural domain, and domains of the same shape are homologous. The light and dark grey squares represent the TPP-binding domains involved in interacting with the pyrimidine ring and diphosphate 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.

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

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 limitatations. 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 ilustrates 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 oxyanion 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 proteins 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 ligase 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

Corteva Exhibit 2027 - Page 25 of 31
Bayer v. Corteva - IPR2023-01038

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.*,

1998; Colovos *et al*., 1998; Zarembinski *et al*., 1998; Volz, 1999), nevertheless new structures almost invariably provide valuable guidance for future biochemical studies.

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 articles referred to in this paper.

Every protein chain in the PDB was matched to its corresponding SWISS-PROT sequence entry, where possible, 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 (Altschul *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 discontiguous 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 ''hypothetical'' 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-SPROT 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

Aevarsson, A., Seger, K., Turley, S., Sokatch, J. R. & Hol, W. G. J. (1999). Crystal structure of 2-oxoisovalerate dehydrogenase and the architecture of 2-oxo acid dehydrogenase multienzyme complexes. *Nature Struct. Biol.* **6**, 785-792.

Alexander, F. W., Sandmeier, E., Metha, P. K. & Christen, P. (1994). Evolutionary relationships among pyridoxal-5′-phosphate-dependent enzymes. *Eur. J. Biochem.* **219**, 953-960.

Altschad, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Aravind, L., Galperin, M. Y. & Koonin, E. V. (1998). The catalytic domain of the P-type ATPase has the haloacid dehalogenase fold. *Trends Biochem. Sci.* **23**, 127-129.

Artymiuk, P. J., Poirrette, A. R., Rice, D. W. & Willett, P. (1997). A polymerase I palm in adenylyl cyclase? *Nature,* **388**, 33-34.

Ashburner, M. & Drysdale, R. (1994). Fly Base: the *Drosophila* genetic database. *Development,* **120**, 2077-2079.

Babbitt, P. C. & Gerlt, J. A. (1997). Understanding enzyme superfamilies - chemistry as the fundamental determinant in the evolution of new catalytic activities. *J. Biol. Chem.* **272**, 30591-30594.

Bairoch, A. (2000). The enzyme database in 2000. *Nucl. Acids Res.* **28**, 304-305.

Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement trEMBL in 2000. *Nucl. Acid Res.* **28**, 45-48.

Bamyard, S. H., Stammers, D. K. & Harrison, P. M. (1978). Electron density map of apoferritin at 2.8 Å resolution. *Nature,* **271**, 282-284.

Barrett, A. J. (1994). Classification of peptidases. *Methods Enzymol.* **244**, 1-15.

Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucl. Acids Res.* **28**, 260-262.

Beaman, T. W., Sugantino, M. & Roderick, S. L. (1998). Structure of the hexapeptide xenobiotic acetyltransferase from *Pseudomonas aervginosa*. *Biochemistry,* **37**, 6689-6696.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D. & Rodgers, J. R. *et al*. (1977). The Protein Data Bank: a computer-based archival file for macro-molecular structures. *J. Mol. Biol.* **112**, 535-542.

Carpenter, E. P., Hawkins, A. R., Frost, J. W. & Brown, K. A. (1998). Structure of dehydroquinate synthase reveals an active site capable of multistep catalysis. *Nature,* **394**, 299-302.

Chabriere, E., Charon, M.-H., Volbeda, A., Pieulle, L., Hatchikian, E. C. & Fontecilla-Camps, J.-C. (1999). Crystal structures of the key anaerobic enzyme pyruvate:ferredoxin oxidoreductase, free and in complex with pyruvate. *Nature Struct. Biol.* **6**, 182-190.

Chem, J. C. H., Miercke, L. J. W., Krucinski, J., Starr, J. R., Saenz, G. & Wang, X. B. *et al*. (1998). Structure of bovine pancreatic cholesterol esterase at 1.6 Å: novel structural features involved in lipase activation. *Biochemistry,* **37**, 5107-5117.

Chen, Z. W., Koh, M., Van Driessche, G., Van Beeumen, J. J., Bartsch, R. G. & Meyer, T. E., *et al*. (1994). The structure of flavocytochrome-*c* sulfide dehydrogenase from a purple phototrophic bacterium. *Science,* **266**, 430-432.

Chothia, C. (1992). Proteins - 1000 families for the molecular biologist. *Nature,* **357**, 543-544.

Colovos, C., Cascio, D. & Yeates, T. O. (1998). The 1.8 Å crystal structure of the ycaC gene product from *Escherichia coli* reveals an octameric hydrolase of unknown specificity. *Structure,* **6**, 1329-1337.

Copley, R. R. & Bork, P. (2000). Homology among $(\beta\alpha)_8$ barrels: implications for the evolution of metabolic pathways. *J. Mol. Biol.* **303**, 627-640.

Devos, D. & Valencia, A. (2000). Practical limits of function prediction. *Proteins: Struct. Funct. Genet.* **41.1**, 98-107.

Erskine, P. T., Newbold, R., Roper, J., Coker, A., Warren, M. J. & Shoolingin-Jordan, P. M. *et al*. (1999). The Schiff base complex of yeast 5-aminolaevulinic acid dehydratase with laevulinic acid. *Protein. Sci.* **8**, 1250-1256.

Ford, G. C., Harrison, P. M., Rice, D. W., Smith, I. M. A., Treffry, A., White, J. L. & Yariv, J. (1984). Ferritin: design and formation of an iron-storage molecule. *Phil Trans. Roy. Soc. London,* **304**, 551-564.

Fraaije, M. W. & Mattevi, A. (2000). Flavoenzymes: diverse catalysts with recurrent features. *Trends Biochem. Sci.* **25**, 126-132.

Galperin, M. Y., Walker, D. R. & Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res.* **8**, 779-790.

Galperin, M. Y., Aravind, L. & Koonin, E. V. (2000). Aldolases of the DhnA family: a possible solution to the problem of pentose and hexose biosynthesis in Archaea. *Fems. Microbial. Letters,* **183**, 259-264.

Gerlt, J. A. & Babbitt, P. C. (1998). Mechanistically diverse enzyme superfamilies: the importance of chemistry in the evolution of catalysis. *Curr. Opin. Chem. Biol.* **2**, 607-612.

Grant, R. A., Filman, D. J., Finkel, S. E., Kolter, R. & Hogle, J. M. (1998). The crystal structure of Dps, a ferritin homolog that binds and protects DNA. *Nature Struct. Biol.* **5**, 294-303.

Grishin, N. V. (2000). Two tricks in one bundle: helix-turn-helix gains enzymatic activity. *Nucl. Acids Res.* **28**, 2229-2233.

Guan, Y., Manuel, R. C., Arvai, A. S., Parikh, S. S., Mol, C. D. & Miller, J. H. *et al*. (1998). MutY catalytic core, mutant and bound adenine structures define specificity for DNA. *Nature Struct. Biol.* **5**, 1058-1064.

Halkier, B. A. (1996). Catalytic reactivities and structure/function relationships of cytochrome *P*490 enzymes. *Phytochemistry,* **43**, 1-21.

Hasson, M. S., Muscate, A., McLeish, M. J., Polovnikova, L. S., Gerlt, J. A. & Kenyon, G. L. *et al*. (1998a). The crystal structure of benzoylformate decarboxylase at 1.6 Angstrom resolution: diversity of catalytic residues in thiamin diphosphate-dependent enzymes. *Biochemistry,* **37**, 9918-9930.

Hasson, M. S., Schlichting, I., Moulai, J., Taylor, K., Barrett, W. & Kenyon, G. L. *et al*. (1998b). Evolution of an enzyme active site: the structure of a new crystal form of muconate lactonizing enzyme compared with mandelate racemase and enolase. *Proc. Natl Acad. Sci. USA,* **95**, 10396-10401.

Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147-164.

Heikinheimo, P., Goldman, A., Jeffries, C. & Ollis, D. L. (1999). Of barn owls and bankers: a lush variety of α/β hydrolases. *Structure,* **7**, r141-r146.

Henrissat, B. (1991). A classification of glycosyl hydrolases based on amino-acid-sequence similarities. *Biochem. J.* **280**, 309-316.

Hoeffken, H. W., Siegward, H. K., Bartlett, P. A. & Huber, R. (1988). Crystal structure determination, refinement and molecular model of creatine amidinohydrolase from *Pseudomonas putida*. *J. Mol. Biol.* **204**, 417-433.

Holrn, L. & Sander, C. (1996). Mapping the protein universe. *Science,* **273**, 595-602.

Hyde, C. C., Ahmed, S. A., Padlan, E. A., Miles, E. W. & Davies, D. R. (1988). Three-dimensional structure of the tryptophan synthase $\alpha_2\beta_2$ multienzyme complex from *Salmonella typhimurivm*. *J. Biol. Chem.* **263**, 17857-17871.

Jayasekera, M. M. K., Shi, W. X., Farber, G. K. & Viola, R. E. (1997). Evaluation of functionally important amino acids in L-aspartate ammonia-lyase from *Escherichia coli*. *Biochemistry,* **36**, 9145-9150.

Jeffery, C. J. (1999). Moonlighting proteins. *Trends Biochem. Sci.* **24**, 8-11.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annu. Rev. Microbiol.* **30**, 409-425.

Jia, J., Huang, W. J., Schorken, U., Sahm, H., Sprenger, G. A., Lindqvist, Y. & Schneider, G. (1996). Crystal structure of transaldolase B from *Escherichia coli* suggests a circular permutation of the α/β barrel within the class I aldolase family. *Structure,* **4**, 715-724.

Juers, D. H., Huber, R. E. & Matthews, B. W. (1999). Structural comparisons of TIM barrel proteins suggest functional and evolutionary relationships between β-galactosidase and other glycohydrolases. *Protein Sci.* **8**, 122-136.

Kai, Y., Matsumura, H., Inoue, T., Terada, K., Nagara, Y. & Yoshinaga, T. *et al*. (1999). Three-dimensional structure of phosphoenolpyruvate carboxylase: a proposed mechanism for allosteric inhibition. *Proc. Natl Acad. Sci. USA,* **96**, 823-828.

Kraulis, P. J. (1991). MOLSCRIPT - a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946 950.

Kuriyan, J., Krishna, T. S. R., ., Wong, L., Guenther, B., Pahler, A., Williams, C. H. & Model, P. (1991). Convergent evolution of similar function in two structurally divergent enzymes. *Nature,* **352**, 172-174.

Laskowski, R. A., Hutchinson, E. G., Michie, A. D., Wallace, A. C., Jones, M. L. & Thornton, J. M. (1997). PDBsum: a web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**, 488-490.

Lee, M., Lenman, M., Banas, A., Bafor, M., Singly, S. & Schweizer, M. *et al*. (1998). Identification of non-heme diiron proteins that catalyze triple bond and epoxy group formation. *Science,* **280**, 915-918.

Levdikov, V. M., Barynin, V. V., Grebenko, A. I., Melikadamyan, W. R., Lamzin, V. S. & Wilson, K. S. (1998). The structure of SAICAR synthase: an enzyme in the *de novo* pathway of purine nucleotide biosynthesis. *Structure,* **6**, 363-376.

Longhi, S., Czjzek, M., Lainzin, V., Nicolas, A. & Cambillau, C. (1997). Atomic resolution (1.0 Å) crystal structure of *Fusarium solani* cutinase: stereochemical analysis. *J. Mol. Biol.* **268**, 779-799.

Lowther, W. T., Zhang, Y., Sampson, P. B., Honek, J. F. & Matthews, B. W. (1999). Insights into the mechanism of *Escherichia coli* methionine aininopeptidase from the structural analysis of reaction products and phosphorus-based transition-state analogues. *Biochemistry,* **38**, 14810-14819.

Makarova, K. S. & Grishin, N. V. (1999). The Zn-peptidase superfamily: functional convergence after evolutionary divergence. *J. Mol. Biol.* **292**, 11-17.

Martin, A. C. R., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M. & Laskowski, R. A., *et al*. (1998). Protein folds and functions. *Structure,* **6**, 875-884.

Martin, J. L. (1995). Thioredoxin - a fold for all reasons. *Structure,* **3**, 245-250.

Mol, C. D., Parikh, S. S., Putman, C. D., Lo T., P. & Tainer, J. A. (1999). DNA repair mechanisms for the recognition and removal of damaged DNA bases. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 101-128.

Muller, Y. A., Lindqvist, Y., Furey, W., Schulz, G. E., Jordan, F. & Schneider, G. (1993). A thiamin diphosphate binding fold revealed by comparison of the crystal structures of transketolase, pyruvate oxidase and pyruvate decarboxylase. *Structure,* **1**, 95-103.

Murphy, M. E. P., Lindley, P. F. & Adman, E. T. (1997). Structural comparison of cupredoxin domains: domain recycling to construct proteins with novel functions. *Protein Sci.* **6**, 761-770.

Murzin, A. G. (1993). Can homologous proteins evolve different enzymatic activities. *Trends Biochem. Sci.* **18**, 403-405.

Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* **6**, 386 394.

Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380-387.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). SCOP - a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-540.

Nardini, M. & Dijkstra, B. W. (1999). α/β hydrolase fold enzymes: the family keeps growing. *Curr. Opin. Struct. Biol.* **9**, 732-737.

Narhi, L. O. & Fulco, A. J. (1987). Identification and characterisation of the two functional domains in cytochrome P450 BM-3, a catalytically self-sufficient monooxygenase induced by barbituates in *Bacillus megaterium*. *J. Biol. Chem.* **292**, 6883-6890.

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.

Nishida, M., Harada, S., Noguchi, S., Satow, Y., Inoue, H. & Takahashi, K. (1998). Three-dimensional structure of *Escherichia coli* glutathione *S*-transferase complexed with glutathione sulfonate: catalytic roles of Cys10 and His106. *J. Mol. Biol.* **281**, 135-147.

Nordlund, P., Sjoberg, B.-M. & Eklund, H. (1990). Three-dimensional structure of the free radical protein of ribonucleotide reductase. *Nature,* **345**, 593-598.

O'Brien, P. J. & Herschlag, D. (1999). Catalytic promiscuity and the evolution of new enzymatic activities. *Chem. Biol.* **6**, r91-r105.

Ollis, D. L., Clteah, E., Cygler, M., Dijkstra, B., Frolow, F. & Franken, S. M. *et al*. (1992). The α/β hydrolase fold. *Protein Eng.* **5**, 197-211.

Orengo, C. A. (1999). CORA - topological fingerprints for protein structural families. *Protein Sci.* **8**, 699-715.

Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature,* **372**, 631-634.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH - a hierarchic classification of protein domain structures. *Structure,* **5**, 1093-1108.

Pearson, W. R. & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA,* **85**, 2444-2448.

Perona, J. J. & Craik, C. S. (1997). Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *J. Biol. Chem.* **272**, 29987-29990.

Petsko, G. A. (1991). Dèja vu all over again. *Nature,* **352**, 104-105.

Petsko, G. A., Kenyon, G. L., Gerlt, J. A., Ringe, D. & Kozarich, J. W. (1993). On the origin of enzymatic species. *Trends Biochem. Sci.* **18**, 372-376.

Piatigorsky, J. & Wistow, G. (1991). The recruitment of crystallins - new functions precede gene duplication. *Science,* **252**, 1078-1079.

Pihlajaniemi, T., Helaakoski, T., Tasanen, K., Myllyla, R., Huhtala, M. L., Koivu, J. & Kivirikko, K. I. (1987). Molecular cloning of the β-subunit of human prolyl-4-hydroxylase. This subunit and protein disulphide isomerase are products of the same gene. *EMBO J.* **6**, 643-649.

Polekhina, G., Board, P. G., Gali, R. R., Rossjohn, J. & Parker, M. W. (1999). Molecular basis of glutathione synthetase deficiency and a rare gene permutation event. *EMBO J.* **18**, 3204-3213.

Poulos, T. L. (1995). Cytochrome P450. *Curr. Opin. Struct. Biol.* **5**, 767-774.

Ren, B., Tibbelin, G., Depascale, D., Rossi, M., Bartolucci, S. & Ladenstein, R. (1998). A protein disulfide oxidoreductase from the archaeon *Pyrococcus furiosus* contains two thioredoxin fold units. *Nature Struct. Biol.* **5**, 602-611.

Rojas, A., Garciavallve, S., Palau, J. & Romeu, A. (1999). Circular permutations in proteins. *Biologia,* **54**, 255-277.

Russell, R. B. & Ponting, C. P. (1998). Protein fold irregularities that hinder sequence analysis. *Curr. Opin. Struct. Biol.* **8**, 364-371.

Russell, R. B., Sasieni, F. D. & Sternberg, M. J. E. (1998). Supersites within superfolds. Binding site similarity in the absence of homology. *J. Mol. Biol.* **282**, 903-918.

Rydén, L. G. & Hunt, L. T. (1993). Evolution of protein complexity - the blue copper-containing oxidases and related proteins. *J. Mol. Evol.* **36**, 41-66.

Schofield, C. J. & Zhang, Z. (1999). Structural and mechanistic studies on 2-oxoglutarate-dependent oxygenases and related enzymes. *Curr. Opin. Struct. Biol.* **9**, 722-731.

Schrag, J. D., Winkler, F. K. & Cygler, M. (1992). Pancreatic lipases - evolutionary intermediates in a positional change of catalytic carboxylates. *J. Biol. Chem.* **267**, 4300-4303.

Schrag, J. D., Vernet, T., Laramee, L., Thomas, D. Y., Recktenwald, A. & Okoniewska, M., *et al*. (1994). Redesigning the active-site of *Geotrichum-candidum* lipase. *Protein, Eng.* **8**, 835-842.

Simpson, A., Bateman, O., Driessen, H., Lindley, P., Moss, D. & Mylvaganam, S. *et al*. (1994). The structure of avian eye lens δ-crystallin reveals a new fold for a superfamily of oligomeric enzymes. *Nature Struct. Biol.* **1**, 724-734.

Taylor, W. R., . & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 208-229.

Terwisscha van Scheltinga, A. C. T., Armand, S., Kalk, K. H., Isogai, A., Henrissat, B. & Dijkstra, B. W. (1995). Stereochemistry of chitin hydrolysis by a plant chitinase lysozyme and X-ray structure of a complex with allosamidin - evidence for substrate assisted catalysis. *Biochemistry,* **34**, 15619-15623.

Terwisscha van Scheltinga, A. C., Henrig, M. & Dijkstra, B. W. (1996). The 1.8 Angstrom resolution structure of hevamine, a plant chitinase/lysozyme, and analysis of the conserved sequence and structure motifs of glycosyl hydrolase family 18. *J. Mol. Biol.* **262**, 243-257.

Thayer, M. M., Ahern, H., Xing, D., Cunningham, R. P. & Tainer, J. A. (1995). Novel DNA binding motifs in the DNA repair enzyme endonuclease III crystal structure. *EMBO J.* **14**, 4108-4120.

Thoden, J. B., Holden, H. M., Wesenberg, G., Raushel, F. M. & Rayment, J. (1997). Structure of carbamoyl phosphate synthase: a journey of 96 Å from substrate to product. *Biochemistry,* **36**, 6305-6316.

Thorn, J. M., Barton, J. D., Dixon, N. E., Ollis, D. L. & Edwards, K. J. (1995). Crystal-structure of *Escherichia coli* QOR quinone oxidoreductase complexed with NADPH. *J. Mol. Biol.* **249**, 785-799.

Van Driessche, G., Koh, M., Chen, Z. W., Mathews, F. S., Meyer, T. E. & Bartsch, R. G. *et al*. (1996). Covalent structure of the flavoprotein subunit of the flavocytochrome *c*: sulfide dehydrogenase from the purple phototrophic bacterium *Chromatium vinosum*. *Protein. Sci.* **5**, 1753-1764.

Van Roey, P., Rao, V., Plummer, T. H. & Tarentino, A. L. (1994). Crystal-structure of endo-β-*N*-acetylglucosarninidase F1, an α/β-barrel enzyme adapted for a complex substrate. *Biochemistry,* **33**, 13989-13996.

Volz, K. (1999). A test case for structure-based functional assignment: the 1.2 Angstrom crystal structure of the yjgF gene product from *Escherichia coli*. *Protein, Sci.* **8**, 2428-2437.

Waldrop, G. L., Rayment, I. & Holden, H. M. (1994). Three-dimensional structure of the biotin carboxylase subunit of acetyl-CoA carboxylase. *Biochemistry,* **33**, 10249-10256.

Wallace, A. C., Laskowski, R. A. & Thornton, J. M. (1996). Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. *Protein Sci.* **5**, 1001-1013.

Wang, S., Trumble, W. R., Liao, H., Wesson, C. R., Dunker, A. K. & Kang, C. (1998). Crystal structure of calsequestrin from rabbit skeletal muscle sarcoplasmic reticulum. *Nature Struct. Biol.* **5**, 476-482.

Webb, E. C. (1992). *Enzyme Nomenclature 1992.* Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, Academic Press, New York.

Wetterau, J. R., Combs, K. A., Spinner, S. N. & Joiner, B. J. (1990). Protein disulphide isomerase is a component of the microsomal triglyceride transfer protein complex. *J. Biol. Chem.* **265**, 9800-9807.

Wilmanns, M., Priestle, J. P., Niermann, T. & Jansonius, J. N. (1992). Three-dimensional structure of the bifunctional enzyme phosphoribosylanthranilate isomerase: indoleglycerolphosphate synthase from *Escherichia coli* refined at 2.0 Å resolution. *J. Mol. Biol.* **223**, 477-507.

Wilson, C. A., Kreychman, J. & Gerstein, M. (2000). Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J. Mol. Biol.* **297**, 233 249.

Wistow, G. & Piatigorsky, J. (1987). Recruitment of enzymes as lens structural proteins. *Science,* **236**, 1554-1556.

Woods, S. A., Miles, J. S., Roberts, R. E. & Guest, J. R. (1986). Structural and functional-relationships between fumarase and aspartase - nucleotide-sequences of the fumarase (fumC) and aspartase (aspA) genes of *Escherichia coli*-K12. *Biochem. J.* **237**, 547-557.

Wu, G., Fiser, A., Terkuile, B., Sali, A. & Muller, M. (1999). Convergent evolution of *Trichomonas vaginalis* lactate dehydrogenase from malate dehydrogenase. *Proc. Natl Acad. Sci. USA,* **96**, 6285-6290.

Wyckoff, T. J. O. & Raetz, C. R. H. (1999). The active site of *Escherichia coli* UDP-*N*-acetylglucosamine acyltransferase - chemical modification and site-directed mutagenesis. *J. Biol. Chem.* **274**, 27047-27055.

Yang, F., Gustafson, K. R., Boyd, M. R. & Wlodawer, A. (1998). Crystal structure of *Escherichia coli* HdeA. *Nature Struct. Biol.* **5**, 763-764.

Zarembinski, T. I., Hung, L. W., Muellerdieckmann, H. J., Kim, K. K., Yokota, H., Kim, R. & Kim, S. H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl Acad. Sci. USA,* **95**, 15189-15193.

Zhang, B. H., Rychlewski, L., Pawlowski, K., Fetrow, J. S., Skolnick, J. & Godzik, A. (1999). From fold predictions to function predictions: automation of functional site conservation analysis for functional genome predictions. *Protein Sci.* **8**, 1104-1115.

Edited by A. R. Fersht

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