doi:10.1016/S0022-2836(02)00649-6 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2002) **321**, 741–765

JMB

Review

One Fold with Many Functions: The Evolutionary Relationships between TIM Barrel Families Based on their Sequences, Structures and Functions

Nozomi Nagano[1,2]*, **Christine A. Orengo**[1] **and Janet M. Thornton**[1,3]

[1]*Biomolecular Structure and Modelling Group, Department of Biochemistry and Molecular Biology, University College London, Gower Street, London WC1E 6BT, UK*

[2]*Molecular Informatics Team Computational Biology Research Center (CBRC) National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064 Japan*

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

**Corresponding author*

The eightfold (βα) barrel structure, first observed in triose-phosphate isomerase, occurs ubiquitously in nature. It is nearly always an enzyme and most often involved in molecular or energy metabolism within the cell. In this review we bring together data on the sequence, structure and function of the proteins known to adopt this fold. We highlight the sequence and functional diversity in the 21 homologous superfamilies, which include 76 different sequence families. In many structures, the barrels are "mixed and matched" with other domains generating additional variety. Global and local structure-based alignments are used to explore the distribution of the associated functional residues on this common structural scaffold. Many of the substrates/co-factors include a phosphate moiety, which is usually but not always bound towards the C-terminal end of the sequence. Some, but not all, of these structures, exhibit a structurally conserved "phosphate binding motif". In contrast metal-ligating residues and catalytic residues are distributed along the sequence. However, we also found striking structural superposition of some of these residues. Lastly we consider the possible evolutionary relationships between these proteins, whose sequences are so diverse that even the most powerful approaches find few relationships, yet whose active sites all cluster at one end of the barrel. This extreme example of the "one fold-many functions" paradigm illustrates the difficulty of assigning function through a structural genomics approach for some folds.

Keywords: TIM barrel; evolution; phosphate-binding motif; functional residues

Introduction

TIM barrel proteins, which have eight β/α motifs folded into a barrel structure, have been widely analysed considering structure, function, folding, and evolution.[1–5] In a recent version of the CATH classification,[6] where structures are grouped by class, architecture, topology and homologous (CATH) family, there are nearly 900 TIM barrel structures. Most TIM barrels function as enzymes, including five of the six primary classes of enzymes, as defined by the enzyme commission (EC).[4,7]

There has been speculation as to whether the TIM barrel structures seen today result from convergent evolution to a stable fold or divergent

Present addresses: N. Nagano, Molecular Informatics Team, Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-41-6 Aomi, Koto-ku, Tokyo 135-0064, Japan; J. Thornton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge CB10 1SD, UK.

Abbreviations used: 7CEL, seven-stranded cellulase; AAMY, α-amylase subgroup; ALD1, aldolase class I family; ALD2, aldolase class II; ALR, alanine racemase; CHTN, chitinase subgroup; CHOB, chitobiase; DHPS, dihydropteroate synthetase; EC, enzyme commission; EG, endoglucanase; ENOL, enolase superfamily; FMN, flavin mononucleotide; FMOP, FMN-dependent oxido-reductase and phosphate binding enzymes; GLYC, eight-stranded glycosidase family; LUCL, luciferase-like pro-teins; MHYD, metal-dependent hydrolases; NADO, NADP-dependent oxidoreductase; NADP, nicotinamide-adenine-dinucleotide phosphate; PDB, Protein Data Bank; PEPE, phosphoenolpyruvate (PEP) binding enzymes; PIPLC, phosphatidylinositol (PI) phospholipase C; PLP, pyridoxal-5′-phosphate; QAPRT, quinolinic acid phospho-ribosyl (QAPR) transferase; PP, phosphate; RUB, ribulose 1,5-bisphosphate carboxylase/oxygenase (rubisco); SPB, standard phosphate binding; TIM, triose phosphate iso-merase; TRGT, tRNA-guanine transglycosylase; XYLL, xylose-isomerase-like enzymes.

E-mail address of the corresponding author: n.nagano@aist.go.jp

Table 1. List of proteins and PDB codes

Family number	Family name (abbreviation)	Number of sequence families	Name of proteins (PDB codes[a] and chain of representative proteins for analysis; EC numbers)	No. of TIM barrels		
				Non-reduntant[b]	All folds[c]	Entries[d]
1	Alanine racemase (ALR)	1	Alanine racemase (1bd0A; EC 5.1.1.1)	2	6	3
2	Dihydropteroate (DHP) synthetase (DHPS)	1	DHP synthetase (1ad4B; EC 2.5.1.15)	2	7	5
3	FMN dependent fluorescent proteins (Luciferase-like proteins) (LUCL)	3	Flavoprotein 390 (1fvpA) Luciferase chain B (1lucB; EC 1.14.14.3) Luciferase chain A (1lucA; EC 1.14.14.3)	4	11	6
4	Seven-stranded glycosidases (cellulases) (7CEL)	2	Endocellulase (1tml; EC 3.2.1.4) Cellobiohydrolase II (1cb2A; EC 3.2.1.91)	3	12	7
5	Phoshoenolpyruvate (PEP) binding enzymes (PEPE)	2	Pyruvate kinase (1pkm; EC 2.7.1.40) Pyruvate phosphate dikinase (1dik04; EC 2.7.9.1)	5	45	12
6	Aldolase class I family (ALD1)	4	*N*-Acetylneuraminate lyase (1nal1; EC 4.1.3.3) Dihydrodipicolinate synthase (1dhpA; EC 4.2.1.52) Fructose-1,6-bisphosphate aldolase (1fbaA; EC 4.1.2.13) Transaldolase B (1onrA; EC 2.2.1.2)	6	35	13
7	Glycosidases (GLYC)	30		53	226	160
	7-1 α-Amylase (AAMY)	(10)	α-Amylase (1ava, 2aaa, 1ppi, 1bli, 1bag; EC 3.2.1.1), oligo-1,6-glucosidase (1uok; EC 3.2.1.10), 1,4-α-D-Glucan maltotetrahydrolase (1amg; EC 3.2.1.60), isoamylase (1bf2; EC 3.2.1.68), α-Amylase II (1bvzA; EC 3.2.1.135), Cyclodextrin glycosyltransferase (1cgt; EC 2.4.1.19)			
	7-2 Endoglucanase (EG)	(14)	β-Amylase (1byb, 1b9z; EC 3.2.1.2), cellulase (1ceo, 1eceA; EC 3.2.1.4), Endoglucanase (1edg, 1egzA; EC 3.2.1.4), Xylanase (1xyzA; EC 3.2.1.8), β-Glycosidase (1gowA; EC 3.2.1.23), β-Galactosidase (1bglA; EC 3.2.1.23), β-Glucuronidase (1bhgA; EC 3.2.1.31), β-Mannanase (1bqcA; EC 3.2.1.78), 6-Phospho-β-D-galactosidase (1pbgA; EC 3.2.1.85), β-1,4-glycanase (2xyl; EC 3.2.1.91), 1,3-β-Glucanase (1ghsA; EC 3.2.1.39), 1,3-1,4-β-Glucanase (1aq0A; EC 3.2.1.73), β-Glucosidase (1cbg; EC 3.2.1.21), myrosinase (2myr; EC 3.2.3.1)			
	7-3 Chitinase (CHTN)	(5)	Chitinase A (1ctn; EC 3.2.1.14), hevamine (2hvm; EC 3.2.1.14 + 3.2.1.17), *Endo*-β-*N*-acetylglucosaminidase (2ebn, 1edt; EC 3.2.1.96), narbonin (1nar)			
	7-4 Chitobiase (CHOB)	(1)	Chitobiase (1qba; EC 3.2.1.52)			
8	Triose phosphate isomerase (TIM) (TIM)	1	TIM (1tpfA; EC 5.3.1.1)	11	91	39
9	NADP-dependent oxidoreductase (NADO)	2	Aldose reductase (1ads; EC 1.1.1.21), [3-α-hydroxysteroid dehydrogenase (1lwiA; EC 1.1.1.50) and Aldehyde reductase (2alr; EC 1.1.1.2)] K + channel (1qrqA)	8	25	20
10	tRNA-guanine (tRNA-G) transglycosylase (TRGT)	1	tRNA-guanine transglycosylase (1wkf; EC 2.4.2.29)	1	6	6
11	Rubisco (RUB)	1	Ribulose 1,5-bisphosphate carboxylase/oxygenase (1rblA; EC 4.1.1.39)	5	67	23
12	Enolase superfamily (ENOL)	4	Enolase (1oneA; EC 4.2.1.11) Mandelate racemase (1mdl; EC 5.1.2.2)	6	36	26

(*continued*)

Table 1 Continued

Family number	Family name (abbreviation)	Number of sequence families	Name of proteins (PDB codes[a] and chain of representative proteins for analysis; EC numbers)	No. of TIM barrels		
				Non-reduntant[b]	All folds[c]	Entries[d]
			Muconate lactonizing enzyme (1mucA; EC 5.5.1.1), (Chloromuconate cycloisomerase (1chrA; EC 5.5.1.7)) D-gucarate dehydratase (1bqg; EC 4.2.1.40)			
13	FMN-dependent oxidoreductase and phosphate (PP) binding enzymes (FMOP)	12	Flavocytochrome *b*2 (1fcbA; EC 1.1.2.3), glycolate oxidase (1gox; EC 1.1.3.15) Trimethylamine dehydrogenase (2tmdA; EC 1.5.99.7) Old yellow enzyme (1oya; EC 1.6.99.1) Dihydroorotate dehydrogenase (1dorA; EC 1.3.3.1) Inosine monophosphate dehydrogenase (1ak5; EC 1.1.1.205) Thiamin phosphate synthase (2tpsA; EC 2.5.1.3) Ribulose-phosphate 3-epimerase (1rpxA; EC 5.1.3.1) Tryptophan synthase (1ubsA; EC 4.2.1.20) Indole-3-glycerol-phosphate synthase (1pii *N*-terminal domain; EC 4.1.1.48) Indole-3-glycerolphosphate synthase (1a53; EC 4.1.1.48) *N*-(5′phosphoribosyl)anthranilate isomerase (1pii C-terminal domain; EC 5.3.1.24) Phosphoribosyl anthranilate isomerase (1nsj; EC 5.3.1.24)	17	71	51
14	Metal-dependent hydrolases (MHYD)	4	Adenosine deaminase (1a4mA; EC 3.5.4.4) Urease (2kauC; EC 3.5.1.5) Phosphotriesterase (1pscA; EC 3.1.8.1) Phosphotriesterase (1bf6A)	5	47	38
15	Divalent-metal-dependent enzymes (xylose isomerase-like proteins) (XYLL)	3	D-Xylose isomerase (1xib; EC 5.3.1.5) Xylose isomerase (1a0dA, EC 5.3.1.5) Endonuclease IV (1qtwA; EC 3.1.21.2)	11	141	64
16	Aldolase class II (ALD2)	1	Fructose-bisphosphate aldolase II (1b57A; EC 4.1.2.13)	3	5	3
17	Phosphatidylinositol (PI) phospholipase C (PIPLC)	3	Phosphatidylinositol-specific phospholipase C (1gym; EC 3.1.4.10) Phospholipase C δ-1 (1qasA; EC 3.1.4.11) Phosphatidylinositol-specific phospholipase C (2plc; EC 3.1.4.10)	3	32	22
18	Quinolinic acid phosphoribosyl(QAPR) transferase (QAPRT)	1	Quinolinic acid phosphoribosyl-transferase (1qpoA; EC 2.4.2.19)	2	26	5
	Total	76		147	889	503

The additional proteins, whose functions are different from the representative in the same sequence family, are in parentheses.
[a] The references are shown in each PDB file.
[b] The number of non-redundant TIM barrel sequences.
[c] The number of all TIM barrel folds in PDB entries.
[d] The number of PDB entries including TIM barrel folds.

evolution from a common ancestor. Notwithstanding the diversity of their catalytic reactions, the active site is always found at the C-terminal end of the barrel sheets,[8] suggesting divergence from an ancestral TIM barrel. These enzymes often have additional domains that precede, interrupt, or follow the barrel, which indicates domain-shuffling events during protein evolution. Ten years ago, Farber & Petsko classified 17 TIM barrel structures into four families on the basis of their different geometries, and suggested divergence from a common ancestor.[1] Bränden also analysed 19 TIM

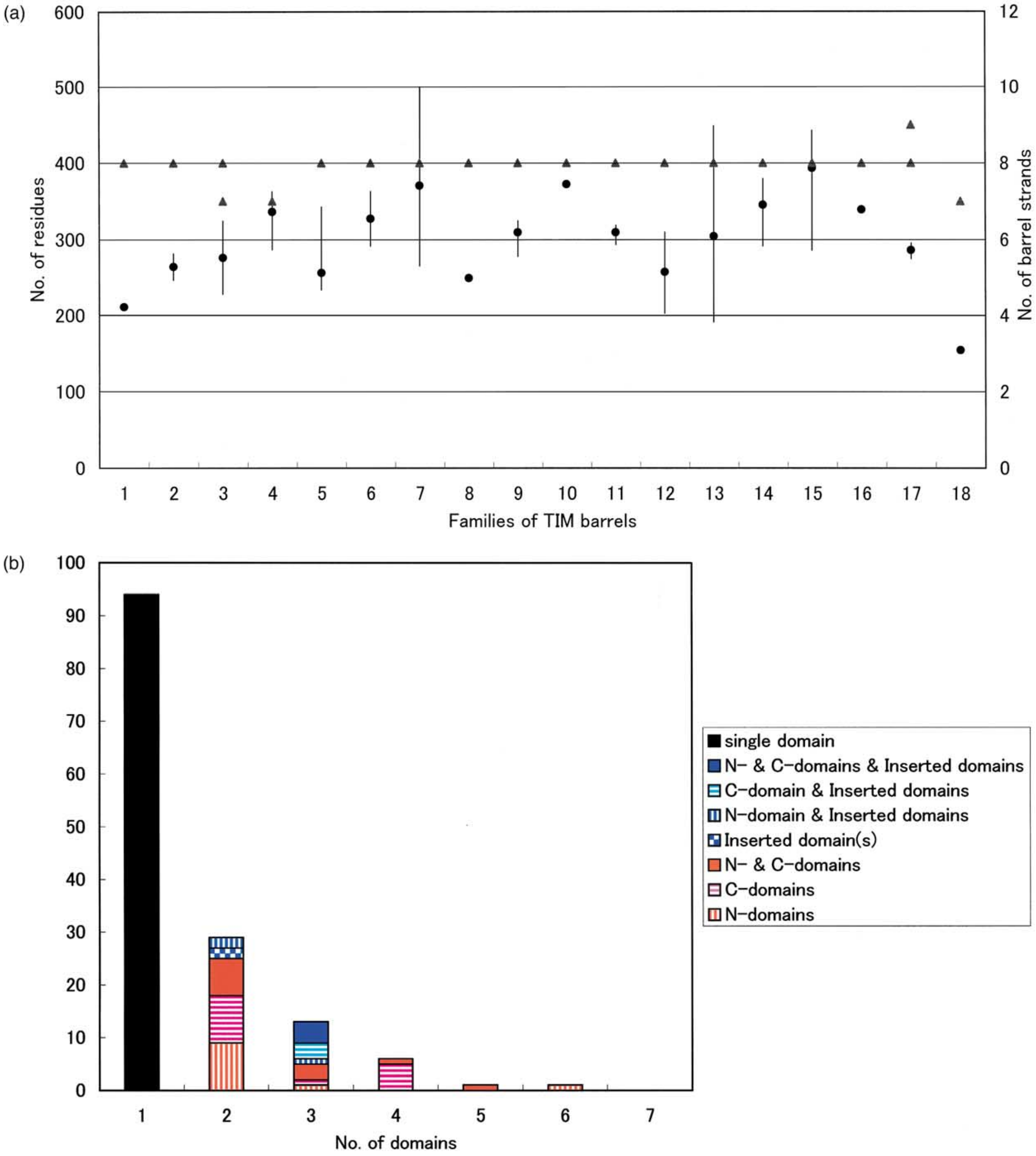

Figure 1 (*legend opposite*)

barrel structures, considering their domain organization, the metal and phosphate binding sites as well as catalytic centres.[2] It was suggested that the presence of the common phosphate-binding site, formed by loop-7, loop-8 and a small helix (helix-8′), is the strongest evidence obtained for the divergent evolutionary history of TIM barrels.[2,3] More recently, using 30 TIM barrel structures and their sequence families, the evolution of TIM barrels was discussed by Reardon & Farber.[9] Although it was still not clear, they concluded that divergent evolution from a common ancestor explained more of the available data than did convergent evolution. More recently Copley & Bork re-analysed a subset of the TIM barrels, including nine of the 21 families considered herein (i.e. those which bind phosphate).[10] They suggest that five of these families (two classes of aldolase; dihydropteroate synthetase; pyruvate kinase and enolase) are distantly related to the enzymes with a common phosphate-binding motif.

Intriguingly the tertiary structures of two enzymes involved in histidine biosynthesis, HisA and HisF, have recently been reported and their

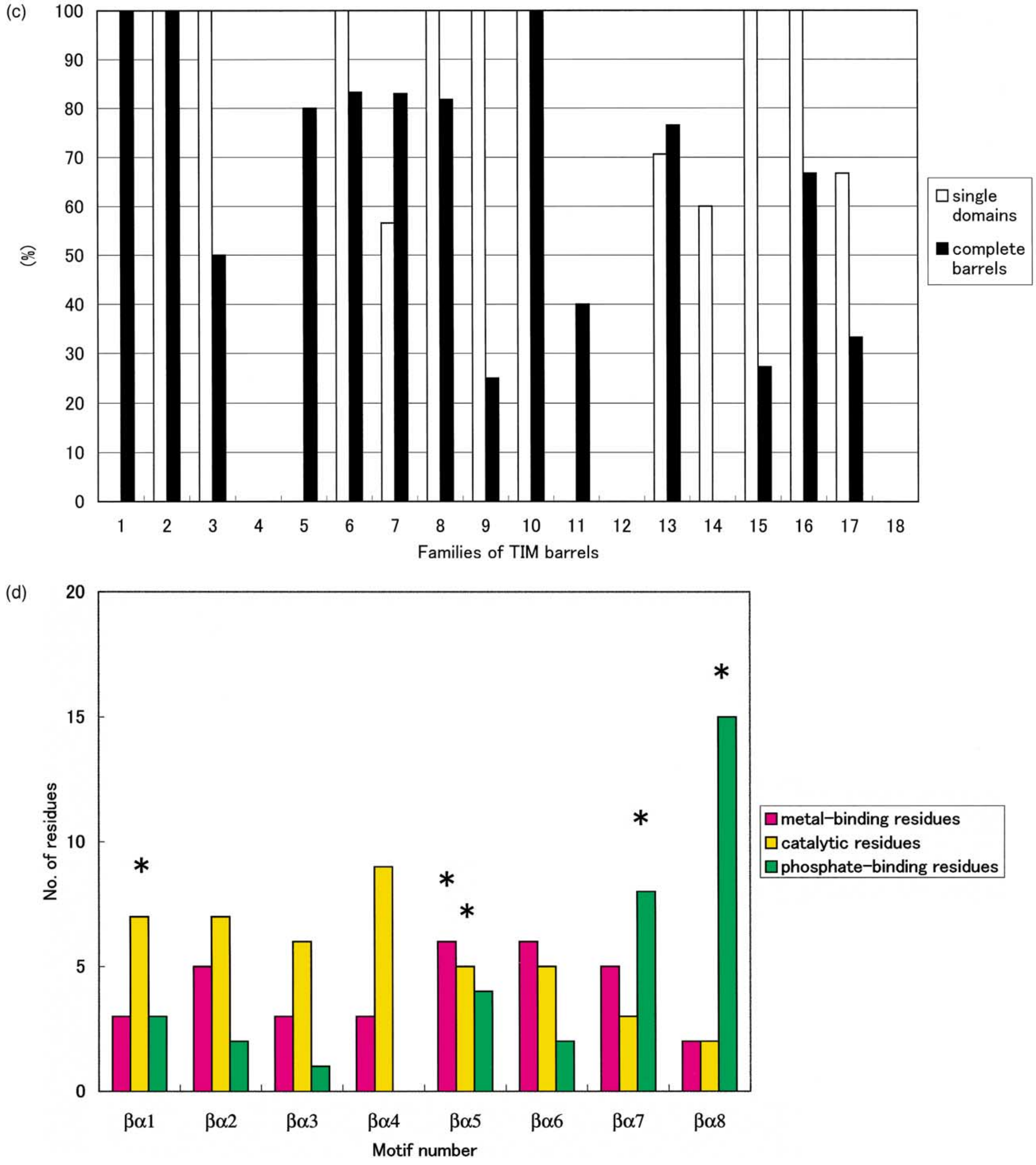

Figure 1. The overview of TIM barrels. (a) The size and the number of barrel strands of TIM barrel domains. The range of the size in residues is indicated by a line, and the average size is shown by a black circular dot (left axis), whilst the number of strands is indicated with a triangular dot (right axis). The domain definition is on the basis of the CATH classification.[39,40] (b) The domain composition of TIM barrels. Single domains are indicated by a black bar, whilst multi-domains are classified and indicated in the following ways. Those domains with inserted domains in the middle of the TIM barrel domain are coloured in blue or cyan, whilst those without such insertions are in red or magenta. Those with additional domains at both the N and C termini of TIM barrels are indicated with plain coloured sectors. Those with domains at the N termini, but not at the C termini of TIM barrel domains are indicated with vertically striped sectors, whilst those with domains at the C termini, but not at the N termini are with horizontally striped ones. Those with only the inserted domains are indicated with blue check sectors. (c) Topologies of TIM barrel families. The percentage of single domain structures in each family is indicated with open bars, whilst that of complete barrels is indicated with black bars. (d) The number of active site residues at the eight βα motifs. Magenta, yellow, and green bars indicate metal-ligating residues, catalytic residues, and phosphate-binding residues, respectively. The active site residues, which could be aligned using multiple structure alignment, are indicated with asterisks (*).

sequence and structure data suggest that they have evolved by twofold gene duplication of a half-barrel ancestor, followed by gene fusion.[5,11] Each half-barrel of these enzymes contains a phosphate-binding motif that is required by the nature of their biphosphate substrate. This motif is found in the C-terminal halves of a number of other TIM barrel enzymes, but not in the N-terminal halves. This could indicate that a half-barrel phosphate-binding precursor might be a common ancestor of all TIM barrels, and that some proteins have diverged so far that all phosphate-binding features have been lost and no similar features can be detected.[5]

In our previous work,[12] the TIM barrel glycosidase superfamily, which is the biggest family in the TIM barrel fold, was analysed and could be clustered systematically, by combining various methods such as PSI-BLAST, structure-comparison methods, and functional analyses. This combined method can be applied to fold-level clustering as well as to superfamily level clustering. In this review, the structure, function and evolution of the TIM barrel enzymes is analysed comprehensively and systematically, using the combined method. With almost 900 such structures in the Protein Data Bank (PDB), we revisit their relationships using the more sophisticated sequence and structure analysis tools now available.

Results

Family size and geometry of TIM barrel enzymes

In CATH version 1.7, there are 889 TIM barrel folds in 503 PDB entries, which are split into 18 homologous families (H-level). Each family (e.g. alanine racemase) is given an abbreviation (ALR) and a family number (F1), which will be used throughout the paper (see Table 1). These include 76 sequence families (S-level in CATH) covering 147 non-identical sequences. Seven of the 18 homologous families consist of only one sequence family (see Table 1). The biggest CATH homologous family is the eight-stranded glycosidase family (GLYC; F7) with 30 sequence families and 22 EC numbers, most of which are 3.2.1.–. In CATH these were clustered together, since they had the same fold and a very similar function. However, in our previous paper,[12] evidence for clustering all these proteins together was limited, so that the GLYC proteins (F7) were divided into four homologous families, α-amylase (AAMY), endoglucanase (EG), chitinase (CHIN), and chitobiase (CHOB). Here these are defined as subgroups, F7-1, F7-2, F7-3, and F7-4, respectively. Dividing this large group of proteins into four families gives 21 homologous families to be considered. The second largest family is the FMN-dependent oxidoreductase and phosphate binding enzymes (FMOP; F13), which covers 12 sequence families

Most of the 21 families comprise only eight-stranded barrels (see Figure 1(a)). However, there are two seven-stranded TIM barrel families: cellulase (7CEL; F4) and quinolinic acid phosphoribosyltransferase (QAPRT; F18). Another family contains one member with a seven-stranded barrel (flavoprotein 390 in FMN-dependent fluorescent proteins (LUCL; F3)) (see Figure 1(a)) and one family PIPLC (F17) has one member with a nine-stranded barrel.

The domain composition was also analysed (Figure 1(b) and (c)). Out of the 147 non-identical TIM barrels, 96 proteins comprise a single domain in the PDB, whilst the remainder are multi-domain, with up to six component domains (Figure 1(b)). In 12 of the 51 multi-domain proteins, one or more domains are inserted within the TIM barrel (Figure 1(b)), e.g. pyruvate kinase (PEPE; F5) and α-amylase (AAMY; F7-1) (see Figure 3). Cross checking against SWISS-PROT[13] and Pfam,[14] we find in addition that 7CEL (F4) has an additional domain, not present in the crystal structure.[15,16] The size of the TIM barrel domains, excluding domain insertions, varies from 150 to 500 residues with an average 298 (see Figure 1(a)).

In six families (ALR (F1), 7CEL (F4), PEP-binding enzymes (PEPE; F5), RUB (F11), enolase superfamily (ENOL; F12), and QAPRT (F18)), all the members have more than one domain, whilst in the remaining families at least some relatives have single domain structures (see Figure 1(c)). The completeness of the barrel was also analysed using our previous method of identifying complete TIM barrel structures.[4] Most families include members with distorted barrels (Figure 1(c)). In some families (7CEL (F4), ENOL (F12), metal-dependent hydrolase (MHYD; F14), and QAPRT (F18)), all proteins have distorted barrel structures, where hydrogen bonds are lost between at least a pair of adjacent barrel strands.

Thus the TIM barrel structures are themselves geometrically diverse and are used in many different combinations with other domains to create functional proteins.

Functions of TIM barrels

Chemical reactions (EC classes) performed by TIM barrels

The distinct functions of all members of each family in the PDB are given in Table 2 and Figure 3. In the 21 homologous families in CATH, there are 61 different EC numbers, which cover primary EC classes 1–5 (Table 2 and Figure 2(a)). Of these, three EC classes (EC 3.2.1.4, 3.2.1.91, 4.1.2.13) occur in two different families (i.e. the same reaction is catalysed by sequentially unrelated proteins), giving a total of 64 EC numbers, as shown in Table 2. In five families (7CEL (F4), PEPE (F5), NADO (F9), MHYD (F14) and PIPLC (F17)), all members belong to the same primary EC class

Table 2. Overview of TIM barrel families

Family name (abbreviation)	No. of EC numbers	EC numbers[a] and types of proteins	Cofactors	Phosphate moiety of ligands	Phosphate[b] binding site
1. Alanine racemase (ALR)	1	5.1.1.1	PLP covalently bound	PLP	PP → L7, L8
2. DHP synthetase (DHPS)	1	2.5.1.15		Substrate	β-PP → β8, L8[c]
3. FMN fluorescent proteins (LUCL)	1	1.14.14.3, Flavoprotein 390	FMA or (FMNH)	FMN	PP → L4, L5
4. Seven-stranded cellulase (7CEL)	2	3.2.1.4, 3.2.1.91			
5. PEP-binding enzyme (PEPE)	2	2.7.1.40, 2.7.9.1	Mg^{2+} or Mn^{2+} (& K^+)	Substrates	3-PP → β2, L2, L3[d]
6. Aldolase class I (ALD1)	4	4.1.2.13, 4.1.3.3, 4.2.1.52, 2.2.1.2		Substrate for (*trans*)aldolase	PP-1 → L7, L8[e]
7. Glycosidases (GLYC)	22	2.4.1.19, 3.2.3.1, 3.2.1.X, (X = 1, 2, 4, 8, 10, 14, 14 + 17, 21, 23, 31 39, 52, 60, 68, 73, 78, 85, 91, 96, 135) Narbonin, concanavalin B		Substrate for 6-Phospho-β-galactosidase (EC 3.2.1.85)	L8
8. TIM (TIM)	1	5.3.1.1		Substrate	✓
9. NADP-oxidoreductase (NADO)	3	1.1.1.2, 1.1.1.21, 1.1.1.50, K^+ channel	NADP	NADP	diPP → L7, PP → L8[c]
10. tRNA-G transglycosylase (TRGT)	1	2.4.2.29		Substrate	Not determined
11. Rubisco (RUB)	1	4.1.1.39	Mg^{2+}	Substrate	✓
12. Enolase superfamily (ENOL)	5	5.1.2.2, 5.5.1.1, 5.5.1.7, 4.2.1.11, 4.2.1.40	Mg^{2+} or Mn^{2+}	Substrate for enolase	PP → L1, L7
13. FMN-oxidoreductase and PP-binding enzymes (FMOP)	11	1.1.2.3, 1.1.3.15, 1.3.3.1, 1.5.99.7, 1.6.99.1, 1.1.1.205, 2.5.1.3, 4.1.1.48, 4.2.1.20, 5.1.3.1, 5.3.1.24	FMN (and heme or 4Fe–4S) NADP, PLP, or no cofactor	FMN or substrates	✓
14. Metal-hydrolase (MHYD)	3	3.1.8.1, 3.5.1.5, 3.5.4.4	Zn^{2+} or $Zn^{2+} \times 2$ or $Ni^{2+} \times 2$	Substrate for Phosphotriesterase	PP → L3, β5[f]
15. Divalent-metal-enzymes (XYLL)	2	3.1.21.2, 5.3.1.5	$Mg^{2+} \times 2$ or $Zn^{2+} \times 3$		
16. Aldolase class II (ALD2)	1	4.1.2.13	$Zn^{2+} \times 2$	Substrate	PP → L7, L8[c]
17. PI phospholipase C (PIPLC)	2	3.1.4.10, 3.1.4.11	Ca^{2+}	Substrate	Not determined
18. QAPR transferase (QAPRT)	1	2.4.2.19	$Mg^{2+} \times 2$	Substrate	
Total	64				

[a] EC numbers that exist in two different families are underlined.
[b] "Standard phosphate binding (SPB) motif": the common phosphate-binding site, ranging from β-7, loop-7, α-7, β-8 to a following small helix, helix-8′. ✓ indicates the existence of such a motif. Loops, β-strands, α-helices of barrels are indicated as L, β, and α, respectively. L1 indicates the binding site on the loop-1 between β-1 and α-1 of barrel structure.
[c] There is not a small helix, helix-8′, between barrel strand 8 (β8) and helix8 (α8) with only one residue inbetween, as well as phosphate binding to the same position as in footnote b.
[d] There is a small helix, in pyruvate kinase, between β8 and α8 with only one residue inbetween, although triphosphate of ATP binds to different site.
[e] In aldolase, one of two phosphate groups in substrate binds to the same position as in footnote b.
[f] No small helix between β8 and α8, and phosphate binds to different site.

(Table 2), whilst others apparently perform multiple different reactions. For example, the large FMOP family (F13) covers primary EC numbers 1, 2, 4, and 5 (Table 2). This analysis only includes proteins found in the PDB. As we have shown,[17] including sequence relatives will often more than double the number of functions associated with a given family.

For comparison to other enzymes, the distribution of chemical functions performed by TIM barrels was compared with all enzymes in *Escherichia coli*. The PEDANT list of 689 *E. coli* genes,[18] with EC numbers and functions assigned automatically, is shown in Figure 2(d). Comparison with Figure 2(a) indicates a dominance of hydrolases (especially glycosidases) and a lack of ligases in the TIM barrel enzymes. Apparently there are many more oxidoreductases and transferases with different folds in *E. coli.*

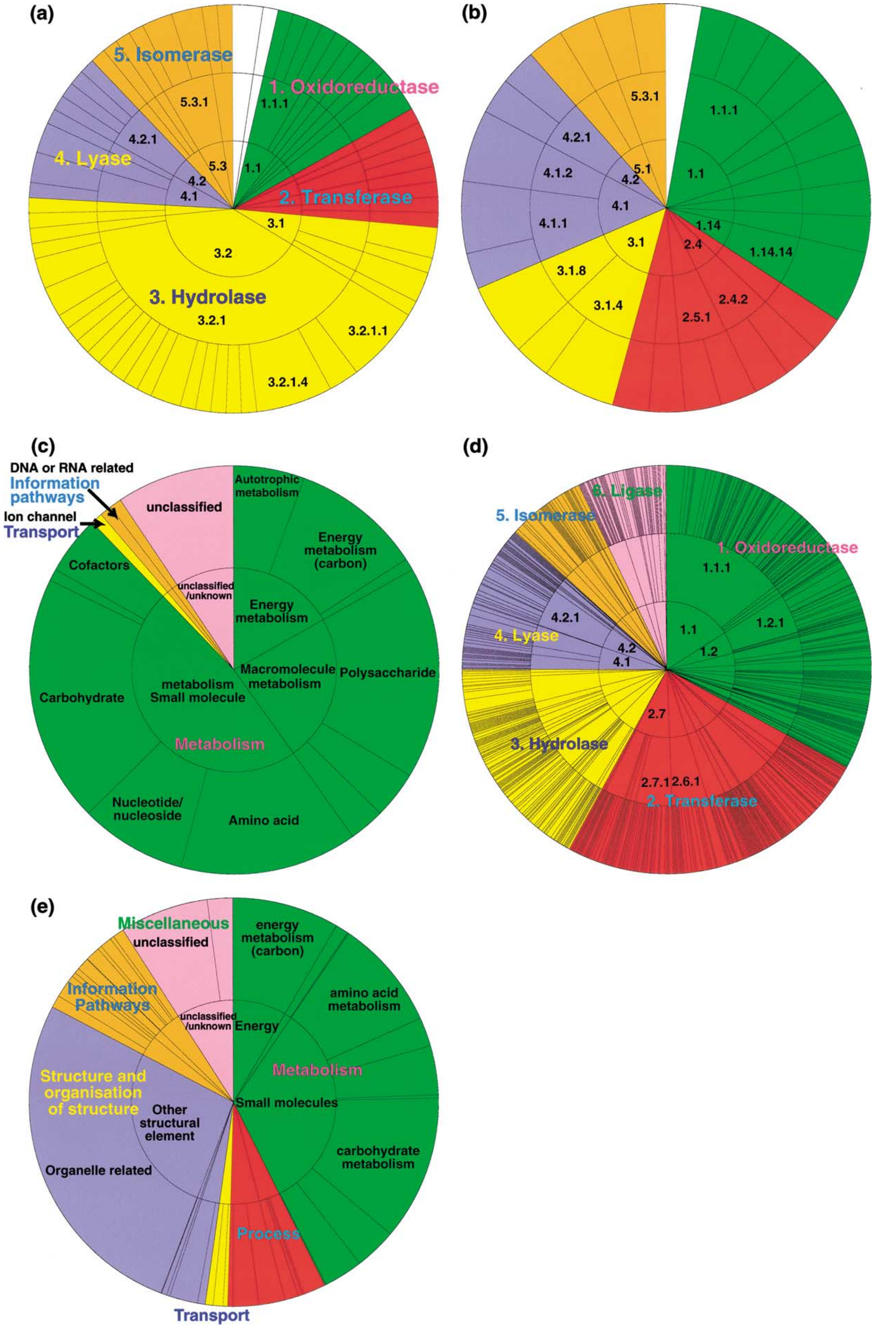

Figure 2 (*legend opposite*)

Biological roles

The biological roles of TIM barrel enzymes in the PDB are shown in Figure 2(c). Some proteins such as luciferase and even some glycosidases are not classified in GenProtEC[19] and KEGG[20] databases, and therefore their biological functions could not be annotated. Figure 2(c) emphasizes that of the 64 enzyme reactions performed by TIM barrels (EC numbers) 85% are involved in energy metabolism, macromolecule metabolism, or small molecule metabolism, as defined by Rison *et al*.[21] The only exceptions are the ion channel transport protein (K^+ channel) (NADO; F9), a DNA-related information pathway protein (endonuclease IV) (XYLL; F15) and an RNA-related information pathway protein (tRNA-guanine transglycosylase) (TRGT; F10). As above these data refer only to proteins of known structures. Expanding these families with sequence relatives will considerably expand the functional repertoire.

As for EC numbers, the distribution of biological functions of TIM barrels were also compared to that of *E. coli* genes deposited in PEDANT[18] (Figure 2(e)) using the scheme proposed by Rison *et al*.[21] The distribution for TIM barrel folds is very different and highlights the importance of the TIM barrel fold for all sorts of metabolism.

Active site: catalytic residues

Active site residues are shown in Figure 3. In two families (ALR (F1) and RUB (F11)), the same type of catalytic residue (carbamated lysine or lysine N^{ζ}-carboxylic acid) is used despite performing different functions (racemase and carbon–carbon lyase, respectively) and lying in a different position (C termini of β-5 and β-2, respectively) (see Figure 3). A carbamated lysine is also used by MHYD (F14) to bind metals. A Schiff-base (a modified lysine) is also important for catalysis in two families (ALR (F1) and ALD1 (F6)). There are five families in which acidic residues or basic residues are involved in catalysis as either acid or base (7CEL (F4), TIM (F8), TRGT (F10), GLYC (F7), and PIPLC (F17)), although they are not positionally equivalent. Inspection of Figure 3 shows that, although the catalytic sites are all located at the C-terminal end of the barrel, the catalytic residues derive from different locations along the sequence. A summary of the location of the active site residues is given in Figure 1(d).

Cofactors to assist catalysis

Twelve of the 18 families bind a cofactor such as FMN, NADP, PLP or divalent metals (Table 2). All the cofactors and substrates are bound to the C termini of the β-strands in all the families (see Figure 3). There are two families, which utilize FMN or flavin (LUCL (F3) and FMOP (F13)), and a family NADO (F9) that uses NADP for catalysing oxido-reductions.

Eight families bind divalent metals (see Table 2 and Figure 3). QAPRT (F18) binds two metal ions through the substrate and a water molecule.[22] For the remainder, the metal ions are bound directly to acidic or basic residues. ENOL (F12) and PEPE(F5) bind either a Mg^{2+} or Mn^{2+}, although PEPE has only two acidic residues and ENOL has three (see Figure 3). The XYLL family (F15) includes two different functions, xylose isomerase (EC 5.3.1.5) and endonuclease IV (EC 3.1.21.2), but both enzymes bind at least two divalent metal ions at the same site (see Figure 3). RUB (F11) uses a Mg ion to assist catalysis and it is bound by a carbamated lysine and two acidic residues (see Figure 3).

The metal hydrolase family, MHYD (F14), catalyses the hydrolytic cleavage of amide, or ester bonds, and structures have been determined for phosphotriesterase (EC 3.1.8.1), urease (EC 3.5.1.5) and adenine deaminase (EC 3.5.4.4), which all hydrolyse ester bonds. These enzymes and their homologues were analysed by Holm & Sander in detail.[23] The members of this family employ a variety of divalent metal ligands for catalysis. Phosphotriesterase and urease contain two zinc ions and two nickel ions, respectively, in the binuclear metal centre where a carbamated lysine acts as bridging ligand, and four histidine residues are involved in binding the metals (Figure 3). In contrast, adenine deaminase binds only a zinc ion in the co-located metal site.

Other metal-binding families (PEPE (F5), ENOL (F12) and XYLL (F15)) all bind divalent metal ions (Mg^{2+}, Mn^{2+}, or Zn^{2+}) using glutamic acid or aspartic acid ligands in loop-5 and loop-6 (see Figure 3). MHYD also has zinc-binding ligands at the same place, but the residues are all histidine. Also in ALD2 (F16), there are zinc-binding ligands

Figure 2. Functional wheels: the distribution of chemical and biological functions represented by sets of concentric pie charts. (a), (b) and (d) The circles, from inner to outer, represent the second, third and fourth levels in the EC hierarchy. White sector indicates non-enzyme proteins. (c) and (e) The circles, from inner to outer, represent the second, and third levels in the hierarchy of biological functions defined by Rison *et al*..[21] Each of these families may map to one or more EC classes and each EC class may map to one or more biological functions in GenProtEC[19] and KEGG.[20] In such cases, the family is represented more than once. (a) and (c) For the all TIM barrels; (b) for the TIM barrels with phosphate-moiety as cofactors or substrates. (d) and (e) for the 689 *E. coli* genes provided by the PEDANT database.[18] The angle subtended by any segment is proportional to the number of TIM barrel sequence families ((a)–(c)) or *E. coli* genes ((c) and (e)) it contains.

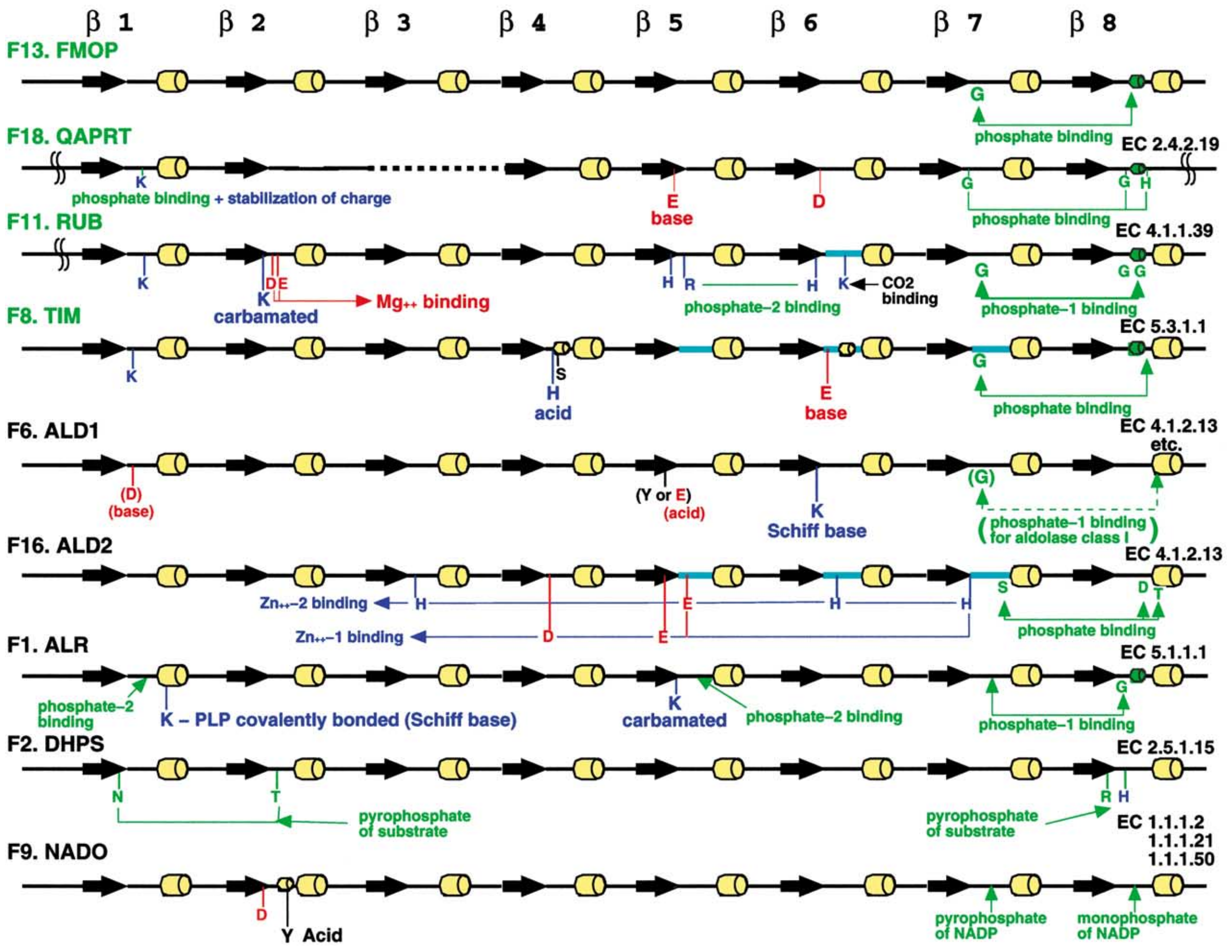

Figure 3 (*legend opposite*)

in the same loop (discussed in detail below). Although the metal-binding site is conserved within each family, the types of metals and residues used for binding can vary (see Figure 3). For example, the third residue used for metal-binding in ENOL (F12) can be aspartic acid, glutamic acid or asparagine (see Figure 3).

Despite this proliferation of metal binding, it is striking that metal-liganding residues occur in all of the eight (βα) motifs, albeit always at the C-terminal end of the barrel (Figures 1(d) and 3). In total there are 33 ligating residues, mostly in loops (19) or strands (12). The summary of metal-binding residues is also given in Figure 1(d), which shows that 58% occur in the C-terminal half of the barrel†.

Common features of ligands: phosphate-moiety in cofactors and substrates

In 12 families (F1–F3, F5, F8–F11, F13, F16–F18), all members have a phosphate-moiety either in their substrates or cofactors such as PLP, FMN and NADP (Table 2). Moreover, another four families have at least one member with a phosphate moiety in the ligand (ALD1 (F6), EG (F7-2), ENOL (F12) and MHYD (F14)). In contrast to the metal-binding sites, there is a strong clustering of phosphate binding sites at the C-terminal end of the sequence. Figure 3 highlights this clustering by showing the location of the phosphate binding sites along the $(\beta\alpha)_8$ barrel sequences marked in green.

Considering first the 12 "obligate" phosphate-binding families, four families (TIM (F8), RUB (F11), FMOP (F13), and QAPRT (F18)) have a common phosphate-binding site on the last two loops of the barrel sheet, as first reported by Wilmanns *et al*. (see Table 2).[3] Here, this common phosphate binding site, ranging from β-7, loop-7, α-7, β-8 to helix-8′, is termed the "standard phosphate binding (SPB) motif". These families will be

† A more detailed description on metal-binding sites can be found at http://www.biochem.ucl.ac.uk/bsm/barrel/tim/metal

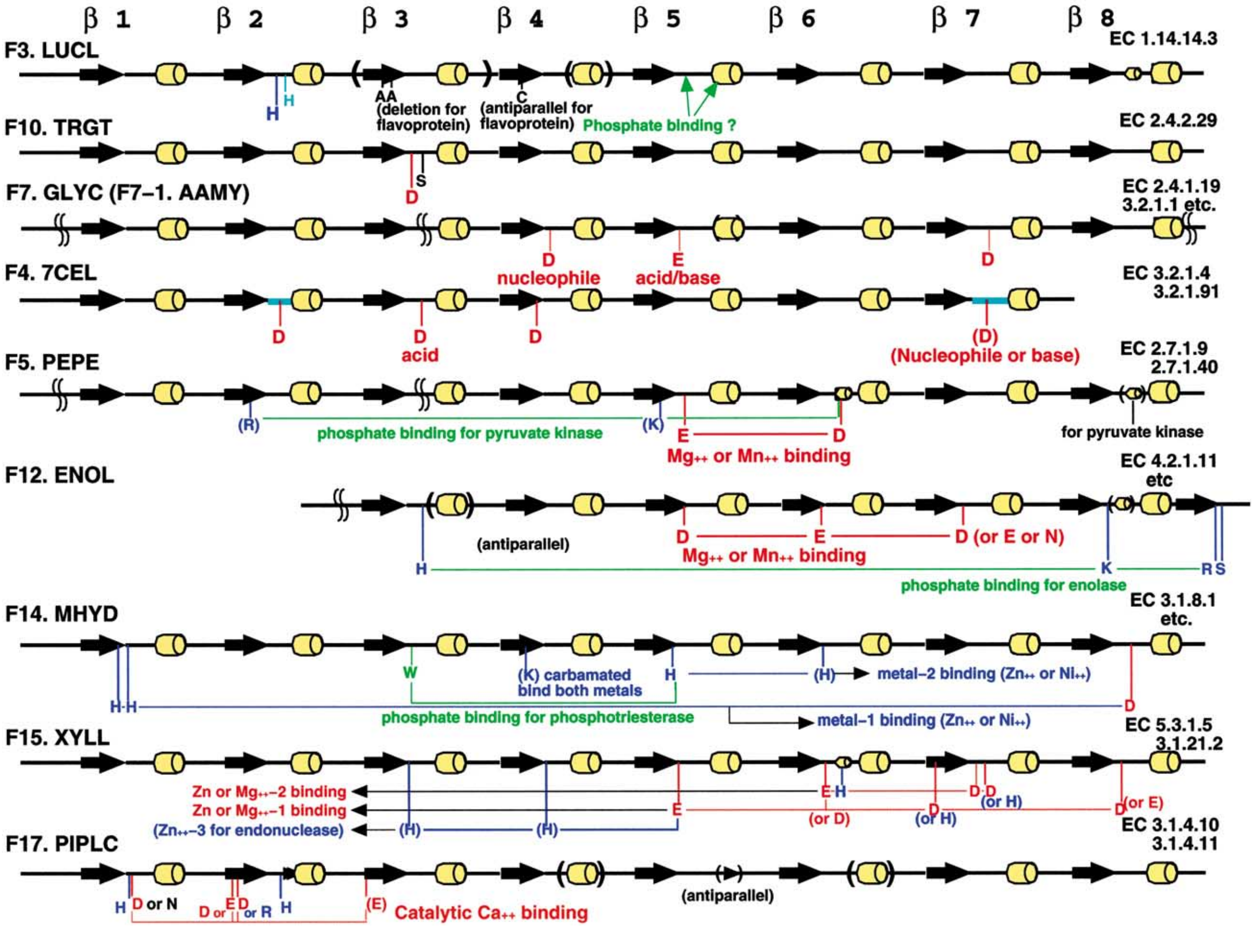

Figure 3. Schematic TIM barrel structures with functional residues. The eight barrel strands and helices are indicated with black arrows and yellow cylinders, respectively. The abbreviation of the family names or enzyme names is indicated on the left-hand of each diagram, whilst EC numbers of the member enzymes are indicated on the right-hand. The phosphate-binding sites are indicated in green. Acidic residues and basic residues involved in catalysis are indicated in red and blue, respectively. The flexible loops are indicated with a cyan blue line. The positions of possible domain insertions are indicated with doubled S-like shapes on loops. The permutations of ENOL (F12) and QAPRT (F18) are also indicated by shifting the graph, and by inserting the dotted line between the secondary structures, respectively. Due to this, the secondary structures following β-7 of ENOL (F12) are not indicated here. In the case of LUCL (F3), the binding position was deduced from that of flavoprotein 390 using multiple-global structure alignment.

described in more detail below, and are indicated in green in Figures 3–7. In addition, we found that ALR (F1) has a similar motif which also binds phosphate on loops-7 and 8 involving also helix-8′. There are four more families (DHPS (F2), ALD1 (F6), NADO (F9) and ALD2 (F16)), which bind the phosphate-moiety on a similar position (loop-7 and loop-8) but the motifs are different and do not include the small helix-8′. A summary of the location of the phosphate-binding sites is given in Figure 1(d), which shows that phosphate binding occurs most frequently at β-7 and β-8.

In contrast, in four more families (LUCL (F3), PEPE (F5), ENOL (F12), and MHYD (F14)), the phosphate-moiety is bound to residues located on different loops in the barrel (Figure 3).

One of the glycosidase enzymes, 6-phospho-β-galactosidase (EC 3.2.1.85) in EG (F7-2), has a phosphate moiety in its substrate, but this is bound to the long loop-8, following β-8.

Although TRGT (F10) has a phosphate moiety in its substrate, its binding site is not known, as the co-crystal structure with the substrate has not been solved.

Within some of the phosphate-binding families, there are variations between members. In ALD1 (F6), only class I aldolase and transaldolase bind a phosphate moiety in their substrate, whilst only enolase binds phosphate in the ENOL family (F12). In the FMOP family (F13), although most members are FMN-dependent, there are several enzymes that do not use any cofactors, whose primary EC numbers are mostly 2, 4 and 5. Moreover, although one protein in NADO (F9) is a voltage-dependent K^+ channel, it nevertheless binds NAD with the three phosphate groups at the same site (PDB code 1qrq).[24] All the enzymes utilizing FMN or NAD are oxidoreductases (primary EC number 1). However, one of the NADP-dependent oxidoreductase enzymes, inosine monophosphate dehydrogenase (IMPDH; EC 1.1.1.205; PDB code 1ak5, 1b3o, 1zfj), is not part of the family NADO (F9). Instead it is grouped into FMOP (F13), as its sequence hits several enzymes from FMOP (F13) using PSI-BLAST, but no proteins from NADO (F9). The phosphate moiety of its NADP is also bound to the SPB motif in loops 7 and 8 (PDB code 1b3o, 1zfj).

As shown in Figure 2(a) and (b), which are sets of concentric pie charts showing the distribution of EC numbers, the proportion of primary EC numbers, 1, 2, and 4 amongst the phosphate-binding barrels is larger than that of all the barrels. Phosphate-binding TIM barrels are predominantly involved in small molecule metabolism and energy metabolism involving oxidoreduction, transferase and lyase enzymes.

PSI-BLAST analyses

Sequence relationships between TIM barrels derived using PSI-BLAST

Since the CATH classification contained families assigned prior to the development of more sensitive analysis tools (i.e. sequence profiles (PSI-BLAST)[25] and 3D-structural profiles (CORA)[26]), the relationships between proteins with TIM barrel structures in CATH were re-analysed, using the sequence profile methods, PSI-BLAST[25] and IMPALA.[27] All the sequence families in the PDB, classified as TIM barrels in CATH, were scanned against all the sequences of PDB and GenBank (10 May 2002 release) using PSI-BLAST.

Only four cross-hits between the 21 different CATH families were directly detected using both PSI-BLAST and IMPALA, which highlights the sequence variability for this fold. In one of these, several sequences from FMOP (F13) (e.g. thiamin phosphate synthase) cross-hit the sequences from QAPRT (F18) (*E*-value 3×10^{-4}; iteration number 3) and *vice versa* (indole-3-glycerol-phosphate synthase (IGP synthase); *E*-value 2×10^{-12}; iteration number 5), aligning the partial sequence of FMOP (F13) covering β-4–β-8 with four strands (β-4–β-7) of QAPRT (F18) by using PSI-BLAST, suggesting an evolutionary relationship between FMOP (F13) and QAPRT (F18). Although QAPRT (F18) is a seven-stranded TIM barrel, it also has the SPB motif, ranging from β-6 to helix-7′, which corresponds to β-7 to helix-8′ of proteins in the FMOP family (F13). The IMPALA analyses, which are supposed to provide a better alignment than PSI-BLAST, also showed significant *E*-values (5×10^{-27}). In this case the alignment of FMOP proteins (F13) as query, aligned six β-strands (β-1, β-3–β-7) of QAPRT (F18), with six strands (β-3–β-8) of proteins from FMOP (F13) (e.g. IMPDH, D-ribulose-5-phosphate 3-epimerase). However, this means β-2 of QAPRT (F18) could not be aligned with any segments of FMOP (i.e. one of the first three βα motifs in QAPRT has clearly been lost during evolution). In the reverse IMPALA alignment of QAPRT (F18) against FMOP (F13), as in the PSI-BLAST results, only the last four or five strand regions of QAPRT could be well-aligned with the corresponding regions of FMOP (F13) (e.g. thiamin phosphate synthase, IGP synthase), whilst the first two or three strands could not be aligned at all with the corresponding secondary structure elements of the FMOP proteins. Thus, it is very difficult to determine which of strands of the first three strand segments had been lost, even using the IMPALA alignment. At best, one of the three segments must have been lost during the evolution. In the second direct hit of PSI-BLAST, the partial sequence of ENOL (F12) (e.g. muconate cycloisomerase), including strands, β-1–β-7, cross-hit the partial sequence (β-4–β-8) of trimethylamine dehydrogenase with *E*-value 5×10^{-8} and iteration number 13 (Figure 4(a)), whilst the IMPALA analyses showed that three proteins from FMOP (F13) (e.g. IMPDH, glycolate oxidase, trimethylamine dehydrogenase) cross-hit proteins from ENOL (F12) (e.g. chloromuconate cycloisomerase, muconate lactonizing enzyme, mandelate racemase) (*E*-value 8×10^{-26}). The best IMPALA alignment showed that six strands (β-3–β-8) of FMOP (F13) were aligned perfectly with six strands (β-1–β-6) of ENOL (F12), which supported the circular permutation of the ENOL family as suggested by Copley & Bork.[10] In the third direct hit, the IMPALA analyses showed that dihydrodipicolinate synthase and *N*-acetylneuraminate lyase (ALD1; F6) cross-hit flavocytochrome *b*2 (FMOP; F13) with significant *E*-value (4×10^{-9} and 6×10^{-7}, respectively; Figure 4(a)), but in reverse flavocytochrome *b*2 did not hit the two proteins, although PSI-BLAST did not detect the relationships, either. Also in the IMPALA alignment, three N-terminal β-strands (β-2–β-4) of *N*-acetylneuraminate lyase (ALD1; F6) were aligned perfectly with three C-terminal β-strands (β-6–β-8) of flavocytochrome *b*2 (FMOP; F13), which suggested the possibility of permutation of ALD1, and will be discussed further below. In the fourth direct hit, the IMPALA analyses showed pyruvate kinase (PEPE; F5) cross-hit thiamin phosphate synthase (FMOP; F13) with significant *E*-value (8×10^{-12}), although PSI-BLAST did not detect the relationships either. The IMPALA alignment showed that five strands (β-4–β-8) of both the families (PEPE and FMOP) were aligned together. Although Copley & Bork found a match between flavocytochrome *b*2 in FMOP (F13) and pyruvate kinase in PEPE (F5) (*E*-value 1×10^{-3}),[10] this fell below our cut-off (*E*-value 5×10^{-4}), which is rather conservative. The *E*-value returned by the IMPALA analysis is much more significant than that by Copley & Bork.[10]

In addition, within the large glycosidase family (GLYC; F7), endoglucanases from EG (F7-2) cross-hit several enzymes from AAMY (F7-1) (Figure 4(a)).[12] The *E*-value (5×10^{-33}) again reflects a clear evolutionary relationship. The IMPALA analyses

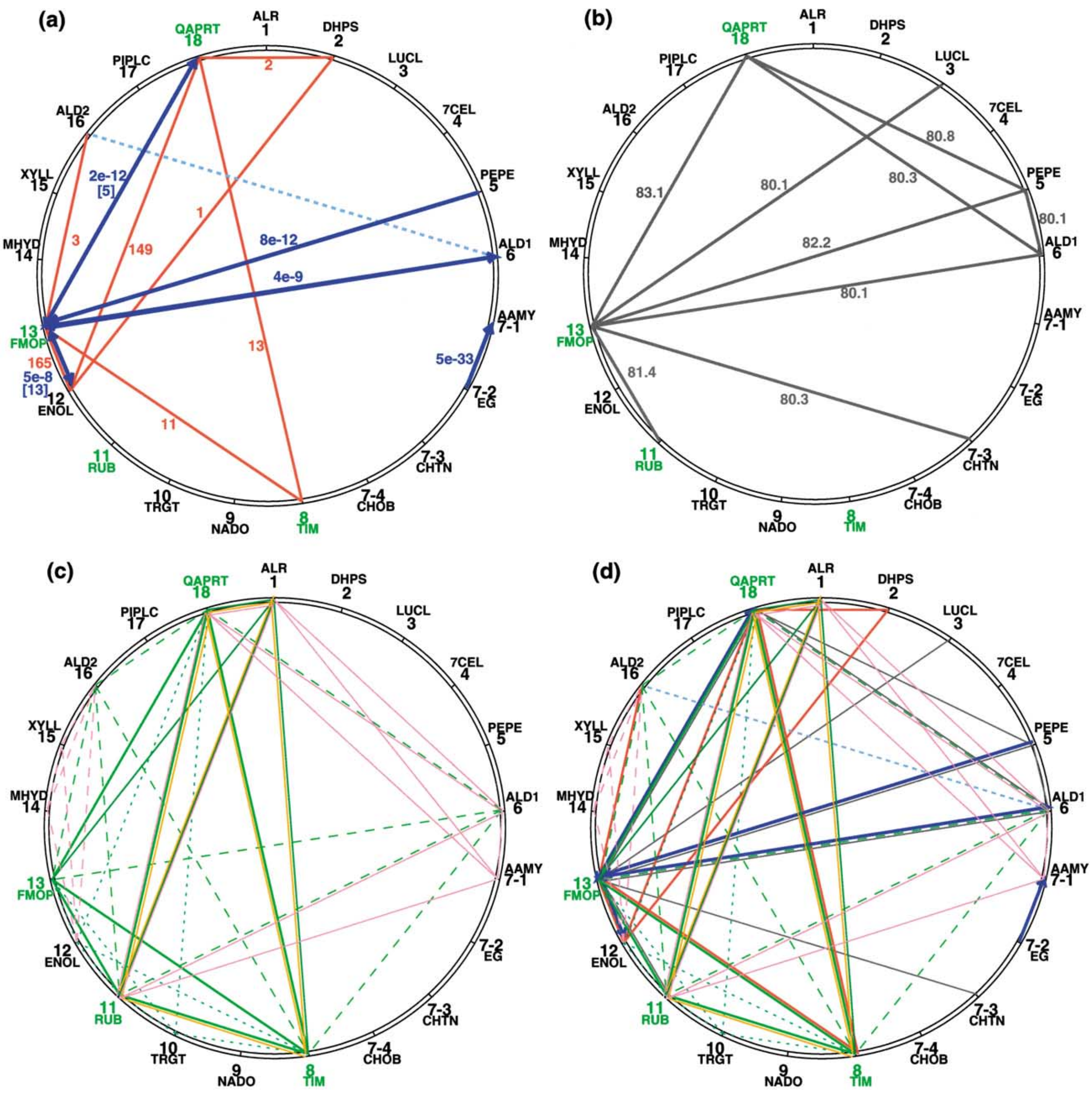

Figure 4. TIM wheels. The relationships between TIM barrel families. The family numbers with the SPB motif are indicated in green. (a) Pairs of families, which could be directly cross-hit using PSI-BLAST[25] and IMPALA,[27] indicated with blue arrows, pointing from seed sequence to hit sequence. The most significant *E*-values of the direct hits are also indicated along the blue arrows. Iteration numbers of PSI-BLAST analyses are indicated in brackets. Pairs of families, for which stepping-stone sequence was identified by the PSI-BLAST analysis, are connected in red, and attached with the number of hit stepping-stone sequences along the line. The two aldolase families (ALD1 (F6) and ALD2 (F16)), for which the stepping-stone sequence was reported by Galperin *et al*.,[28] is indicated with a broken cyan line. (b) SSAP scores higher than 80.0 are indicated with a grey line and the SSAP scores along the line. (c) Functional relationships between families. Those families with catalytic lysine at β-1 aligned by CORA are connected by thick yellow lines. Those with catalytic residues at β-5 aligned are connected by thin pink lines, whilst those with metal-ligating residues at β-5 aligned are connected by thin broken pink lines. The families with the SPB motifs are connected by thick green lines. The family with a similar structure to the SPB motif and phosphate bound to the same site (ALR; F1) is connected to the SPB families with thin green lines. The families with phosphate bound to the same site as the SPB motif (ALD1; F6 and ALD2; F16) are connected to the SPB families by light green broken lines, whilst those with a similar structure to the SPB motif and phosphate not bound to the site (TRGT; F10 and ENOL; F12) are connected to the SPB families by green dotted lines. (d) The superposition of (a), (b) and (c).

showed that AAMY (F7-1) cross-hit EG (F7-2) (*E*-value 1×10^{-26}), and five strands (β-2, β-3 and β-6–β-8) of both the families were aligned perfectly. Thus, the direct PSI-BLAST analysis merges (PEPE (F5), ALD1(F6), ENOL (F12), FMOP (F13) and QAPRT (F18)) and the glycosidases (AAMY (F7-1) and EG (F7-2)) leaving 16 distinct families.

Table 3. A comparison matrix

	1	2	3	4	5	6	7-1	7-2	7-3	7-4	8	9	10	11	12	13	14	15	16	17	18
1		15	9	8	11	10	9	9	9	7	10	11	8	10	10	12	9	10	5	7	7
2	75.4		8	8	12	10	10	10	8	7	11	12	12	8	*12	12	10	10	6	8	*13
3	75.7	74.6		8	8	7	9	9	9	6	9	7	6	7	7	10	9	8	7	6	13
4	69.1	69.4	66.0		9	8	8	9	9	6	10	7	6	7	9	9	7	9	6	6	10
5	74.9	79.3	75.7	71.7		10	14	12	10	10	9	12	14	11	11	*14	9	13	9	8	13
6	73.0	77.6	70.6	68.8	80.1		10	9	9	7	14	8	6	8	10	*14	8	10	8	7	14
7-1	71.6	74.7	68.7	68.0	75.1	72.4		*9	12	8	9	8	7	8	11	12	8	9	8	7	12
7-2	73.5	75.7	72.0	73.7	79.6	73.2	74.1		12	8	7	12	8	10	11	14	8	11	8	7	11
7-3	72.2	72.7	72.0	67.5	76.1	72.0	74.3	73.2		8	7	11	9	9	12	12	7	10	8	7	10
7-4	73.0	73.3	69.8	65.7	74.9	74.5	72.8	73.6	73.7		7	10	4	6	8	9	7	8	5	6	7
8	74.4	75.7	75.0	70.7	75.4	78.9	72.3	68.0	68.2	66.2		8	11	12	11	*13	8	10	11	6	*14
9	72.5	74.0	71.6	67.4	77.5	72.1	69.3	76.7	78.7	76.7	72.7		6	7	10	13	8	10	5	9	15
10	73.9	76.5	68.9	65.8	79.6	71.5	64.4	70.9	68.8	64.7	73.0	72.2		8	8	12	6	7	8	4	10
11	74.4	78.2	70.4	68.1	79.6	74.8	71.4	72.3	72.8	71.2	75.8	71.6	71.8		10	14	8	10	8	7	15
12	75.3	78.2	77.2	72.7	78.7	78.8	75.8	73.9	75.6	71.9	74.8	78.0	75.7	76.1		*13	9	12	8	7	*13
13	76.8	79.4	80.0	73.9	82.2	80.1	78.3	78.8	80.3	76.6	79.2	78.5	78.5	81.4	78.7		11	12	*12	8	*18
14	71.7	73.1	67.3	66.9	73.0	71.2	67.1	69.9	70.0	72.0	71.5	69.0	65.0	71.2	72.1	75.6		10	8	7	9
15	71.6	72.9	70.0	70.5	76.4	73.3	71.4	74.1	71.8	68.1	71.4	71.2	69.6	72.0	77.1	77.6	70.6		7	8	12
16	74.1	77.0	68.6	64.7	79.0	74.2	67.5	74.4	70.4	68.9	75.7	70.5	69.9	71.8	78.2	79.7	67.9	69.1		7	13
17	67.2	70.8	66.5	66.4	72.8	65.6	68.0	68.2	66.2	64.8	66.4	68.4	66.3	69.4	70.1	73.9	64.6	65.6	68.1		9
18	77.8	76.5	80.1	73.5	80.8	80.3	76.7	78.7	76.7	73.9	80.0	79.0	79.9	80.0	77.8	83.1	76.2	74.6	79.0	72.9	

*1 Family pairs detected by PSI-BLAST are attached with asterisk (*) in the sequence identity column.
*2 Correlation coefficient between maximum SSAP and sequence identities: 0.74.

Stepping-stone sequences between homologous families: indirect PSI-BLAST analysis

PSI-BLAST was also used to find pairs of yet more distantly related families that have a common intermediate sequence, dubbed as stepping-stone sequences.[25] This revealed seven additional pairs as illustrated in Figure 4(a), involving DHPS (F2), TIM (F8), ENOL (F12), FMOP (F13), ALD2 (F16) and QAPRT (F18). Three of these families contain the SPB motif (Figure 4(a)).

In addition, since an intermediate sequence between ALD1 (F6) and ALD2 (F16) has recently been found,[28] the ALD1 and ALD2 families are also connected, as shown in Figure 4(a), although our PSI-BLAST analysis did not detect the intermediate. According to the recent analyses,[28] the relationships between the intermediate and ALD1 (F6) is on the basis of its catalytic residues, whilst the relationship with ALD2 (F16) was found by sequence analysis (PSI-BLAST). Although these two classes of aldolase have totally different reaction mechanisms (Figure 3), their ancestral aldolase might have been a Schiff-base-forming enzyme that gave rise to several lines of descent, one of which evolved into the metal-dependent class II aldolase (ALD2 (F16)).[28]

These data provide weaker evidence that it may be possible to enlarge the FMOP family further to include the DHPS (F2), TIM (F8) and ALD2 (F16) homologous families, as well as PEPE (F5), ALD1 (F6), ENOL (F12) and QAPRT (F18).

Global structure comparisons

As all these proteins adopt the same fold, we used global structure comparison scores (SSAP)[29] to identify yet more distant relatives, generating a matrix of SSAP scores and sequence identities derived from the structural alignment. A comparison matrix (Table 3) shows the maximum structure-based sequence identities (right-upper diagonal) and the maximum SSAP scores (lower-left diagonal). FMOP (F13) and QAPRT (F18) showed the highest scores in both indices, with a sequence identity of 18%, and SSAP score of 83.1. This correlates well with the PSI-BLAST results and reinforces the clear evolutionary relationship between FMOP (F13) and QAPRT (F18). In this Table, the top 13 sequence identities (>13%) and the best nine of SSAP scores (>80.0) are shown in red. Pairs of homologous families with SSAP scores >80.0 are connected, as shown in Figure 4(b). Table 3 shows that some, though not all, high SSAP scores correlate with "high" sequence identities (correlation coefficient 0.74 for all the pairs in matrix). However, this is perhaps not surprising, given the problems inherent to sequence identity for very distant relatives. Pairs of families, which are detected by stepping-stone PSI-BLAST, tend to show a sequence identity greater than 11%, in line

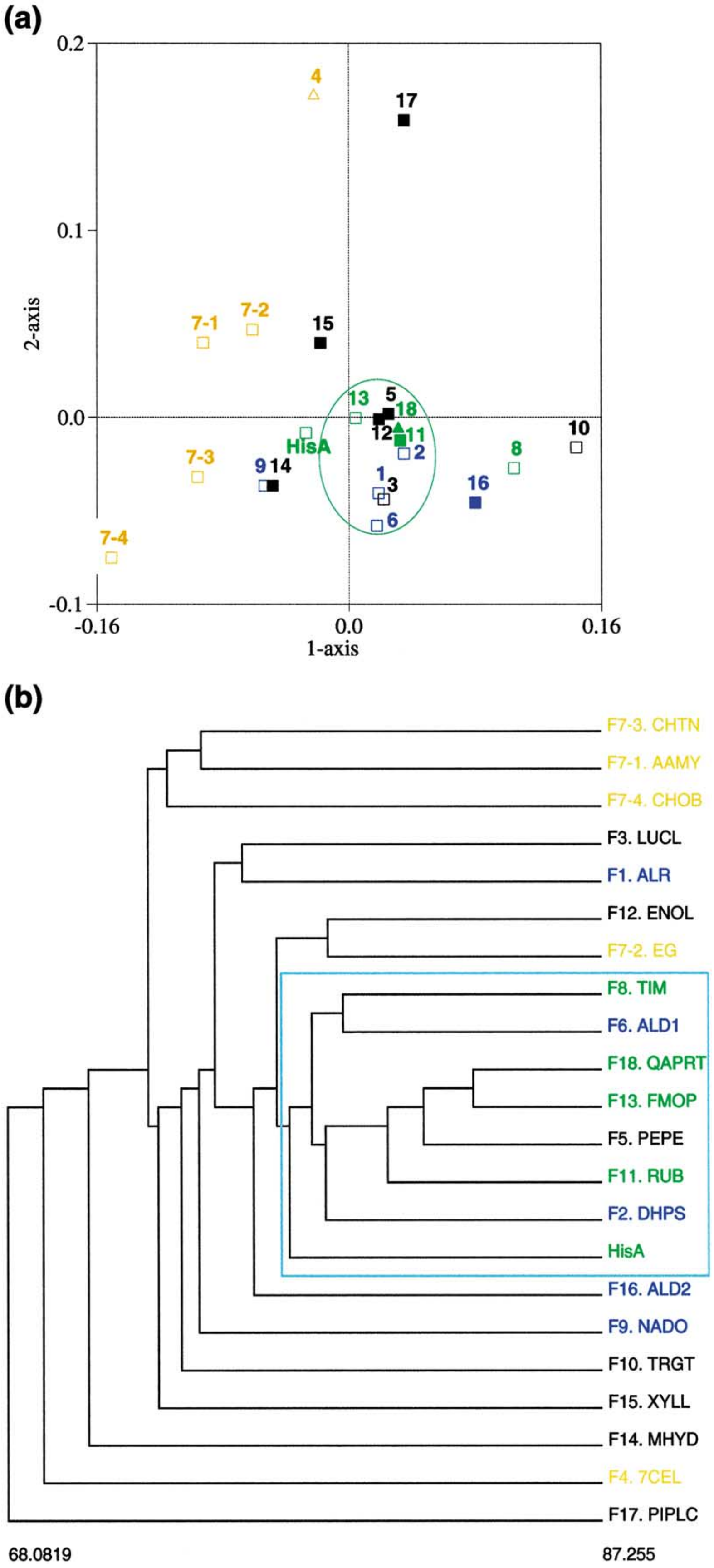

Figure 5. The clustering of the TIM barrel homologous families on the basis of the maximum SSAP scores. (a) The PCA analysis. The yellow markers indicate the glycosidase subgroups or family (F4 and F7). The green colour shows the enzyme families with the SPB motif, whilst the blue one shows those with phosphate-moiety of the ligands in the same position as the SPB motif. Filled markers indicate the metal-binding families, whilst open markers indicate the remainder. The triangles show seven-stranded TIM barrel family. The central cluster is encircled in green. The plot obtained from the maximum matrix of the SSAP scores corresponds to 22.2% of the total information content. Although NADO (F9) and MHYD (F14) appeared overlapped on the plot, they are in fact separated in the third dimension of the PCA plot (data not shown). (b) The phylogenetic tree analysis with means linkage method. Colours are the same as for (a). The cluster that includes all the families with the SPB motif is enclosed in a cyan box.

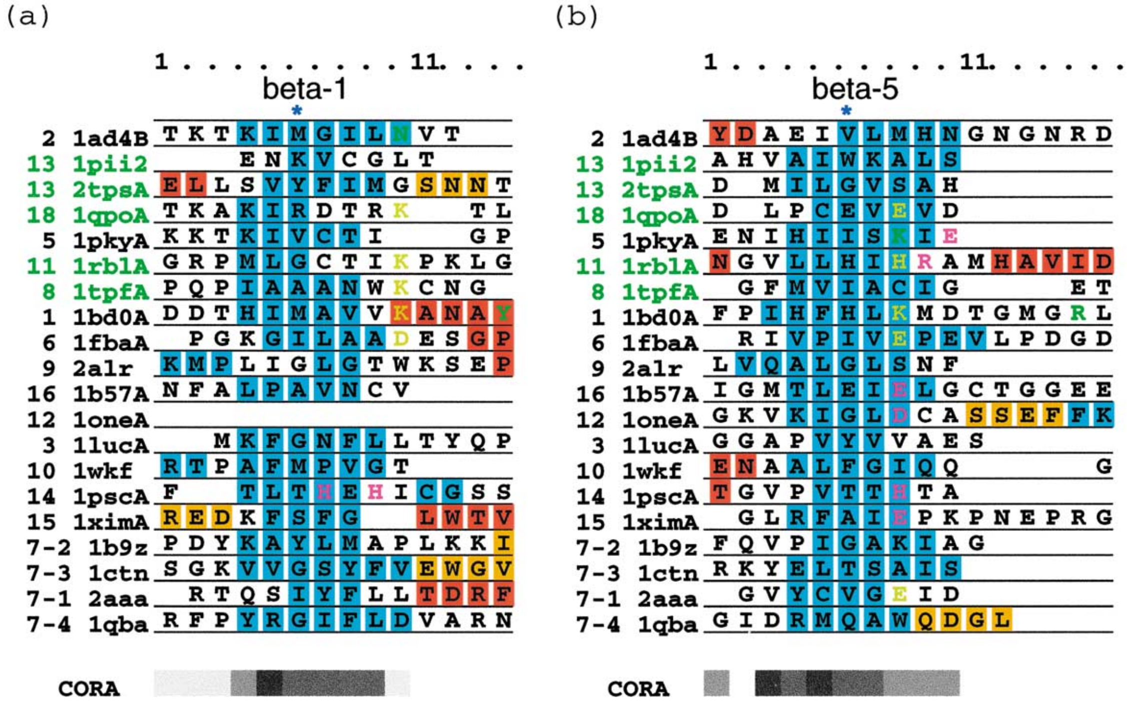
(a)
1 . . . . 11. . . .
beta-1
2 1ad4B T K T K I M G I L N V T
13 1pii2 E N K V C G L T
13 2tpsA E L L S V Y F I M G S N N T
18 1qpoA T K A K I R D T R K T L
5 1pkyA K K T K I V C T I G P
11 1rblA G R P M L G C T I K P K L G
8 1tpfA P Q P I A A A N W K C N G
1 1bd0A D D T H I M A V V K A N A Y
6 1fbaA P G K G I L A A D E S G P
9 2alr K M P L I G L G T W K S E P
16 1b57A N F A L P A V N C V
12 1oneA
3 1lucA M K F G N F L L T Y Q P
10 1wkf R T P A F M P V G T
14 1pscA F T L T H E H I C G S S
15 1ximA R E D K F S F G L W T V
7-2 1b9z P D Y K A Y L M A P L K K I
7-3 1ctn S G K V V G S Y F V E W G V
7-1 2aaa R T Q S I Y F L L T D R F
7-4 1qba R F P Y R G I F L D V A R N
CORA
(b)
1 . . . . 11. . . . . .
beta-5
2 1ad4B Y D A E I V L M H N G N G N R D
13 1pii2 A H V A I W K A L S
13 2tpsA D M I L G V S A H
18 1qpoA D L P C E V E V D
5 1pkyA E N I H I I S K I E
11 1rblA N G V L L H I H R A M H A V I D
8 1tpfA G F M V I A C I G E T
1 1bd0A F P I H F H L K M D T G M G R L
6 1fbaA R I V P I V E P E V L P D G D
9 2alr L V Q A L G L S N F
16 1b57A I G M T L E I E L G C T G G E E
12 1oneA G K V K I G L D C A S S E F F K
3 1lucA G G A P V Y V V A E S
10 1wkf E N A A L F G I Q Q G
14 1pscA T G V P V T T H T A
15 1ximA G L R F A I E P K P N E P R G
7-2 1b9z F Q V P I G A K I A G
7-3 1ctn R K Y E L T S A I S
7-1 2aaa G V Y C V G E I D
7-4 1qba G I D R M Q A W Q D G L
CORA

Figure 6. The catalytic and metal-ligating residues aligned by global structural alignment. CORA alignments of (a) β-1 and (b) β-5. The residues of α-helices, 3/10-helices, and β-strands are indicated with red, yellow, and blue boxes, respectively. The following TIM barrel domains are indicated: 1ad4A (F2), 1pii (the C-terminal domain; PRA isomerase) (F13), 2tpsA (F13), 1qpoA (F18), 1pkyA (F5), 1rblA (F11), 1tpfA (F8), 1bd0A (F1), 1fbaA (F6), 2alr (F9), 1b57A (F16), 1oneA (F12), 1lucA (F3), 1wkf (F10), 1pscA (F14), 1ximA (F15), 1b9zA (F7-2], 1ctn (F7-3), 2aaa (F7-1) and 1qba (F7-4). The names of proteins, which have got the SPB motif, are indicated in green. Due to the shift of strands, the sequence of 1one (F12) is not aligned in (a). The residues, which bind phosphate moiety, are also indicated with green characters, whilst the catalytic and metal-ligating residues are indicated in yellow and magenta, respectively. The core residues of TIM barrels, which contribute to the packing inside the barrel, are indicated with blue asterisks (*). At the bottom of alignment, the monochromatic bar indicates the extent of conservation at each residue site. The darker the portion of the bar, the more conserved the site.

with previous observations that distant homologues often present sequence identities in this range.[30] Considering that the threshold for merging proteins into one homologous family in SSAP scores is generally 80, the figures in the matrix are significant. However, most of the close structural similarities were not detected using sequence analysis, notably FMOP (F13) with LUCL (F3), CHTN (F7-3) and RUB (F11). Among these only RUB has the SPB motif, whilst LUCL binds its phosphates elsewhere.

To identify the most similar structures and explore the extent of structural diversity, we performed a principal component analysis (PCA) of the SSAP scores and also clustered the data using a cluster analysis program OC† to generate a phylogenetic tree (see Figure 5). The SSAP scores and sequence identities were converted to indices, which measure the distance between the families, using the formulae described in Materials and Methods. However, this approach was only illuminating for the SSAP scores, since these very distant relationships are hidden at the sequence level.

Inspection of the PCA plot in Figure 5(a) shows that two families are distant outliers (7CEL (F4) and PIPLC (F17)). Both these homologous families also show very low maximum sequence identities to all the other families. The first 7CEL (F4) is a seven-stranded incomplete barrel, whilst the second PIPLC (F17) is also distorted, as one of the member structures (PDB code 1gym) has an extra antiparallel strand inserted between β-5 and β-6, and another has no helix at all between β-5 and β-6. TRGT (F10) and the other glycosidases (AAMY (F7-1); EG (F7-2); CHTN ((F7-3); CHOB (F7-4)) are also outliers.

The central cluster includes most of the SPB families and most of the families that bind a phosphate moiety in a similar position (C termini of β-7 and β-8), even though they do not have the conserved structural motif (green and blue in Figure 5(a)). TIM (F8), which also has the SPB motif, is a surprising outlier.

The tree, generated from maximum SSAP scores (Figure 5(b)), is mostly consistent with the PCA plot. Families that include an SPB motif are clustered together and they have been enclosed in a cyan box. This group includes DHPS (F2), PEPE (F5) and ALD1 (F6) (Figure 5(b)). As in the PCA plot (Figure 5(a)), TIM (F8) is an outlier

† Barton, G. J. 1993. OC, a cluster analysis program. http://barton.ebi.ac.uk/new/software.html

and is placed on a different branch of the tree from the other three SPB families (RUB (F11), FMOP (F13), QAPRT (F18)). In the centre FMOP (F13) and QAPRT (F18) are tightly clustered as expected.

The histidine biosynthesis enzymes (HisA and HisF)

Recently, Lang *et al*. have reported the tertiary structures of two enzymes involved in histidine biosynthesis, the products of genes HisA and HisF, which evolved by twofold gene duplication and gene fusion from a common half-barrel ancestor.[5] Each half-barrel of these enzymes contains a phosphate-binding motif that is imposed by the nature of their biphosphate substrate. These enzymes have been suggested as strong candidates for the "closest living relative" to a common ancestor of the TIM barrels.[5]

Using PSI-BLAST, the sequence of HisA detected the sequences of several enzymes from FMOP (F13) such as IGP synthase (*E*-value 1×10^{-12}; iteration number 4) and ribulose-phosphate 3-epimerase (*E*-value 8×10^{-8}; iteration number 5). Moreover, structure comparison of HisA (PDB code 1qo2) with the other TIM barrels revealed *N*-(5′-phosphoribosyl) anthranilate isomerase (PRA isomerase) (PDB code 1pii) from FMOP (F13) (SSAP score 80.3) as the most similar structure, and the proteins with the second and third highest SSAP scores are from QAPRT (F18) (SSAP scores 79.6 and 79.0, respectively). In contrast, the sequence identities of HisA (PDB code 1qo2) with the proteins analysed here were not significantly high. Even the highest sequence identity was only 13% (with tryptophan synthase (PDB code 1ubsA; F13) and DHPS (1ad4B; F2)). In the PCA plot (Figure 5(a)), HisA is slightly outside the central cluster, including three of the four SPB families, and relatively close to FMOP (F13). In the SSAP-based phylogenetic tree (Figure 5(b)), HisA lies close to the families with an SPB motif (Figure 5(b)).

In addition, Lang *et al*. pointed out that HisA and HisF are very similar to two enzymes, the gene products of TrpC (IGP synthase; EC 4.1.1.48) and TrpF (*N*-5′-phosphoribosyl anthranilate isomerase (PRA isomerase); EC 5.3.1.24), which catalyse two successive reactions in tryptophan biosynthesis.[5] This highlights the inter-pathway similarity between TrpF and HisA, which both catalyse isomerization reactions. These two enzymes, from TrpC and TrpF, belong to the FMOP (F13) family.

Multiple alignments of 3D structures of TIM barrels, CORA analysis

Shifting strands. All the representative proteins, except 7CEL (F4) and PIPLC (F17), whose structures are very distorted and very distant from the other families (see above, Figure 5), were further analysed by generating a multiple structure alignment using the CORA technique.[26] Extracts from the global CORA alignment of regions β-1, β-5, and β-7 to helix-8′, which corresponds to the SPB motif, are shown in Figures 6(a), (b) and 7(a), respectively†.

From a structural perspective, the third βα motif in the seven-stranded QAPRT (F18) was found to be missing (see Figure 3), although the IMPALA alignment only showed that QAPRT lost one of the first three β-strands. In addition, five strands, β-1, β-3, β-4, β-5 and β-6 of enolase (ENOL; F12) were aligned with β-3, β-5, β-6, β-7 and β-8 of the eight-stranded SPB families (TIM (F8), FMOP (F13) and RUB (F11)). As detected by the IMPALA alignment, this suggests that the ENOL family (F12) might have resulted from a major deletion in an ancestral protein. As β-2 of enolase is antiparallel with the other β-strands, this strand was not aligned with any other strands. This result is consistent with the PSI-BLAST results by Copley & Bork.[10] In the literature there have been various reports of circular permutations within the TIM barrels.

Transaldolase B (EC 2.2.1.2) in ALD1 (F6) was proposed to be a circular permutation of class I aldolase (EC 4.1.2.13).[31] The Schiff-base forming lysine in transaldolase is shifted to β-4 from β-6 in aldolase class I.[31] Surprisingly neither the sequence profile analyses (PSI-BLAST; IMPALA) nor the CORA alignment identified this shift, presumably because they both favour complete global alignments, rather than partial matches. However, a different structure comparison method, Matras,[32] could detect this permutation very clearly. In the Matras alignment, the Schiff-base forming lysine residues of both the proteins were aligned perfectly, and the phosphate binding site (β-7 to β-8 of class I aldolase and β-5 to β-6 of transaldolase B) could be aligned and superimposed‡,§.

Sergeev & Lee have also predicted various circular permutations in ENOL (F12), xylose isomerase from XYLL (F15), taka-amylase from AAMY (F7-1), fructose biphosphate aldolase from ALD1 (F6) on the basis of a local multiple-alignment of β-barrels with TIM (F8), RUB (F11) and FMOP (F13).[33] However, we have no evidence to support this from global sequence and structure comparisons, except the two relationships (FMOP with ALD1, and FMOP with ENOL). Only the results from PSI-BLAST and IMPALA that three strands, β-6, β-7, and β-8 of FMOP (F13) were aligned with β-2, β-3, and β-4 of ALD1 (F6) suggest a circular permutation by four β-strands. The circular permutation by four strands is such a symmetric permutation that it is difficult from these results

† The complete alignment can be found at http://www.biochem.ucl.ac.uk/bsm/barrel/tim/whole/

‡ Matras program: publicly available. http://bongo.lab.nig.ac.jp/~takawaba/matras_pair.html

§ The 3D comparison of the local structures of the phosphate-binding site can be found at http://www.biochem.ucl.ac.uk/bsm/barrel/tim/phos/

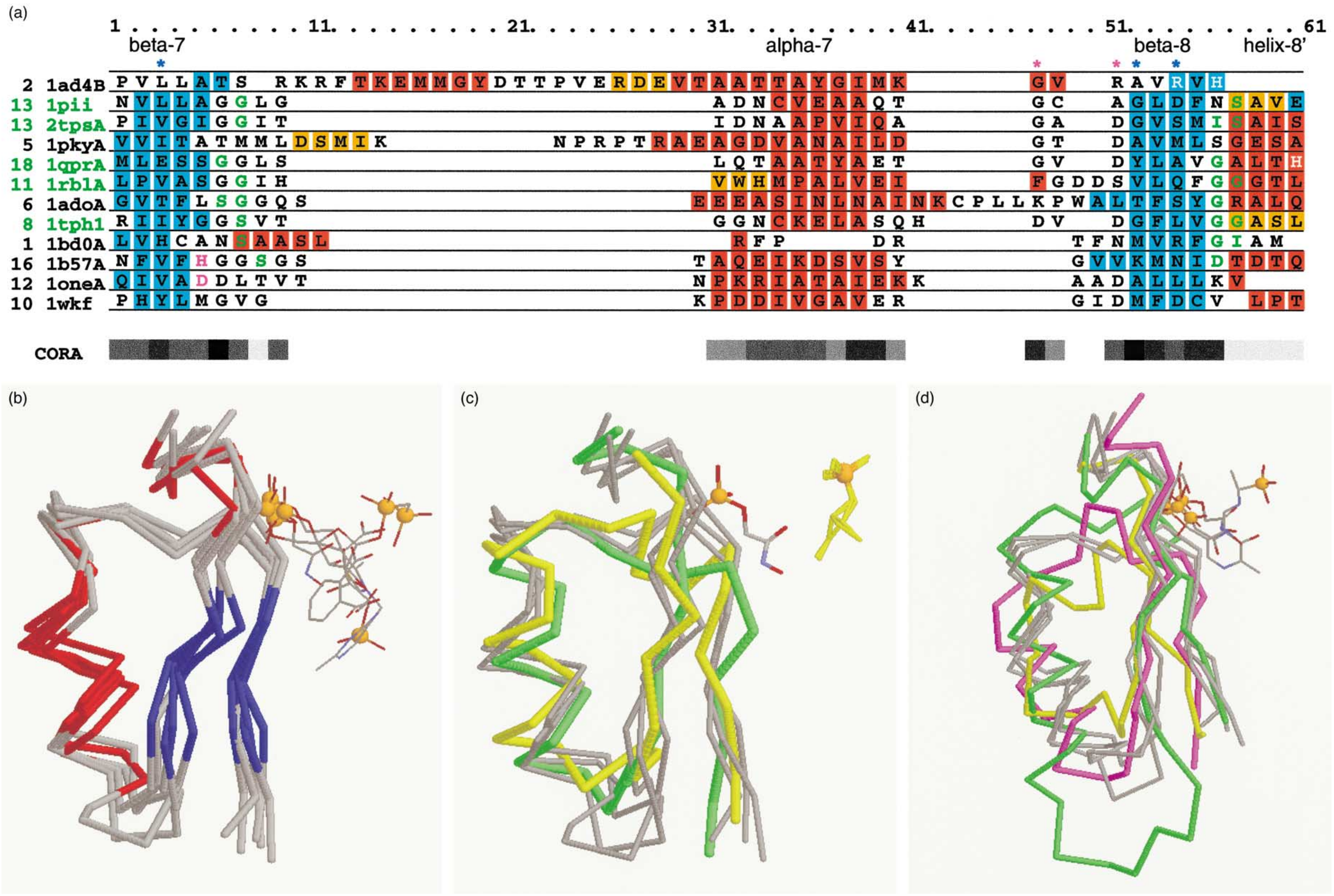
(a)
1 . . . . 11. . . . . 21. . . . . 31. . . . . 41. . . . . 51. 61
beta-7
alpha-7
beta-8
helix-8'
*
2 1ad4B PVLLATS RKRFTKEMMGYDTTPVERDEVTAATTAYGIMK GV RAVRVH
13 1pii NVLLAGGLG ADNCVEAAQT GC AGLDFNSAVE
13 2tpsA PIVGIGGIT IDNAAPVIQA GA DGVSMISAIS
5 1pkyA VVITATMMLDSMIK NPRPTRAEAGDVANAILD GT DAVMLSGESA
18 1qprA MLESSGGLS LQTAATYAET GV DYLAVGALTH
11 1rblA LPVASGGIH VWHMPALVEI FGDDSVLQFGGGTL
6 1adoA GVTFLSGGQS EEEASINLNAINKCPLLKPWALTFSYGRALQ
8 1tph1 RIIYGGSVT GGNCKELASQH DV DGFLVGGASL
1 1bd0A LVHCANSAASL RFP DR TFNMVRFGIAM
16 1b57A NFVFHGGSGS TAQEIKDSVSY GVVKMNIDTDTQ
12 1oneA QIVADDLTVT NPKRIATAIEKK AADALLLKV
10 1wkf PHYLMGVG KPDDIVGAVER GIDMFDCV LPT
CORA
(b)
(c)
(d)

Figure 7 (*legend opposite*)

to be certain whether it has occurred or not. In contrast, as described above, both the CORA and IMPALA alignments support the circular permutation of ENOL (F12) by two β-strands.

Given the similarity in function between the unusual seven-stranded glycosidase (7CEL; F4) and the eight-stranded glycosidase (GLYC; F7) families (F7-1, 7-2, 7-3 and 7-4), we separately used CORA to align the structures. Although the catalytic aspartic acid on the C terminus of β-3 of 7CEL (F4) was aligned with the catalytic glutamic acid on the C terminus of β-4 of EG (F7-2), the global alignment was poor. Thus, there is little evidence from sequence or structural similarities that they have a common ancestor.

Structural features revealed by the multiple structure alignment

Previous structural analyses have suggested that TIM barrels have 12 residues that contribute to the core of the barrel in three layers.[34] They suggest two types of packing depending on whether the core residues of the central layer lie on the odd or even-numbered strands. Using an automatic method for identifying the layers and the core residues involved,[4] we found that the core residues of β-1 and β-5 aligned perfectly in CORA and are shown with blue asterisks in Figure 6(a). In contrast to the previous analyses,[34] our previous results show that even within homologous families, the number of layers can be different and there is no substantive evidence for two different types of packing.[4]

In addition, from the CORA alignment we find that a common sequence motif (Gly-X-Asp) occurs in some of the loops connecting a helix to a strand (α1-β2; α3-β4; α5-β6; α7-β8). In total, there are 16 examples of G-X-D motifs, 26 examples of X-X-D, and 33 examples of G-X-X out of 80 even-numbered strands (20 structures multiplied by four even-numbered strands). This means that 20% of even-numbered strands have got the preceding G-X-D loops, which is a large percentage considering 400 pairwise types of amino acids are possible. These are at the opposite end of the barrel from all the catalytic residues and ligand-binding sites. This may just reflect a preferred geometry for short loops,[35,36] but it is not observed before the odd-numbered strands, so it may suggest 4-fold symmetry, whereby a barrel is constructed by a 4-fold duplication of an ancestral $(\beta\alpha)_2$ motif. This feature is also observed in HisF.

Catalytic residues and metal-ligating residues aligned by multiple-alignment

Using the structure-based sequence alignments we have compared the location of catalytic and metal-ligating residues. We find many catalytic residues and metal-ligating residues which are aligned, the most remarkable regions being in β-1 and β-5 (Figure 6). At the C-terminal end of β-1, catalytic lysine residues from four different families (ALR (F1); TIM (F8); RUB (F11); QAPRT (F18)) are all aligned, together with a catalytic aspartic acid in ALD1 (F6) (Figure 6(a)). At the C termini of β-5, catalytic residues from five families (ALR (F1); ALD1 (F6); AAMY (F7-1); RUB (F11); QAPRT (F18)) and metal-ligating residues from four families (ENOL (F12); MHYD (F14); XYLL (F15); ALD2 (F16)) are aligned together, although the residues vary from acidic (Glu or Asp) to basic (Lys or His) (Figure 6(b)). The phosphate-binding lysine of PEPE (F5) is also aligned at the same site (Figure 6(b)), which is very structurally conserved. Among four of these metal-binding families, three families (ENOL (F12); XYLL (F15); ALD2 (F16)) also have aligned metal-ligating residues in β-7 as well. These functional "matches" between families are indicated with yellow and pink lines in Figure 4(c). (See the following section for discussion of these data.)

In the previous paper by Janecek & Balaz, invariant and functional glutamate residues at the C-terminal end of β-5 from nine TIM barrel

Figure 7. The structural comparison of the SPB motif. (a) CORA alignment of the SPB motif and corresponding regions on the other TIM barrels (β-7 and β-8, helix-8′). The residues of α-helices, 3/10-helices, and β-strands are indicated as for Figure 6. The following domains are indicated: 1ad4A (F2), 1pii (the C-terminal domain; PRA isomerase) (F13), 2tpsA (F13), 1pkyA (F5), 1qprA (F18), 1rblA (F11), 1adoA (F6), 1tph1 (F8), 1bd0A (F1), 1b57A (F16), 1oneA (F12), and 1wkf (F10). The names of proteins are indicated as for Figure 6. The residues, whose main-chain atoms bind phosphate moiety, are also indicated in green, whilst the residues, whose side-chain atoms bind phosphate moiety, are indicated with white letters. The core residues are indicated with blue asterisks (*), whilst the preceding residues of the core residues on the β-8 are indicated with magenta asterisks. (b), (c) and (d) Superimposition of the phosphate binding sites. (b) The SPB motifs from β-7 to helix-8′ in 1pii (F13), 2tpsA (F13), 1qprA (F18), 1tph1 (F8) and 1rblA (F11). The structures were fitted on all main-chain atoms. The C^{α} traces of the protein structures are shown with their ligands that are indicated with stick model. The secondary structures of proteins, helices and strands, are indicated in red and blue, respectively, whilst the remainder are in grey. Here, the phosphorus atoms of ligands are emphasized with orange spheres. (c) C^{α} traces of the corresponding structure of 1wkf (F10) and structure from β-5 to helix-6′ of 1oneA (F12) are shown in green and yellow, respectively, together with the SPB motifs (1pii (F13), 2tpsA (F13), and 1tph1 (F8)) in grey. The ligands of 1tph1 and 1oneA are also indicated with wire frame models, whilst the phosphorus atoms are emphasized with orange spheres. The ligand of 1one is indicated in yellow. (d) Structures from β-7 to helix-8′ of 1bd0A (F1), 1adoA (F6) and 1b57A (F16) are shown in yellow, green, and pink, respectively, as in (c).

families (PEPE (F5), ALD1 (F6), AAMY (F7-1), EG (F7-2), TIM (F8), ENOL (F12), FMOP (F13), MHYD (F14), XYLL (F15)) were reported.[37] However, their results were slightly different from ours. Among the nine families, only the glutamate of three families (ALD1 (F6), AAMY (F7-1), XYLL (F15)) were multiply-aligned using CORA. In contrast, additional functional residues of ALR (F1), RUB (F11), ALD2 (F16) and QAPRT (F18) were multiply-aligned at β-5 in our work, as detailed below.

(1) Instead of glutamate, the phosphate-binding lysine of PEPE (F5) and metal-ligating histidine of MHYD (F14) were multiply-aligned in our work.
(2) The functional glutamate residues of EG (F7-2) and FMOP (F13) are not conserved within their families. In the case of β-amylase of EG (F7-2), which was used as the representative for our CORA alignment, a glutamate residue is not present in in β-5. (Janecek & Balaz mentioned the glutamate of β-galactosidase.[37]) The corresponding glutamate residues of IGP synthase, used by Janecek & Balaz[37] at β-5 of FMOP (F5) were aligned using CORA, but they are not functional
(3) The glutamate residues of TIM (F8) were not aligned at all using CORA alignment. Instead, it was aligned with the second glutamate of ALD2 (F16).
(4) As Janecek & Balaz did not consider the permutation of ENOL (F12), they aligned the metal-ligating glutamic acid at β-5 with other aligned glutamate residues.[37] By detecting the permutation by CORA, however, the metal-ligating aspartate at β-3 could be aligned with the other functional residues (see Figure 6(b)).

Structural analysis of phosphate binding motifs

The phosphate binding sites of the TIM barrels were analysed by superposing the 3D structures of the SPB motifs, found in four families (TIM (F8), RUB (F11), FMOP (F13), QAPRT (F18)), on the basis of the global CORA alignment. In addition the structures of the phosphate-binding families in the core cluster of the PCA plot (ALR (F1), DHPS (F2), PEPE (F5), and ENOL (F12)) were analysed further (Figure 7(a)), together with the phosphate-binding families (ALD1 (F6), and ALD2 (F16)), and TRGT (F10), whose phosphate binding site is unknown.

The SPB motifs in β-7, β-8 and helix-8′ of the four families superposed well (RMSD 1.2–2.2 Å) (Figure 7(b)). The phosphate moieties of the ligands fit between the C termini of β-7 and β-8, and the N terminus of helix-8′ (Figure 7(b)). The residues on the C termini of β-7 and β-8 interact with the phosphate moiety through their main-chain atoms and sometimes also through the side-chain oxygen atoms of a serine residue. Therefore, the type of amino acid is not well conserved, although glycine is widely used (Figure 7(a)).

The region ranging from β-5 and β-6 to helix-6′ of enolase (1one; ENOL; F12) could also be fitted well onto the SPB motif ((RMSD 1.8–2.7 Å) (Figure 7(c)), which supports the alignment shift described above. However, the position of the phosphate moiety on the substrate of enolase (ENOL; F12) is different from that of the phosphate in the classic SPB motif (β-5 and β-6 and helix-6′) (Figures 3 and 7(c)).

The structure of β-7, β-8 and helix-8′ of TRGT (1wkf; F10) superposed well onto the SPB motif (RMSD 1.7–2.0 Å) (Figure 7(c)). Although the phosphate-binding site for TRGT (F10) is not known, the binding site can be predicted from the positions of equivalent phosphate binding residues in the multiple-alignment. This will require experimental validation.

The corresponding motifs (β-7, β-8 and helix-8′) of two classes of aldolases, ALD1 (F6) and ALD2 (F16), and ALR (F1) were also superimposed onto the SPB motifs (Figure 7(d)). Although α-7 of ALR (1bd0; F1) is shortened, the motif was relatively well-fitted (RMSD 2.5–3.0 Å). Of the two aldolases, ALD2 (1b57; F16) fitted reasonably to the phosphate-binding motifs (RMSD 2.4–3.0 Å), although some local structures on loops were distorted. In contrast, ALD1 (1ado; F6) showed larger RMSD values (3.0–3.3 Å). However, all the three structures bind a phosphate moiety at the same position as the SPB motif, using the main-chain atoms (see Figure 7(a) and (d)). (However, considering that ALD1 (F6) might have resulted from the permutation by four β-strands, the original phosphate-binding site might have been at β-3 and β-4, which can support the SPB motif at the N-terminal half-barrel of HisA/HisF.) In addition, TIM and ALD2 have the same substrate, phosphoglycohydroxamate (PGH), and three flexible loops on the same position, loop-5, loop-6 and loop-7, which are rearranged on PGH binding (see Figure 3).[38] Furthermore, in support of this putative relationship, the two aldolase enzymes are positioned adjacent to TIM in the fructose metabolic pathway.

Although pyruvate kinase (1pyk; PEPE (F5)) does not bind phosphate at the same position, the corresponding motif of pyruvate kinase from PEPE (F5) fitted well onto the SPB motif (RMSD 1.9–2.7 Å) (data not shown). In contrast, although DHPS (1ad4B; F2) binds phosphate in the same region, the structures are disrupted (RMSD 2.3–2.6 Å). However, in DHPS, binding occurs through the side-chain atoms, which could lead to the slightly different site for phosphate binding (Figure 7(a))†.

† The structure of the corresponding regions of PEPE (F5) and DHPS (F2) can be found at http://www.biochem.ucl.ac.uk/bsm/barrel/tim/phos/

Amongst the phosphate-binding families, NADO (F9), which binds three phosphate groups between β-7 and β-8, was excluded from this analysis, as it does not give a good alignment in this region, and there is a long insertion. Visual inspection shows that NADO (F9) has a very different structure from the SPB motif, in order to accommodate three phosphate groups in the same place‡.

These functional and structural relationships between families on the basis of the SPB motif are also indicated in green in Figure 4(c). ALR (F1) is strongly connected to the SPB families in this Figure, both in terms of phosphate-binding sites and catalytic residues.

These data provide support for merging those families, which bind phosphate at the same site, especially those with the SPB motif. We can confidently propose a cluster extending FMOP (F13) and QAPRT (F18) to include RUB (F11), TIM (F8) and ALR (F1). The stepping-stone PSI-BLAST hits support the inclusion of TIM, but none of the rest. The global structural comparisons support the RUB linkage but none of the rest. The presence of the functional lysine residues in β-1 in RUB, TIM and ALR provides additional support for merging (see Figures 4(c) and 6(a)). However, the structural details of the phosphate binding sites are difficult to interpret in terms of their evolutionary relationships. To summarise, if we cluster all the families, which either bind phosphate at β-7 and β-8 and/or have a structure in this region, which is similar to the SPB motif, we can group the following.

(1) TIM (F8), RUB (F11), FMOP (F13), QAPRT (F18): SPB motif.
(2) ENOL (β-5, β-6) (F12): SPB motif, though the phosphate is slightly displaced.
(3) TRGT (F10): SPB motif, but no complex structure determination.
(4) ALR (F1), ALD1 (F6) ALD2 (F16), DHPS (F2): phosphate binding at β-7, β-8.
(5) PEPE (F5): structural similarity to SPB, but no phosphate binding.

This clusters 11 of the 21 homologous families. However, it is currently not possible to provide a good statistical basis for this structural/functional clustering, which could have arisen by convergent, rather than divergent evolution.

The superposed catalytic and metal-binding residues in β-1 and β-5 weakly link many of the families hitherto isolated (e.g. MHYD (F14), XYLL (F15) and ALD2 (F16)). The class 1 aldolase family (ALD1 (F6)) has many tantalising but inconclusive similarities to other proteins. The PSI-BLAST hit to FMOP (F13) is strong, but shifted along the structure, superposing the second half of ALD1 (F6) onto the first half of the FMOP family. In contrast, the global structure alignment is not shifted, yet the functional residues overlap (see Figures 4(c), 6(a) and (b)).

In several families (DHPS (F2), LUCL (F3), PEPE (F5), NADO (F9), PIPLC (F17) and the glycosidases (F4 and F7)), the functional residues provide no evidence for ancestral linkage (see Figure 4(c)).

Discussion

This paper highlights the structural and functional diversity of the proteins that adopt a TIM barrel fold. All these proteins are enzymes, or are clearly related to an enzyme, and all the enzymes are involved in molecular or energy metabolism, which are considered to be amongst the oldest biological functions. As such they are ubiquitous in all the kingdoms of life, and are amongst the most frequent folds observed in all genomes. The 889 TIM barrel structures in CATH 1.7 represent 147 non-identical sequences, which were clustered into 76 sequence (>35% sequence identity) families and 21 homologous superfamilies. Here we have considered detailed sequence, structure and functional comparisons to look for evidence for further evolutionary relationships.

The results are summarised in Figure 4(d), which attempts to provide a simple representation of all the linkages found. The more closely related pairs of families are, the more lines connect them in the Figure. The problem in interpreting these data is the lack of good statistical measures to evaluate the significance of the global structural comparisons, the local 3D motif comparisons and the superposition of catalytic functional residues. Global structure comparison scores do not correlate particularly well with the sequence-based linkages shown in Figure 4(a), nor with the local motif and residue site superpositions in Figure 4(c). More sophisticated methods which focus on conserved structural segments, as identified in CORA, may provide a more powerful approach for tracking the evolution of structure.

This work builds on and extends previous analyses†. The reports by Wilmanns *et al.* and Bränden are consistent with our results,[2,3] as these are on the basis of the common phosphate-binding motif (SPB motif in this work). However, the analysis by Reardon & Farber,[9] which divided the TIM barrels into six different groups on the basis of barrel shape, domain composition or insertions and chemical mechanisms, are not consistent with our results. For example, the families with the SPB motif were divided into two different categories in their results. Furthermore, ALD1 (F6) was a separate family from others in their work, although it is connected to many other families herein (Figure 4(d)). Other than for ENOL (F12) and QAPRT

‡ An example of the phosphate-binding site of NADO (F9) can be found at http://www.biochem.ucl.ac.uk/bsm/barrel/tim/phos/

† A summary can be found at http://www.biochem.ucl.ac.uk/bsm/barrel/tim/evol/

(F18) and transaldolase, we did not find any evidence for "circular permutations" or shifts, contrary to previous suggestions.[33]

Copley & Bork have also reported on the evolution of some TIM barrel families linking nine out of 21 homologous families (F2, F5, F6, F8, F11–F13, F16 and F18, which are also grouped in this work).[10] Their work is on the basis of analysis of sequence and metabolic pathways, whilst our study concentrates on the structural relationships and active site residues aligned in the multiple global alignment. They have focused on the phosphate-binding enzymes and the related enzymes such as aldolases, pyruvate kinase, and enolase superfamily (ALD1 (F6); ALD2 (F16); PEPE (F5); ENOL (F12)).[10] Although their results are largely consistent with ours, there are some differences.

Rather than compare individual proteins, we have analysed homologous families, allowing the identification of structural outliers, 7CEL (F4) and PIPLC (F17), and the detection of some distantly related TIM barrel structures, such as alanine racemase (ALR; F1). They identified PEPE (F5) as related to other phosphate-binding TIM barrels using PSI-BLAST, but the *E*-value was low (1×10^{-3}). Although our more strict PSI-BLAST analysis did not detect this, the IMPALA analysis detected this relationship with a more significant *E*-value (8×10^{-12}). Moreover, the global structural comparisons reported here weakly support this relationship. Copley & Bork suggested that ENOL (F12) and PEPE (F5) might have arisen from a common ancestral PEP and Mg ion-binding enzyme, as both bind PEP and either Mg^{2+} or Mn^{2+},[10] although PEPE has only two acidic residues and ENOL has three for metal binding (see Figure 3). Taking into account the shift of ENOL (F12), the metal-binding positions on the enolase sequence are similar to those in PEPE (F5), but neither our 3D multiplealignment nor PSI-BLAST analyses aligned these residues. In addition the binding of PEP is different in the two enzymes. The relationship between PEPE (F5) and ENOL (F12) remains tenuous. Indeed, we found that ENOL (F12) has more similarities with other metal-binding TIM barrel families (MHYD (F14), XYLL (F15), and ALD2 (F16)) (see Figures 4(c) and 6(b)). Although both analyses detected the shifted alignment of the enolase family (ENOL; F12), CORA provided an accurate structural alignment, which allowed the identification of a more conserved phosphate binding motif in β-7 to helix-8′, rather than the site in β-6 and β-7 which they proposed.[10]

Analysis of the catalytic and metal-binding residues herein has found many co-located functional sites, which are common to many TIM barrels.

There are some families which do not appear to be related to the others by any of the data. These may be extremely distant relatives, or alternatively they may represent examples of convergent evolution. Two of the outlying families, PIPLC (F17) and 7CEL (F4), have nine and seven strands, respectively, differing from the geometrically favoured eight-stranded barrel, which is so ubiquitous. In the case of seven-stranded cellulase (7CEL; F4), although its function is similar to GLYC (F7), there is no strong evidence to suggest it has arisen from the same ancestor except the multiply aligned catalytic residue.

At the most conservative, using sequence comparisons only, we can link PEPE 9F5), ALD1 (F6), FMOP (F13), QAPRT (F18), ENOL (F12) and TIM (F8) in one group and AAMY (F7-1) with EG (F7-2) in another group (see Figure 4(a)). The first grouping is supported by the similarity of the SPB motif.

At the most extreme (where all links are considered to indicate an evolutionary relationship) we can link all the families into one large cluster, apart from 7CEL (F4), CHOB (F7-4), NADO (F9) and PIPLC (F17). Among the 17 linked families, ten families ALR (F1), DHPS (F2), PEPE (F5), ALD1 (F6), TIM (F8), RUB (F11), ENOL (F12), FMOP (F13), ALD2 (F16) and QAPRT (F18) seem more closely related (Figures 4(d), 5(a) and (b)). The co-location of active site residues and phosphate binding sites suggested that AAMY (F7-1), TRGT (F10), MHYD (F14), and XYLL (F15) might also be distant relatives to these ten families. Although CHOB (F7-4) could not be connected to this large cluster in Figure 4(d), our previous work found that one of the catalytic residues at β-4 of this family was aligned with that of CHTN (F7-3) in the global structural alignment.[12] In addition they have similar substrates, *N*-acetylated sugars. Similarly, although the phosphate-binding structure of NADO (F10) is very distorted, the phosphate groups are bound to the same place, and it has a catalytic residue at the same position as the catalytic carbamated lysine at β-2 of RUB (F11) in the global alignment. Considering these data, these two families (CHOB, F7-4; NADO, F10) might possibly be connected to the large cluster. However, these functional links are fragile, since we know from other protein families that similar metal binding sites (e.g. zinc binding sites) and functional catalytic sites (e.g. the catalytic triad) can evolve independently. However, in the TIM barrels the additional factor is that these sites are co-located and of course the folds are all the same.

Therefore at this point in time, when our knowledge of sequence and structure and function is still limited, there are hints of a common ancestry for 17 of the 21 TIM barrel families, although these links are much stronger for ten families than the rest. As more data are collected, in particular more "stepping-stone" sequences and structures, only more linkages can be found.

Attempts to identify a common ancestral protein are fraught with difficulties. The observation of the recurrent loop sequence motifs (G-X-D) preceding the even-numbered strands may suggest evolution from a common $(\beta\alpha)_2$ motif, with two duplications to create the eightfold barrel.

Similarly, although the data for gene duplication in HisA and HisF are clear-cut, we did not find evidence of duplication in the other TIM barrels, nor that HisA/HisF are necessarily the closest extant relative to a common ancestral protein. Other than in HisA and HisF, the SPB motif has been found only in the C-terminal half of barrels.

In conclusion, this paper highlights the amazing sequence and functional diversity of TIM barrels, the most ubiquitous and undoubtedly one of the most ancient folds. From a practical perspective, predicting the biochemical function for a TIM-barrel fold will be difficult, unless proteins of known function and close sequence similarities (>30% sequence identity) are available. This will be a challenge for structure genomics.

Material and Methods

The TIM barrel structures in the Protein Data Bank (PDB) were clustered into homologous superfamilies and sequence families according to version 1.7 (22562 domains) of the CATH classification.[39] In CATH, multidomain proteins are subdivided into their constituent domains using a consensus procedure.[40] Structures from the PDB are grouped into homologous families and sequence families, using two different sequence comparison methods; the modified version of the pairwise sequence comparison method by Needleman & Wunsch,[41] and the profile-based method, PSI-BLAST.[25] Structure comparisons are performed to detect more distant homologues and analogues, using the SSAP program.[29] CATH has several levels of the classification such as homologous superfamily (H) and sequence family (S).[6] In CATH, a protein is clustered into a homologous superfamily (H-level) if:

(1) the sequence identity of any member of the cluster is equal to or larger than 35%, and 60% of the larger structure is equivalent to the smaller one;
(2) the SSAP score is equal to or larger than 80.0, sequence identity is larger than 20%, and 60% of the larger structure is equivalent to the smaller protein; or
(3) 60% of the larger structure is equivalent to the smaller, the SSAP score is equal to or larger than 80.0, and the domains have related functions.

For new TIM barrel structures, it is often necessary to check the classification manually.

Within each H-level, protein structures are subdivided into sequence families, the S-level of CATH, on the basis of sequence identities. Domains clustered in the same sequence families have sequence identities larger than 35% to another member of the family, with at least 60% of the larger domain equivalent to the smaller, indicating highly similar structures and functions.

The representative proteins of each homologous family are listed in Table 1. The completeness of TIM barrels was analysed using the previously developed method.[4] Chemical and biological functions of all the sequence families were analysed on the basis of EC classes of enzymes, as most TIM barrels are enzymes and have been assigned an EC classification. Whilst the chemical reaction of an enzyme can be inferred directly from its EC class, its biological function must be annotated separately using databases such as GenProtEC[19] and Kyoto Encyclopedia of Genes and Genomes (KEGG),[20] where each EC class of enzyme is assigned to biological functions or metabolic pathways. Each protein is annotated according to the scheme proposed by Rison *et al*.[21] which integrates the biological functions or metabolic pathways of several databases into the hierarchical levels of the biological. As some proteins have more than one function, all the functions are indicated in this work. The distributions of chemical and biological functions were displayed using a modified version of the software used to generate "CATH wheels".[42] To compare with the functions of TIM barrels, 689 genes from *E. coli* at the PEDANT database[18] were also analysed according to the scheme[21]. In the PEDANT database,[18] EC numbers and biological functions (FunCat in PEDANT) are assigned by BLASTing the organisms' protein sequences against master-lists, and by assigning the function/EC number of the top hits.

Principal component analyses (PCA) were carried out, in order to analyse the relationships between the families, using the SSAP scores and sequence identities of non-identical proteins in the dataset taken from version 1.7 of the CATH classification. SSAP scores and sequence identities could be converted to indices showing the distances between the families. As sequence identities beyond homologous families are generally lower than 20, the distances, D_{seq}, were obtained using the following formula:

$$D_{\text{seq}} = 20 - (\%\ \text{sequence identity}) \qquad (1)$$

In contrast, as the SSAP scores are between 0 and 100 with higher scores for larger similarity, the distances, D_{SSAP}, were calculated using the following formula:

$$D_{\text{SSAP}} = (100 - \text{SSAP\_Score})/100 \qquad (2)$$

The structural alignments of TIM barrels were performed using CORA technique,[26] which generates multiple-alignment of 3D structures of proteins. The aligned structures were superimposed and viewed using the 3D viewer from the SAS package.[43] The Matras program,[32] which adopts a different scoring system from SSAP[29] by considering the Markov Transition Model of Evolution, was also used, especially to obtain local structure alignment and superimposition†.

The PSI-BLAST program[25] was also used to detect weak but biologically relevant sequence similarities by following the procedure to assign genomic sequences to the CATH database.[39,44] Each non-identical TIM barrel sequence was PSI-BLASTed against GenBank (10 May 2002 release). Stepping-stone sequences are then identified as those sequences matched by PSI-BLAST[25] in common to two TIM barrel families. The maximum number of iterations allowed is 20, and the E-value for inclusion in the next pass is 0.0005. All sequence segments with E-values <0.0005 and of similar length to the probe sequence (overlap >80%) were identified as putative homologous domains. The IMPALA suite,[27] which is a refined version of PSI-BLAST and supposed to provide more optimal alignments than original PSI-BLAST, was also applied to get alignments against GenBank.

† This program is publicly available at http://bongo.lab.nig.ac.jp/~takawaba/Matras_pair.html

Phylogenetic tree analysis of the families was performed using OC, a cluster analysis program, which clusters using a means linkage algorithm§.

Acknowledgements

This research was funded by the Japan Society for the Promotion of Science (JSPS). We acknowledge support for computing from the BBSRC. We thank Mr James Bray, Mr Ian Sillitoe, Dr Andrew Harrison and Dr Roman Laskowski for useful advice on SSAP and CORA analysis; Dr Frances Pearl and Dr David Lee on PSI-BLAST analyses; and Stuart Rison on functional classification. Moreover, we thank Dr Dimitrij Firshman and Dr Martin Mokrej for providing us with the PEDANT data.

References

1. Farber, G. K. & Petsko, G. A. (1990). The evolution of α/β barrel enzymes. *Trends Biochem. Sci.* **15**, 228–234.
2. Bränden, C. I. (1991). The TIM barrel—the most frequently occurring folding motif in proteins. *Curr. Opin. Struct. Biol.* **1**, 978–983.
3. Wilmanns, M., Hyde, C. C., Davies, D. R., Kirschner, K. & Jansonius, J. N. (1991). Structural conservation in parallel β/α barrel enzymes that catalyze three sequential reactions in the pathway of tryptophan biosynthesis. *Biochemistry*, **30**, 9161–9169.
4. Nagano, N., Hutchinson, E. G. & Thornton, J. M. (1999). Barrel structures in proteins—automatic identification and classification including a sequence analysis of TIM barrels. *Protein Sci.* **8**, 2072–2084.
5. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546–1550.
6. 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.
7. 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.
8. Bränden, C. I. & Tooze, J. (1991). α/β structures. In *Introduction to Protein Structure* (Bränden, C. I. & Tooze, J., eds), pp. 43–57, Garland Publishing, New York.
9. Reardon, D. & Farber, G. K. (1995). The structure and evolution of α/β barrel proteins. *FASEB J.* **9**, 497–503.
10. 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.
11. Fani, R., Lio, P., Chiarelli, I. & Bazzicalupo, M. (1994). The evolution of the histidine biosynthetic genes in prokaryotes: a common ancestor for the hisA and hisF genes. *J. Mol. Evol.* **38**, 489–495.
12. Nagano, N., Porter, C. T. & Thornton, J. M. (2001). The $(\beta\alpha)_8$ glycosidases: sequence and structure analyses suggest distant evolutionary relationships. *Protein Eng.* **14**, 845–855.
13. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45–48.
14. Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Finn, R. D. & Sonnhammer, E. L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucl. Acids Res.* **27**, 260–262.
15. Rouvinen, J., Bergfors, T., Teeri, T., Knowles, J. K. & Jones, T. A. (1990). Three-dimensional structure of cellobiohydrolase II from *Trichoderma reesei*. *Science*, **249**, 380–386.
16. Spezio, M., Wilson, D. B. & Karplus, P. A. (1993). Crystal structure of the catalytic domain of a thermophilic endocellulase. *Biochemistry*, **32**, 9906–9916.
17. Todd, A. E., Orengo, C. A. & Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J. Mol. Biol.* **307**, 1113–1143.
18. Frishman, D. & Mewes, H. W. (1997). PEDANTic genome analysis. *Trends Genet.* **13**, 415–416.
19. Riley, M. (1998). Genes and proteins of *Escherichia coli* K-12 (GenProtEC). *Nucl. Acids Res.* **26**, 54.
20. Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H. & Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucl. Acids Res.* **27**, 29–34.
21. Rison, S. C. G., Hodgman, T. C. & Thornton, J. M. (2000). Comparison of functional annotation schemes for genomes. *Funct. Integr. Genom.* **1**, 56–69.
22. Sharma, V., Grubmeyer, C. & Sacchettini, J. C. (1998). Crystal structure of quinolinic acid phosphoribosyltransferase from *Mmycobacterium tuberculosis*: a potential TB drug target. *Structure*, **6**, 1587–1599.
23. Holm, L. & Sander, C. (1997). An evolutionary treasure: unification of a broad set of amidohydrolases related to urease. *Proteins: Struct. Funct. Genet.* **28**, 72–82.
24. Gulbis, J. M., Mann, S. & MacKinnon, R. (1999). Structure of a voltage-dependent K^+ channel β subunit. *Cell*, **97**, 943–952.
25. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., 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.
26. Orengo, C. A. (1999). CORA—topological fingerprints for protein structural families. *Protein Sci.* **8**, 699–715.
27. Schäffer, A. A., Wolf, Y. I., Ponting, C. P., Koonin, E. V., Aravind, L. & Altschul, S. F. (1999). IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
28. 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 Microbiol. Letters*, **183**, 259–264.
29. Taylor, W. R. & Orengo, C. A. (1989). Protein structure alignment. *J. Mol. Biol.* **208**, 1–22.
30. Russell, R. B., Saqi, M. A., Sayle, R. A., Bates, P. A. & Sternberg, M. J. (1997). Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.* **269**, 423–439.
31. Jia, J., Huang, W., 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.

§Barton, G. J. 1993. OC, a cluster analysis program. http://barton.ebi.ac.uk/new/software.html

32. Kawabata, T. & Nishikawa, K. (2000). Protein tertiary structure comparison using the Markov transition model of evolution. *Proteins: Struct. Funct. Genet.* **41**, 108–122.
33. Sergeev, Y. & Lee, B. (1994). Alignment of β-barrels in (β/α)8 proteins using hydrogen-bonding pattern. *J. Mol. Biol.* **244**, 168–182.
34. Lesk, A. M., Bränden, C. I. & Chothia, C. (1989). Structural principles of α/β barrel proteins: the packing of the interior of the sheet. *Proteins: Struct. Funct. Genet.* **5**, 139–148.
35. Edwards, M. S., Sternberg, M. J. E. & Thornton, J. M. (1987). Structural and sequence patterns in the loops of βαβ units. *Protein Eng.* **1**, 173–181.
36. Scheerlinck, J. P., Lasters, I., Claessens, M., De Maeyer, M., Pio, F., Delhaise, P. & Wodak, S. J. (1992). Recurrent αβ loop structures in TIM barrel motifs show a distinct pattern of conserved structural features. *Proteins: Struct. Funct. Genet.* **12**, 299–313.
37. Janecek, S. & Balaz, S. (1995). Functionally essential, invariant glutamate near the C terminus of strand β 5 in various (α/β)8-barrel enzymes as a possible indicator of their evolutionary relatedness. *Protein Eng.* **8**, 809–813.
38. Hall, D. R., Leonard, G. A., Reed, C. D., Watt, C. I., Berry, A. & Hunter, W. N. (1999). The crystal structure of *Escherichia coli* class II fructose-1, 6-bisphosphate aldolase in complex with phosphoglycolohydroxamate reveals details of mechanism and specificity. *J. Mol. Biol.* **287**, 383–394.
39. Pearl, F. M. G., Lee, D., Bray, J. E., Sillitoe, I., Todd, A. E., Harrison, A. P. *et al.* (2000). Assigning genomic sequences to CATH. *Nucl. Acids Res.* **28**, 277–282.
40. Jones, S., Stewart, M., Michie, A., Swindells, M. B., Orengo, C. & Thornton, J. M. (1998). Domain assignment for protein structures using a consensus approach: characterization and analysis. *Protein Sci.* **7**, 233–242.
41. Needleman, S. B. & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
42. 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.
43. Milburn, D., Laskowski, R. A. & Thornton, J. M. (1998). Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.* **11**, 855–859.
44. Pearl, F. M. G., Lee, D., Bray, J. E., Buchan, D. W. A., Shepherd, A. J. & Orengo, C. A. (2001). The CATH extended protein-family database providing structural annotations for genome sequences. *Protein Sci.* **11**, 233–244.

Edited by B. Honig

(Received 26 November 2001; received in revised form 27 June 2002; accepted 27 June 2002)