Identification and analysis of catalytic TIM barrel domains in seven further glycoside hydrolase families

Daniel J. Rigden[a,*], Mark J. Jedrzejas[b], Luciane V. de Mello[a]

[a] *Embrapa Genetic Resources and Biotechnology, Cenargen/Embrapa, Estação Parque Biológico, Final W5, Asa Norte, 70770-900, Brasília-DF, Brazil*
[b] *Children's Hospital Oakland Research Institute, Oakland, CA 94609, USA*

Received 2 April 2003; revised 25 April 2003; accepted 25 April 2003

First published online 13 May 2003

Edited by Robert B. Russell

Abstract Fold recognition results allocate catalytic triose phosphate isomerase (TIM) barrels to seven previously unassigned glycoside hydrolase (GH) families, numbers 29, 44, 50, 71, 84, 85 and 89, enabling prediction of catalytic residues. Modelling of GH family 50 suggests that it may be the common evolutionary ancestor of families 42 and 14. TIM barrels now comprise the catalytic domains of more than half of the assigned GH families, and catalyse a much larger variety of GH reactions than any other catalytic domain architecture. Only 327 GH sequences still have no structurally identified catalytic domain.

Key words: Glycosidase; Fold recognition; Evolutionary relationship; Triose phosphate isomerase barrel; Catalytic residue

1. Introduction

Glycoside hydrolases (GHs; EC 3.2.1.x) are a remarkably diverse group of enzymes and catalyse the degradation of a huge variety of naturally occurring carbohydrates and glycoconjugates. As a means to organise knowledge of GH enzymes and to direct further research, a database, named CAZY, has been introduced and developed [1,2] which is now available for public access on the World Wide Web [3]. The CAZY database is structured into evolutionarily related GH families, each of which may contain members with different catalytic activities. These families may be grouped into clans when crystal structures or computational analyses provide clear evidence of an evolutionary relationship, even in the absence of significant overall sequence similarity. This contrasts with the general classification of catalytic activities in the EC scheme which assigns a number to each chemical reaction catalysed. In the case of GH enzymes, the CAZY mode of classification is by far the more useful since structurally unrelated enzymes may catalyse the same chemical reaction, while closely homologous proteins can catalyse different reactions. Grouping by evolutionary relationships enables inferences to be drawn for entire GH families, and to a lesser extent for clans, based on experimental data derived from study of an individual protein. Key characteristics amenable to this treatment are the identities of catalytic and substrate binding residues and the nature of the mechanism, whether retaining or inverting the anomeric configuration of the carbon atom of the cleaved bond.

Analysis of structurally diverse GHs shows that they invariably possess an elongated cleft to accommodate the glycan molecules [4]. Catalysis nearly always involves two conserved acidic residues, one acting as acid, the other as base. Inverting and retaining GH enzymes have characteristic spacings between these two acidic residues of around 10 Å and 5.5 Å, respectively [5]. Exceptions to this pattern are the GH families 18 and 20 which work by substrate-assisted catalysis [6,7]. In these cases there is only a single acidic catalytic residue, although transition state stabilisation is often provided by another acidic residue at the catalytic site. Recent work suggests that a hydrophobic platform, acting to stabilise the transition state, is also a ubiquitous feature of GHs [8]. These relatively simple catalytic requirements may offer part of the explanation for the remarkable structural diversity of GHs [9]. Although triose phosphate isomerase (TIM) barrels are conspicuously abundant [10], and β-helical proteins also show a tendency for carbohydrate binding [11], GHs span the full range of protein architectures [9] from all-α to all-β proteins, and including α/β, α+β and multidomain enzymes. Many GH families contain varying numbers of additional carbohydrate binding domains to enhance catalytic efficiency [12]. This modular nature represents a challenge to functional annotation [13], probably best addressed by piecemeal analysis of the various components in the context of the CAZY hierarchical database.

Encouraged by the insights previously gained from computational analysis of GH families [14–16], we have systematically analysed those families for which catalytic domain architecture is currently unknown. Here we present evidence for the presence of TIM barrel catalytic domains in seven of these families, numbers 29, 44, 50, 71, 84, 85 and 89. Members of these families are implicated in human diseases (families 29 and 89 [17,18]), basic cytosolic carbohydrate recycling (family 85 [19]) and cellular development (families 29 and 71 [20,21]). Others are noteworthy for their biotechnological potential (family 50 [22]) or possible role in bacterial infections (family 84 [23,24]). The localisation of TIM barrel catalytic domains enables the identities of catalytic acidic residues to be predicted and will aid in their eventual structural determination. These data also improve our understanding of TIM barrel GH evolution.

*Corresponding author. Fax: (55)-61-340 3658.
E-mail address: daniel@cenargen.embrapa.br (D.J. Rigden).

0014-5793/03/$22.00 © 2003 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.
doi:10.1016/S0014-5793(03)00481-2

2. Materials and methods

Members of GH families 29, 44, 50, 71, 84, 85 and 89 were located in the CAZY database ([3] and retrieved from GenBank). Alignments of each family were made with T-Coffee [25]. Jalview (http://www.ebi.ac.uk/~michele/jalview) was used to manipulate alignments and for calculation of maximally diverse representative sets of three sequences for each family. Examination of sequence conservation along the alignment was used to determine the core conserved region common to all members of a particular family. All sequences were screened against the PFAM [26], SMART [27] and CDD [28] domain databases to search for the presence of known domains. Domain alignments meeting the respective default *E*-value cut-offs of these databases were taken as reliable, but checked for coverage of the entire length of the domain database entry. Secondary structure predictions were made with PSI-PRED [29]. The core conserved regions of the diverse three sequences representing each family (along with CDD-defined consensus sequence when available) were submitted for fold recognition at the Structure Prediction META server [30]. Regions of the core conserved regions outside of the predicted TIM barrel were also submitted. Most attention was paid to the scores produced by the consensus fold recognition method Pcons2 [31] which outperforms individual methods [32]. The results were compared with those ongoing fold recognition server evaluation experiment LiveBench [32] in order to assess their significance. Iterated sequence database searches were carried out using PSI-BLAST [33] at the NCBI (http://www.ncbi.nlm.nih.gov/BLAST/). Model building was carried out with MODELLER 6 [34], employing an iterative modelling scheme in which sets of models were analysed for packing and solvent exposure with PROSA II analysis [35], and for stereochemical properties with PROCHECK [36]. Relationships of known structures were analysed using the SCOP database [37].

3. Results

3.1. Presence of TIM barrels

The fold recognition scores presented in Table 1 clearly demonstrate the presence of catalytic TIM barrel domains in the seven GH families analysed here. Currently, in the LiveBench-6 experiment [30], in which 98 targets of varying degrees of difficulty are subjected to a variety of fold recognition methods, the highest three Pcons2 scores given to wrong folds are 2.17, 1.58 and 1.35, respectively. All the Pcons2 scores in Table 1 comfortably exceed the second worst incorrect Pcons2 score of 1.58 and approach or exceed the top scoring incorrect fold value of 2.17. The results of one of the most sensitive individual methods, 3D-PSSM [38], shown for comparison in Table 1, are also convincing. The idea that these seven families contain catalytic TIM barrels is additionally supported by the alternating helices and strands that are generally conspicuous in their predicted secondary structures (e.g. Fig. 1) and, most obviously, by the fact that all top scoring hits are other GHs. Typically, newly reported GH TIM barrels are more similar to other GH TIM barrels than to non-GH proteins sharing the same architecture [39].

In previous similar studies, PSI-BLAST has proved capable of demonstrating evolutionary links between GH families [15,16]. Here, this approach was only successful in the case of family 50. At the second iteration, using a strict *E*-value cut-off of 0.001, family 42 proteins appeared with significant scores of up to $4\times e^{-10}$. For all the other families no interfamily links could be demonstrated by this method, suggesting that they bear only distant relationships to better characterised families. There are relatively few experimental data available for the families in question that might support the TIM domain assignment. However, it is relevant to mention that the agarase (GH family 50) portion shown to have activity when heterologously expressed [41] contains the entire predicted TIM barrel domain.

3.2. Content of the common conserved regions

The length of the common conserved region of each family varied. In some cases it corresponded to a typical TIM barrel size while in other families additional domains were also invariably present in conjunction with the TIM barrel. These were identified through fold recognition experiments, both the original submissions of the entire common regions and in individual analyses of the regions in question.

Family 29 apparently contains only a catalytic TIM barrel in the common conserved region. Upstream of the TIM barrel match lie 30–110 residues in the different members of the family which are less well-conserved and unlikely to represent an additional domain. In contrast, family 44 members have a clear additional domain following the catalytic TIM domain. This contained seven predicted β-strands and was convincingly assigned to the Greek key fold, of unknown function, following the TIM barrel in α-amylases, α-*N*-acetylgalactosaminidase, oligo-1,6-glucosidase and some glucanotransferases (SCOP family b 71.1.1). The GH members of this group belong to families 13 and 27 in clans H and D, respectively, so it is interesting to see that family 44 seems to bear closest similarity to family 5 in clan A (Table 1). In family 71 there are around 100 residues in the common conserved region following the assigned TIM barrel. These were not convincingly aligned in initial submissions of the entire common conserved regions, nor in subsequence experiments with just this C-terminal portion. Nevertheless, the largely β-strand secondary structure predictions, size, and position are consistent with a Greek key fold, as seen for family 44.

In family 50, an additional domain is also present in the common conserved region, but this time preceding the TIM barrel. Separate fold recognition experiments suggested that this contains the Ig-like domain found at the C-terminus of bacterial chitobiases and variously positioned in other carbohydrate active enzymes (SCOP family b 1.1.5). In maltogenic amylase from GH family 13 this domain has been predicted to interact with substrate [40] and likely serves the same purpose in GH family 50.

In families 84 and 85 the TIM barrel domain comprises the complete common conserved region. In contrast, the common conserved region of family 89 matches convincingly all three domains of the family 67 structures giving best fold recognition results. Preceding the TIM barrel is an α+β domain of unknown function similar to that seen at the N-termini of GH family 20 chitobiases (SCOP d 92.2.1) while a largely helical domain (not yet in the SCOP database) follows the TIM domain and seems to function mainly to form the dimer interface [39].

3.3. Other domains

GH enzymes often show a modular structure [12] with catalytic domains found in combination with other domains, many binding carbohydrate. The patterns of domain combination often vary between different species. Searches in the PFAM [26], SMART [27] and CDD [28] databases were therefore carried out in order to analyse domain contents of the families considered here.

The simplest domain combinations are shown by GH family 71, whose members invariably contain only the catalytic

Table 1
GH families from the CAZY database containing newly discovered TIM barrels

CAZY family	Activities	Number in CAZY	Mech-anism	Representative[a] (limits of common conserved region)[b]	Domain contents of common conserved regions	Pcons2 results (value; PDB code; protein; GH family; GH clan)[c]	3D-PSSM results (value; PDB code; GH family; GH clan)[c]	Number of conserved acidic residues in common conserved region	Predicted catalytic residues (position in TIM barrel)[d]
29	α-L-fucosidase (EC 3.2.1.51)[e]	34	unknown	178409; *Homo sapiens*; α-L-fucosidase (28–393)	catalytic TIM barrel	2.58; 1ece; *Acidothermus cellulolyticus* endocullulase; family 5; clan A	0.007; 1ece; *A. cellulolyticus* endocullulase; family 5; clan A	2	**Asp225**
44	endoglucanase (EC 3.2.1.4)	7	inverting	144291; *Caldicellulosiruptor saccharolyticus*; β-mannanase/ endoglucanase[f] (785–1308)	catalytic TIM barrel followed by Greek key domain	1.94; 1qnr; *Trichoderma reesei* β-mannanase; family 5; clan A	0.001; 1ece; *A. cellulolyticus* endocullulase; family 5; clan A	18	**Glu964**, **Glu1148**
50	agarase (EC 3.2.1.81)	6	unknown	9946959; *Pseudomonas aeruginosa*; hypothetical protein (233–734)	catalytic TIM barrel preceded by Ig-like domain	7.25; 1kwg; *T. thermophilus* A4 β-galactosidase; family 42; clan A	8.6×10^{-10}; 1kwg; *T. thermophilus* A4 β-galactosidase; family 42; clan A	10	**Glu479**, **Glu641**
71	α-1,3-glucanase (EC 3.2.1.-)	7	unknown	27263094; *Penicillium funiculosum*; α-1,3-glucanase (38–431)	catalytic TIM barrel followed by possible Greek key domain	2.10; 1h1n; *Thermoascus aurantiacus* endoglucanase; family 5; clan A	0.009; 1bgl; *Escherichia coli* β-galactosidase; family 2; clan A	7	among Asp79, Asp105, Asp266 and Glu269
84	*N*-acetyl-β-glucosaminidase (EC 3.2.1.52), hyaluronidase (EC 3.2.1.35)	19	unknown	21110332; *Xanthomonas axonopodis*; conserved hypothetical protein (178–476)	catalytic TIM barrel	2.01; 1qbd; *Serratia marcescens* chitobiase; family 20; clan K	0.013; 1qbd; *S. marcescens* chitobiase; family 20; clan K	5	**Asp295**; **Asp296**
85	endo-β-*N*-acetyl-glucos-aminidase (3.2.1.96)	17	probably retaining	28269953; *Lactobacillus plantarum*; endo-β-*N*-acetyl-glucos-aminidase (56–385)	catalytic TIM barrel	2.10; 1d2k; *C. immitis* chitinase; family 18; clan K	0.21; 2ebn; *Flavobacterium meningosepticum* endo-β-*N*-acetyl-glucos-aminidase F1; family 18; clan K	2	**Glu222**; Asp276
89	α-*N*-acetyl-glucos-aminidase (3.2.1.50)[e]	11	unknown	4505327; *H. sapiens*; α-*N*-acetyl-glucos-aminidase (3–735)	catalytic TIM barrel preceded by α+β domain and followed by all-α domain	2.19; 1gqi; *Pseudomonas cellulosa* α-glucuronidase; family 67; no clan	0.022; 1gqk; *P. cellulosa* α-glucuronidase; family 67; no clan	6	among Asp312, Glu316 and Asp382

[a] The GenBank ID is followed by species name and enzyme activity.
[b] Obtained from PFAM in the case of GH family 29, from experimental data in the case of GH family 71 [60] and from visual inspection in the remaining cases.
[c] Best score obtained from analysis of maximally diverse three members of family and, where available, the CDD [28] consensus sequence.
[d] Bold indicates more confident assignments where the residue in question aligned with an experimentally determined catalytic residue of a known structure.
[e] Activities newly associated with the TIM barrel fold.
[f] Contains two catalytic domains [44].

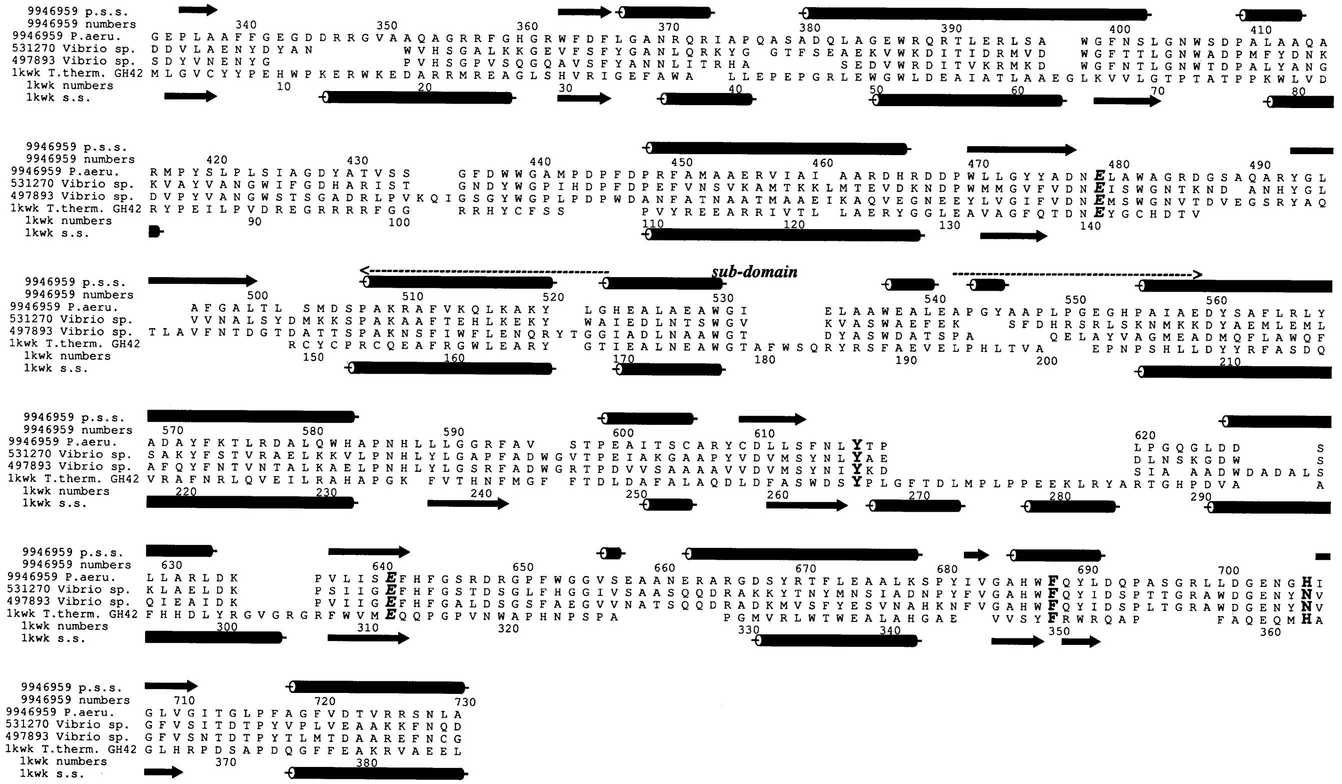
9946959 p.s.s.
9946959 numbers
9946959 P.aeru.
531270 Vibrio sp.
497893 Vibrio sp.
1kwk T.therm. GH42
1kwk numbers
1kwk s.s.
340 350 360 370 380 390 400 410
GEPLAAFFGEGDDRRGVAAQAGRRFGHGRWFDFLGANRQRIAPQASADQLAGEWRQRTLERLSA WGFNSLGNWSDPALAAQA
DDVLAENYDYAN WVHSGALKKGEVFSFYGANLQRKYG GTFSEAEKVWKDITIDRMVD WGFTTLGNWADPMFYDNK
SDYVNENYG PVHSGPVSQGQAVSFYANNLITRHA SEDVWRDITVKRMKD WGFNTLGNWTDPALYANG
MLGVCYYPEHWPKERWKEDARRMREAGLSHVRIGEFAWA LLEPEPGRLEWGWLDEAIATLAAEGLKVVLGTPTATPPKWLVD
10 20 30 40 50 60 70 80
420 430 440 450 460 470 480 490
RMPYSLPLSIAGDYATVSS GFDWWGAMPDPFDPRFAMAAERVIAI AARDHRDDPWLLGYYADNELAWAGRDGSAQARYGL
KVAYVANGWIFGDHARIST GNDYWGPIHDPFDPEFVNSVKAMTKKLMTEVDKNDPWMMGVFVDNEISWGNTKND ANHYGL
DVPYVANGWSTSGADRLPVKQIGSGYWGPLPDPWDANFATNAATMAAEIKAQVEGNEEYLVGIFVDNEMSWGNVTDVEGSRYAQ
RYPEILPVDREGRRRRFGG RRHYCFSS PVYREEARRIVTL LAERYGGLEAVAGFQTDNEYGCHDTV
90 100 110 120 130 140
sub-domain
500 510 520 530 540 550 560
AFGALTL SMDSPAKRAFVKQLKAKY LGHEALAEAWGI ELAAWEALEAPGYAAPLPGEGHPAIAEDYSAFLRLY
VVNALSYDMKKSPAKAAFTEHLKEKY WAIEDLNTSWGV KVASWAEFEK SFDHRSRLSKNMKKDYAEMLEML
TLAVFNTDGTDATTSPAKNSFIWFLENQRYTGGIADLNAAWGT DYASWDATSPA QELAYVAGMEADMQFLAWQF
RCYCPRCQEAFRGWLEARY GTIEALNEAWGTAFWSQRYRSFAEVELPHLTVA EPNPSHLLDYYRFASDQ
150 160 170 180 190 200 210
570 580 590 600 610 620
ADAYFKTLRDALQWHAPNHLLLGGRFAV STPEAITSCARYCDLLSFNLYTP LPGQGLDD S
SAKYFSTVRAELKKVLPNHLYLGAPFADWGVTPEIAKGAAPYVDVMSYNLYAE DLNSKGDW S
AFQYFNTVNTALKAELPNHLYLGSRFADWGRTPDVVSAAAAVVDVMSYNIYKD SIA AADWDADALS
VRAFNRLQVEILRAHAPGK FVTHNFMGF FTDLDAFALAQDLDFASWDSYPLGFTDLMPLPPEEKLRYARTGHPDVA A
220 230 240 250 260 270 280 290
630 640 650 660 670 680 690 700
LLARLDK PVLISEFHFGSRDRGPFWGGVSEAANERARGDSYRTFLEAALKSPYIVGAHWFQYLDQPASGRLLDGENGHI
KLAELDK PSIIGEFHFGSTDSGLFHGGIVSAASQQDRAKKYTNYMNSIADNPYFVGAHWFQYIDSPTTGRAWDGENYNV
QIEAIDK PVIIGEFHFGALDSGSFAEGVVNATSQQDRADKMVSFYESVNAHKNFVGAHWFQYIDSPLTGRAWDGENYNV
FHHDLYRGVGRGRFWVMEQQPGPVNWAPHNPSPA PGMVRLWTWEALAHGAE VVSYFRWRQAP FAQEQMHA
300 310 320 330 340 350 360
710 720 730
GLVGITGLPFAGFVDTVRRSNLA
GFVSITDTPYVPLVEAAKKFNQD
GFVSNTDTPYTLMTDAAREFNCG
GLHRPDSAPDQGFFEAKRVAEEL
370 380

Fig. 1. Alignment of three diverse GH 50 family members (labelled with GenBank number and species) with the GH family 42 enzyme *Thermus thermophilus* β-galactosidase (PDB code 1kwk). PSI-PRED [29] derived predicted secondary structure and numbering for the representative GH family 50 protein (see Table 1) are shown above the alignment while 1kwk numbering and STRIDE-defined [54] secondary structure are shown below. The subdomain inserted into the TIM barrel is labelled. The figure was made with ALSCRIPT [55].

TIM domain and possible Greek key domain, as described above. Similarly, for family 29, with a single exception, just the catalytic TIM domain is present. The exception is the *Clostridium perfringens* homologue (GenBank ID18145540) which contains a Calx-β domain (PFAM entry PF03160) whose function, even in the eukaryotes in which it was first described, remains obscure [42].

Within GH family 44, dockerin domains (PFAM entry PF00404) are found in both clostridial enzymes and known carbohydrate binding domains of two kinds (cellulose binding domain and PKD domain; PFAM entries PF00942 and PF00801, respectively) in others. A novel carbohydrate binding domain has also been identified in a cellulase from *Ruminococcus flavefaciens* [43]. Most interesting are two different combinations of separate catalytic domains, one, with a GH family 5 [44] and one with GH family 9 [45]. In contrast, no domains could be located in any family 50 members, although some of these sequences contain almost 1000 residues. Evidently the non-TIM portions of family 50 contain divergent examples of known domains or novel domains.

Again with family 84, additional domains with probable carbohydrate binding function are found (F5/8 type C domain (structurally related to galactose binding domains – see [16]), and FnIII domain; PFAM entries PF00754 and PF00041, respectively). Additionally, many family 84 members contain the same α+β domain, originally seen in GH families 20 and 67, and invariably present in GH family 89 (see above). In family 85, three kinds of presumed carbohydrate binding domains are represented (F5/8 type C, FnIII and PKD domains), invariably C-terminal to the catalytic TIM domain. Although some members of GH family 89 are very large (up to 2100 residues) recognisable additional domains are less in evidence. Only the *C. perfringens* sequence contains two clear carbohydrate binding domains (F5/8 type C and FnIII).

3.4. Predicted catalytic and binding residues

Several sources of information were used to help locate probable catalytic residues, invariably Glu or Asp in GHs. Firstly, it is reasonable to consider only positions at which a Glu or Asp is entirely conserved, although the possibility exists of non-catalytic members within a given GH family [10] and of non-transcribed genome sequences, conceivably having accumulated mutations, being included in the sequence databases. Secondly, the alignments resulting from fold recognition or sequence searches may align known catalytic residues in determined structures with conserved acidic residues in the family being analysed, provided the evolutionary relationship between aligned sequences is sufficiently close. Thirdly, catalytic sites in TIM barrels are invariably located at the C-terminal end of the barrel [46] so that putative catalytic residues should, allowing for errors in predicted secondary structure, appear towards the end of predicted β-strands or in the loops which follow them. Fourthly, GH TIM domains seem to invariably place one catalytic residue at the end of β-strand 4. The families considered here contain between two and 18 (mean seven) conserved acidic residues (Table 1) so that experimental determination of catalytic residues through their systematic mutation would be rather laborious.

The most reliably predicted catalytic residues are those in GH family 50, which could be linked to other GH families in

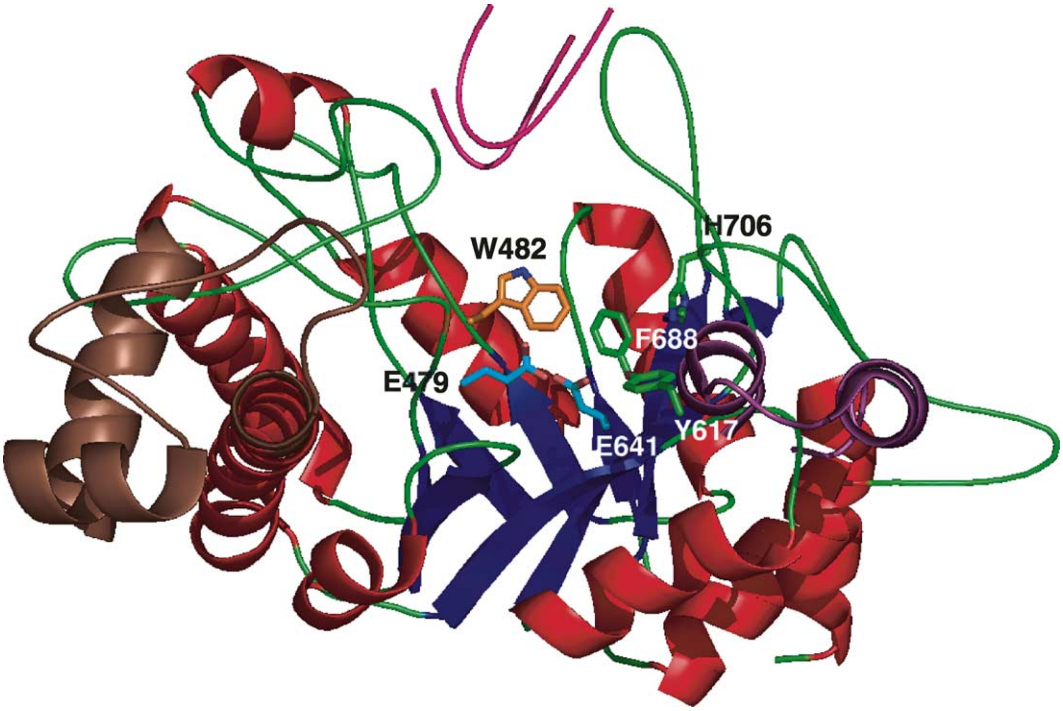

Fig. 2. Pymol [56] figure of the final model of the TIM barrel of the representative GH family 50 protein viewed along the substrate binding cleft. The cartoon is coloured by secondary structure and the subdomain, towards the left, coloured brown. Stick representations are shown for predicted catalytic acidic residues (cyan), the hydrophobic platform ([8]; green) and the conserved Trp482 (orange) which presumably binds substrate. Also shown at the right and coloured violet are the 17-residue loops, present in the template from GH family 42, but deleted in GH family 50, whose absence broadens the near side of the cleft in the latter family. The loops from the alternate subunit in the GH family 42 trimer which block the far side of the cleft, but which are not predicted to be present in GH family 50, are shown and coloured magenta.

clan A using PSI-BLAST. However, the better alignments, matching catalytic residues and predicted with actual secondary structure more satisfactorily, were obtained from fold recognition rather than from PSI-BLAST. In this case the alignments (Fig. 1) were deemed to be of sufficient quality for the construction of a useful molecular model. Indeed, the pG [47] score of the model – 0.98 – approached the maximum attainable 1.0, despite the GH family 50 representative and the GH42 template sharing only 18% sequence identity. Stereochemical quality was also good with the model placing 89% of residues in most favoured regions of the Ramachandran plot, and possessing only a single disallowed residue, corresponding to an unusually placed residue in the template. As shown in Fig. 2, the final structure illustrates conservation, between model and template, not only of catalytic residues but also of the transition state-stabilising hydrophobic platform, apparently ubiquitous among GHs [8], although one of the three platform positions varies in other members of the family (Fig. 1). A conserved Trp, numbered 482 in the representative sequence (see Table 1), borders the catalytic cleft and is well placed to interact with substrate, as often observed for aromatic residues in carbohydrate binding proteins [48]. Interestingly, at one end of the catalytic cleft, a 17-residue deletion at position 619 relative to the template of family 42, lead to a more open cleft in the model structure (Fig. 2). This may be related to the bulky, double helical structure of agarose, the substrate of GH family 50 [49]. It would presumably lead to catalytic advantage if only the portion to be cleaved, binding at the centre of the cleft, had to be unwound from its helical structure and that, outside this region, the enzyme had a more open structure capable of accommodating the intact helix. The fact that GH family 50 lacks the C-terminal domain of the template GH family 42 structure that contributes around 45% of the trimer interface means that GH family 50 is unlikely to exist as a similar trimer. Thus, the blockage of one end of the catalytic cleft, responsible for the exo-style β-galactosidase activity in GH42, is not present. The unimpeded catalytic cleft predicted by modelling (Fig. 2) is in accord with the predominance of larger tetraose and hexaose fragments among the products of GH family 50 agarases [22,41]. GH family 42, the closest neighbour of GH family 50 (Table 1), share an unexpectedly close structural similarity to family 14 β-amylases which, in contrast to family 42, catalyse hydrolysis with inversion of configuration [49]. Families 42 and 14 share a subdomain, not present in other TIM barrels of clan A, but have evolved pocket-type catalytic sites, from the presumed ancestral cleft-type site, through different mechanisms – trimerisation and loop addition, respectively [50]. The proposed characteristics of the shared ancestor are therefore monomericity, a cleft-type site and similar subdomain to GH families 42 and 14. These considerations are in line with known (Figs. 1 and 2) and predicted characteristics (Fig. 2) of family 50, making it an attractive candidate for shared ancestor of families 42 and 14, and a particularly interesting case for structural determination.

Other cases in which reliable catalytic residue assignments could be made were GH families 84 and 85, both matched to TIM barrels of clan K. For both families, 18 and 20, making up this clan, there is now clear evidence that, unlike the majority of GHs, catalysis proceeds with the involvement of a single catalytic Glu residue, acting as proton donor, and the acetamido group of the substrate acting as nucleophile [6,7]. A

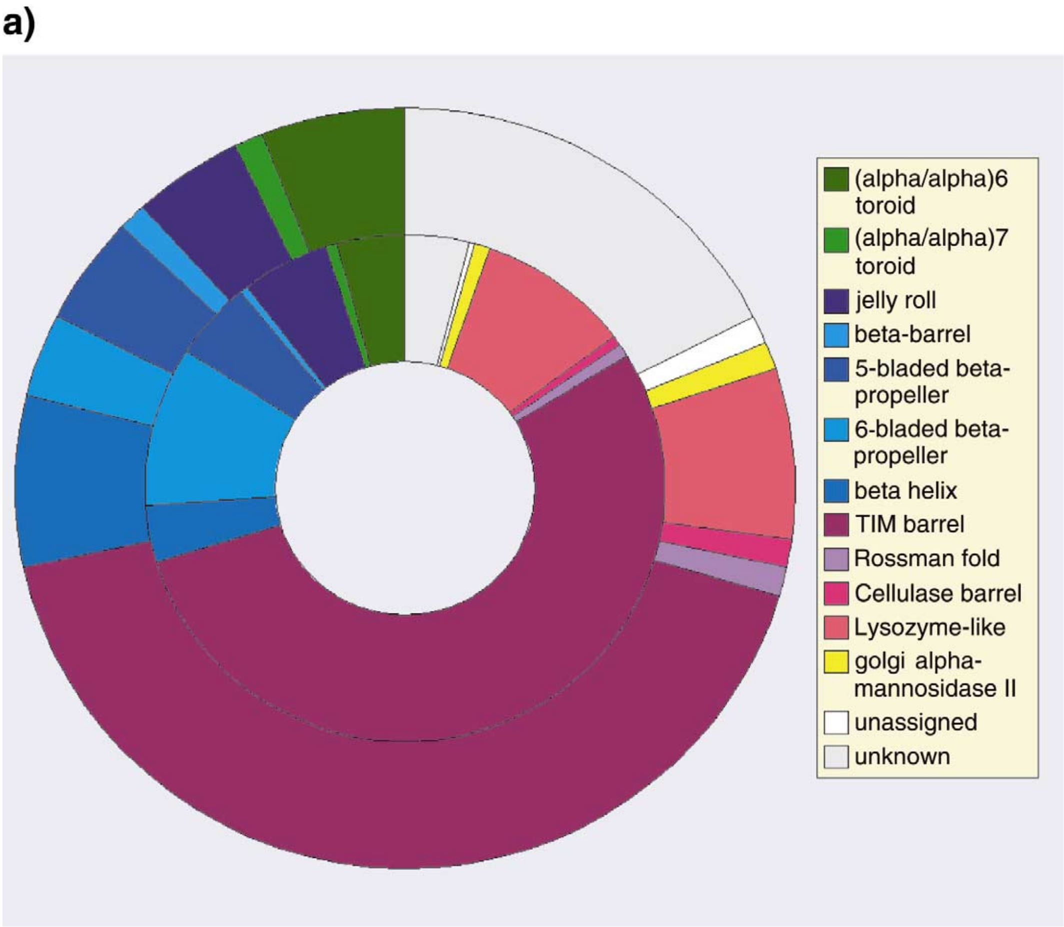

Fig. 3. a: SCOP-defined [37] fold distribution among GH families (outer ring) and sequences (inner ring). b: Number of EC classified GH catalytic activities associated with each fold. Within SCOP, TIM barrels include the distorted barrels present in family 56 [57]. In family 3, the catalytic site lies between the TIM barrel and a second domain [58]. Family 67 GH structures are not yet in SCOP but may be confidently included among TIM barrels from their published description [39]. Family 57 structures are known [59] but are not yet available and no confident fold assignment can be made.

conserved Asp, either one residue (GH family 20) or two residues before the conserved Glu plays a key role, amongst other possible functions, in stabilisation of the positively charged intermediate [51]. In family 84 a completely conserved Asp–Asp pair is aligned with the catalytic Asp–Glu of GH family 20 and presumably represents the catalytic acidic pair. In family 85, a conserved Glu aligns with the proton donor identified in GH family 18. However, the conserved Asp two residues prior to this in GH family 18 is aligned with a conserved Asn in family 85. Mutation of this Asp in *Coccidioides immitis* chitinase leads to complete loss of activity [51], suggesting that, although the Asn replacement might be capable of taking over the additional orientation functions assigned to the corresponding Asp in GH family 18, stabilisation of the positively charged intermediate should be provided by a differently placed conserved acidic residue in GH family 85. In this case, only one other totally conserved acidic residue is present in all members of the family – Asp276. However, its positioning towards the end of predicted strand 5 of the TIM barrel is consistent with a proposed catalytic role. In assigning a substrate-assisted catalytic mechanism to GH families 84 and 85, a requirement for a acetamido (or similar) group on the substrate is introduced. Indeed, all the known substrates of families 84 and 85 (Table 1) possess such a group. Furthermore, this predicted mechanism is in line with data suggesting a retaining mechanism for GH family 85 [3]. Once confirmed, the mechanism predicted here would enable the use of mechanism-based inhibitors, already of proven effectiveness against other GH families (e.g. [52]), against the possible virulence gene products secreted by *C. perfringens* and *Enterococcus faecium* [23,24].

Three of the families considered here matched most closely to GH family 5. The clearest of these is family 44, where conserved Glu residues (see Table 1) align well with both conserved Glu residues of GH family 5. In GH family 29, completely conserved Asp228 aligned with catalytic residues situated after β-strand 4 of the TIM barrel. However, the only other completely conserved acidic residue, Asp158, lies within a predicted helix and therefore seems not to be catalytic. In this case, other acidic catalytic residues must lie among non-conserved positions. For GH family 71, no conserved acidic residues could be aligned with catalytic residues of known structures. Therefore the four candidates in Table 1 are again those suitable placed relative to predicted secondary structure. Also less strongly predicted are the likely catalytic residues for GH family 89, where different alignments with other TIM barrels produced different registers of actual and predicted secondary structure elements. Nevertheless, only three of the

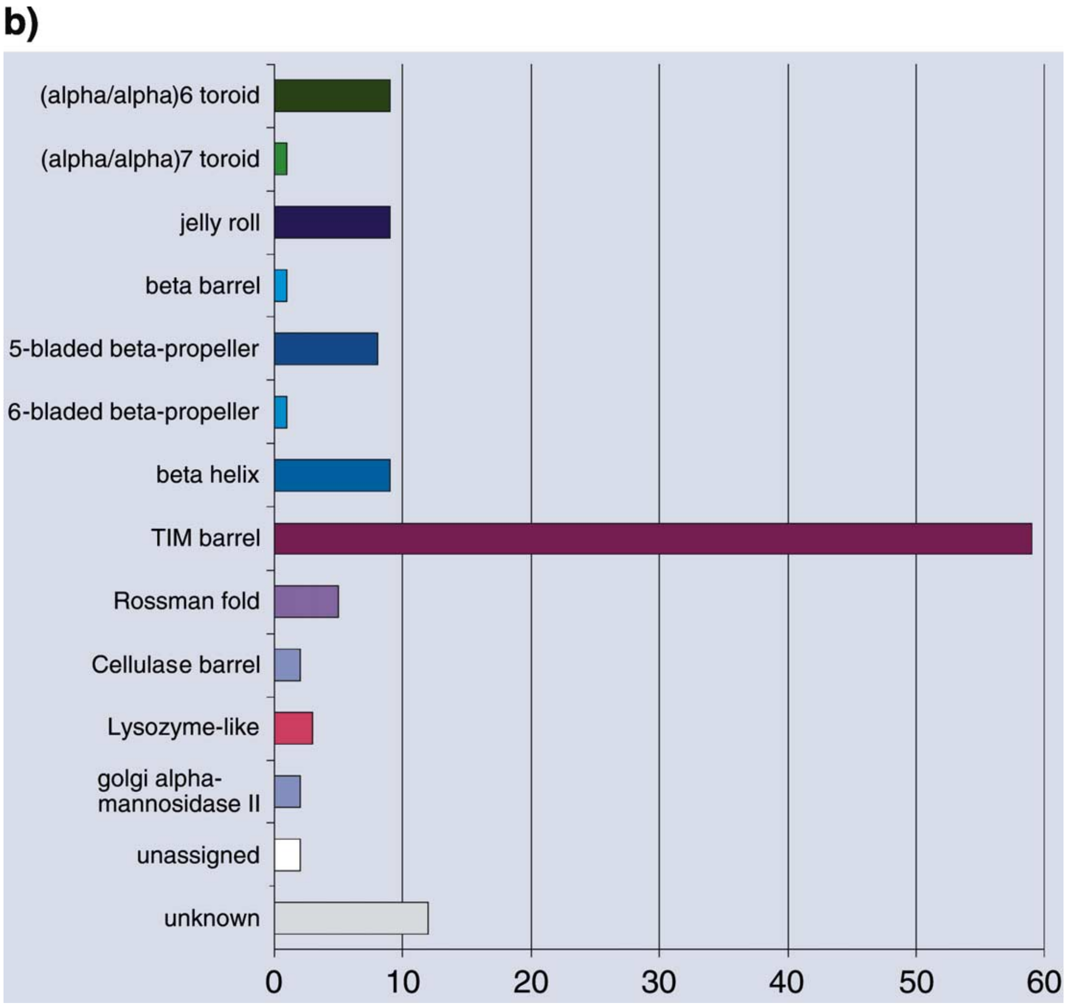

Fig. 3 (*Continued*).

six conserved acidic residues are suitable positioned towards the end of predicted β-strands or in the following loops (Table 1).

3.5. *Distribution of GH families and catalytic activities among folds*

The recent demonstrations, both crystallographic and computational ([50,39,15], this work), of catalytic TIM barrels for more GH families offer an opportunity to reexamine the relationship between GH activity and the TIM barrel. While TIM barrels catalyse activities in five of the six major divisions of the EC classification, around half of known TIM barrels are hydrolases (EC numbers 3.x.x.x; [10]) and most of these are GHs (EC 3.2.1.x; [10]). Using the present CAZY database as a foundation and adding in families with catalytic domains reliably computationally identified ([14–16], this work), the fold distribution of GH families and sequences, employing the SCOP classification [37], can be analysed (Fig. 3a). GH activity is present in five of the seven principal SCOP classes, these being, in order of abundance, α/β > all-β > α+β > all-α > multidomain. Even allowing for possible distortions caused by some families having received more attention than others, the contribution of TIM barrels is impressive. There are 86 GH families in the current CAZY database, of which 71 have folds assigned to their catalytic domains, either through crystallography or through computational studies. Of the 71 structurally assigned families, 36 are now seen to contain TIM barrels. 54% of known GH sequences (57% of those for which catalytic domain structure is known) contain TIM barrels. Evidently, the contribution of TIM barrels to GH activity is even more pronounced than previously believed [10]. Fig. 3a shows that a reasonably complete structural perspective of GH families is therefore already available. When analysed by sequence instead of by families, only a relatively small number (327, 3.8%) of GHs are of unknown catalytic domain fold. As Fig. 3b shows, TIM barrels also catalyse by far the largest range of GH reactions. TIM barrels catalyse around two thirds of the catalogued GH reactions in the EC classification – 59 out of 89 numbers.

The striking contribution of TIM barrels to known total GH activity is thought-provoking. Does the TIM barrel have particular characteristics that have led to its being favoured to such an extent? In this regard, ability to bind diverse substrates through modulation of the loops at the C-terminal end of the barrel and through recruitment of other domains is already evident [9]. The various mechanisms [50] by which a ‘cleft’-type site may be transformed into a ‘pocket’-type site (thereby producing exo rather than endo activity) are also worth noting. However, it is difficult to imagine that other folds are not capable of such adaptation. Alternatively, is the TIM barrel simply an ancient domain in which GH activity appeared early, so that the present day diversity just reflects elapsed evolutionary time? In order for this to be true, it is necessary to prove evolutionary links between most or all of GH TIM barrels to be able to discard the alternative hypothesis that TIM barrels arose independently several times. It must be remembered that repetitive structures such as the TIM barrel may be particularly easy to evolve through duplication and fusion of short peptide ancestors [53]. In the easiest cases, evolutionary relationships between GH families may be apparent from more sensitive sequence comparisons to the BLAST routinely used to assign families [10]. Structural comparisons may also reveal similarities at key points in the TIM barrel indicative of an evolutionary relationship [10]. The results of a recent study applying such sequence and structure analyses were insufficient to prove a single evolutionary origin for the TIM barrel [10]. A more definitive answer to the question of evolutionary relatedness will likely be accessible after the structural determination of representatives of more GH families. It is worth noting that computational analyses of the kind presented here assist this process in several ways – firstly by highlighting particularly interesting families that may be worthy of priority treatment (such as GH family 50 mentioned here), secondly by defining limits for the domains of interest, thereby facilitating their functional, modular characterisation, and thirdly by highlighting related structures that may aid in crystallographic structure solution by molecular replacement.

3.6. *Conclusions*

The data presented here show the value of computational exploration using fold recognition tools, especially in the context of reliable, hierarchical databases such as CAZY. The information that can be extracted from these fold recognition results depends on the closeness of the structural relationships identified, but valuable data regarding domain limits, domain composition, catalytic and binding residues and evolutionary links were all forthcoming from the analyses reported here. These new fold assignments reveal that TIM barrels are the catalytic domains for more than half the known current GH families and sequences. Patterns evident among known GH TIM barrel structures are largely suggestive of a single evolutionary ancestor, but further evolutionary links are required to bridge gaps and demonstrate relationships that are not currently reliably established. Computational studies will help determine the most interesting targets and facilitate their study, thereby enabling a more rapid arrival at a solution to the question of GH TIM barrel evolutionary relatedness.

References

[1] Henrissat, B. (1991) Biochem. J. 280, 309–316.

[2] Henrissat, B. and Bairoch, A. (1993) Biochem. J. 293, 781–788.

[3] Coutinho, P.M. and Henrissat, B. (1999) Carbohydrate-Active Enzymes server at URL: http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html.

[4] Jedrzejas, M.J. (2000) Crit. Rev. Biochem. Mol. Biol. 35, 221–251.

[5] McCarter, J.D. and Withers, S.G. (1994) Curr. Opin. Struct. Biol. 4, 885–892.

[6] Terwisscha van Scheltinga, A.C., Armand, S., Kalk, K.H., Isogai, A., Henrissat, B. and Dijkstra, B.W. (1995) Biochemistry 34, 15619–15623.

[7] Mark, B.L., Vocadlo, D.J., Knapp, S., Triggs-Raine, B.L., Withers, S.G. and James, M.N. (2001) J. Biol. Chem. 276, 10330–10337.

[8] Nerinckx, W., Desmet, T. and Claeyssens, M. (2003) FEBS Lett. 538, 1–7.

[9] Bourne, Y. and Henrissat, B. (2001) Curr. Opin. Struct. Biol. 11, 593–600.

[10] Nagano, N., Porter, C.T. and Thornton, J.M. (2001) Protein Eng. 14, 845–855.

[11] Ciccarelli, F.D., Copely, R.R., Doerks, T., Russell, R.B. and Bork, P. (2002) Trends Biochem. Sci. 27, 59–62.

[12] Henrissat, B. and Davies, G.J. (2000) Plant Physiol. 124, 1515–1519.

[13] Davies, G.J. and Henrissat, B. (2002) Biochem. Soc. Trans. 30, 291–297.

[14] Naumoff, D.G. (2001) Proteins 42, 66–76.

[15] Rigden, D.J. (2002) FEBS Lett. 523, 17–22.

[16] Rigden, D.J. and Franco, O.L. (2002) FEBS Lett. 530, 225–232.

[17] Michalski, J.C. and Klein, A. (1999) Biochim. Biophys. Acta 1455, 69–84.

[18] Zhao, H.G., Li, H.H., Bach, G., Schmidtchen, A. and Neufeld, E.F. (1996) Proc. Natl. Acad. Sci. USA 93, 6101–6105.

[19] Suzuki, T., Yano, K., Sugimoto, S., Kitajima, K., Lennarz, W.J., Inoue, S., Inoue, Y. and Emori, Y. (2002) Proc. Natl. Acad. Sci. USA 99, 9691–9696.

[20] Muller-Taubenberger, A., Westphal, M., Noegel, A. and Gerisch, G. (1989) FEBS Lett. 246, 185–192.

[21] Wei, H., Scherer, M., Singh, A., Liese, R. and Fischer, R. (2001) Fungal Genet. Biol. 34, 217–227.

[22] Sugano, Y., Matsumoto, T., Kodama, H. and Noma, M. (1993) Appl. Environ. Microbiol. 59, 3750–3756.

[23] Canard, B., Garnier, T., Saint-Joanis, B. and Cole, S.T. (1994) Mol. Gen. Genet. 243, 215–224.

[24] Rice, L.B., Carias, L., Rudin, S., Vael, C., Goossens, H., Konstabel, C., Klare, I., Nallapareddy, S.R., Huang, W. and Murray, B.E. (2003) J. Infect. Dis. 187, 508–512.

[25] Notredame, C., Higgins, D.G. and Heringa, J. (2000) J. Mol. Biol. 302, 205–217.

[26] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) Nucleic Acids Res. 30, 276–280.

[27] Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P. and Bork, P. (2002) Nucleic Acids Res. 30, 242–244.

[28] Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J. and Bryant, S.H. (2003) Nucleic Acids Res. 31, 383–387.

[29] Jones, D.T. (1999) J. Mol. Biol. 292, 195–202.

[30] Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Bioinformatics 17, 750–751.

[31] Rychlewski, L., Jaroszewski, L., Li, W. and Godzik, A. (2000) Protein Sci. 9, 232–241.

[32] Bujnicki, J.M., Elofsson, A., Fischer, D. and Rychlewski, L. (2001) Proteins 45 (Suppl. 5), 184–191.

[33] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.

[34] Sali, A. and Blundell, T.L. (1993) J. Mol. Biol. 234, 779–815.

[35] Sippl, M.J. (1993) Proteins 17, 335–362.

[36] Laskowski, R., Macarthur, M., Moss, D. and Thornton, J. (1993) J. Appl. Crystallogr. 26, 283–290.

[37] Lo Conte, L., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2002) Nucleic Acids Res. 30, 264–267.

[38] Kelley, L.A., MacCallum, R.M. and Sternberg, M.J.E. (2000) J. Mol. Biol. 299, 499–520.

[39] Nurizzo, D., Nagy, T., Gilbert, H.J. and Davies, G.J. (2002) Structure 10, 547–556.

[40] Kim, J.S., Cha, S.S., Kim, H.J., Kim, T.J., Ha, N.C., Oh, S.T.,

Cho, H.S., Cho, M.J., Kim, M.J., Lee, H.S., Kim, J.W., Choi, K.Y., Park, K.H. and Oh, B.H. (1999) J. Biol. Chem. 274, 26279–26286.

[41] Sugano, Y., Matsumoto, T. and Noma, M. (1994) Biochim. Biophys. Acta 1218, 105–108.

[42] Schwarz, E.M. (1997) Proc. Natl. Acad. Sci. USA 94, 10249–10254.

[43] Rincon, M.T., McCrae, S.I., Kirby, J., Scott, K.P. and Flint, H.J. (2001) Appl. Environ. Microbiol. 67, 4426–4431.

[44] Gibbs, M.D., Saul, D.J., Luthi, E. and Bergquist, P.L. (1992) Appl. Environ. Microbiol. 58, 3864–3867.

[45] Ahsan, M.M., Kimura, T., Karita, S., Sakka, K. and Ohmiya, K. (1996) J. Bacteriol. 178, 5732–5740.

[46] Wierenga, R.K. (2001) FEBS Lett. 492, 193–198.

[47] Sanchez, R. and Sali, A. (1998) Proc. Natl. Acad. Sci. USA 95, 13597–13602.

[48] Quicho, F.A. and Vyas, N.K. (1999) in: Bioinorganic Chemistry: Carbohydrates, pp. 441–457, Oxford University Press, New York.

[49] Arnott, S., Fulmer, A., Scott, W.E., Dea, I.C., Moorhouse, R. and Rees, D.A. (1974) J. Mol. Biol. 90, 269–284.

[50] Hidaka, M., Fushinobu, S., Ohtsu, N., Motoshima, H., Matsuzawa, H., Shoun, H. and Wakagi, T. (2002) J. Mol. Biol. 322, 79–91.

[51] Bortone, K., Monzingo, A.F., Ernst, S. and Robertus, J.D. (2002) J. Mol. Biol. 320, 293–302.

[52] Houston, D.R., Eggleston, I., Synstad, B., Eijsink, V.G. and van Aalten, D.M. (2002) Biochem. J. 368, 23–27.

[53] Lupas, A.N., Ponting, C.P. and Russell, R.B. (2001) J. Struct. Biol. 134, 191–203.

[54] Frishman, D. and Argos, P. (1995) Proteins 23, 566–579.

[55] Barton, G.J. (1993) Protein Eng. 6, 37–40.

[56] DeLano, W.L. (2002) The PyMOL Molecular Graphics System on World Wide Web http://www.pymol.org.

[57] Markovic-Housley, Z., Miglierini, G., Soldatova, L., Rizkallah, P.J., Muller, U. and Schirmer, T. (2000) Structure 8, 1025–1035.

[58] Varghese, J.N., Hrmova, M. and Fincher, G.B. (1999) Structure 7, 179–190.

[59] Imamura, H., Fushinobu, S., Yamamoto, M., Kumasaka, T., Jeon, B.S., Wakagi, T. and Matsuzawa, H. (2003) J. Biol. Chem., in press.

[60] Fuglsang, C.C., Berka, R.M., Wahleithner, J.A., Kauppinen, S., Shuster, J.R., Rasmussen, G., Halkier, T., Dalboge, H. and Henrissat, B. (2000) J. Biol. Chem. 275, 2009–2018.