Iterative database searches demonstrate that glycoside hydrolase families 27, 31, 36 and 66 share a common evolutionary origin with family 13

Daniel J. Rigden*

Embrapa Genetic Resources and Biotechnology, Cenargen/Embrapa, S.A.I.N. Parque Rural, Final W5, Asa Norte, 70770-900 Brasília, Brazil

Received 28 March 2002; revised 17 May 2002; accepted 22 May 2002

First published online 20 June 2002

Edited by Robert B. Russell

Abstract Classification of glycoside hydrolases (GHs) into families, along with the structure-based grouping together of families into clans, improve our understanding of the evolution of the large natural variety of these enzymes, help rationalise experimental data and guide further studies. Here we identify triose phosphate isomerase (TIM) barrels in GH families 27, 31, 36 and 66. We further show that iterated sequence database searches provide evidence for their sharing a common evolutionary origin with GH family 13. The catalytic, nucleophilic residue common to all these families is thereby determined and candidate catalytic proton donors identified within each family. © 2002 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

Key words: Glycosidase; Iterative database search; Clan; Evolutionary relationship

1. Introduction

The remarkable variety of naturally occurring carbohydrates and glycoconjugates has produced a correspondingly large diversity in the enzymes which act upon them. In response, a glycoside hydrolase (GH) classification system was introduced and developed [1,2] which is now available as a CAZY public WWW resource [3]. In contrast to the more general EC classification [4] which groups enzymes solely by catalytic activity, the CAZY classification also takes into account sequence and structural relationships thereby grouping together enzymes of common evolutionary origin, irrespective of differences in the reactions that they catalyse. Categorisation of enzymes in this way enables inferences to be drawn for entire protein families based on experimental study of a few members. Important characteristics to be determined are mechanism (retaining or inverting) and the identity of the two acidic groups near-universally involved in the catalysis of glycoside bond cleavage [5].

Within the CAZY system there are now more than 80 families defined on the basis of sequence similarity [3]. With the ongoing determinations of GH structures [6], structural similarities occasionally reveal evolutionary relationships between different families that were not apparent from simple sequence analysis. These data allow CAZY families to be grouped together into higher-level clans, improving our understanding of GH evolution and enabling testable hypotheses regarding key catalytic residues to be made from cross-family comparison. These benefits have stimulated several in silico studies, utilising sensitive sequence comparisons or fold recognition tools, revealing distant relationships between GH families [7–10]. In the current CAZY database, 12 clans from GH-A to GH-L are defined, eight of known overall protein fold, within which five different architectures are represented – the $(\beta/\alpha)_8$ triose phosphate isomerase (TIM) barrel, the $(\alpha/\alpha)_6$ toroid, the β-propeller, the β-jelly roll and an α+β fold.

Clan GH-H contains the TIM barrels of GH families 13, 70 and 77. By far the largest and best studied of these families is family 13 which contains α-amylase and other catalytic activities related to hydrolysis or transglycosylation of α-linked glucans [11]. Clan GH-D contains families 27 and 36 [12] while many other families, including numbers 31 and 66 are not yet grouped into clans. Here we show that iterated PSI-BLAST database searches reveal a distant evolutionary kinship of GH family 13 with families 27, 31, 36 and 66 which catalyse diverse reactions (Table 1). The similarity centres on a single nucleophilic Asp residue, but each of these four families clearly contains an entire TIM barrel. This knowledge helps locate candidate proton donors within each family.

2. Materials and methods

Members of GH families 27, 31, 36 and 66 were located in the CAZY database and retrieved using Entrez (http://www.ncbi.nlm.nih.gov:80/entrez) yielding sets of 51, 106, 43 and six sequences, respectively. Groups of sequences were aligned with T-Coffee [13], where the programme's capacity allowed, and otherwise with CLUSTALW [14]. Manipulation and limited hand-editing of alignments were performed with Jalview (http://www.ebi.ac.uk/∼michele/jalview/), as was the determination of the four maximally diverse representatives of each family. Intra- and inter-family sequence motifs were sought using MEME [15]. The seed alignment of family 13 in the PFAM database [16] was used for the calculation of family 13 sequence motifs. Secondary structure prediction was carried out using PSI-PRED [17]. Fold recognition experiments made use of the Structure Prediction META server [18]. The principal indicator used to measure success of fold recognition was the 'Shotgun on 3' consensus prediction (D. Fischer, unpublished) which produces a score based on the results of three independent fold recognition methods, FFAS [19], 3D-PSSM [20] and Inbgu [21], and is currently the method that differentiates best between true and false positives (see http://bioinfo.pl/LiveBench/ [22]). Iterated sequence database searches were carried out using PSI-BLAST [23] at the NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) and PDB-BLAST servers (http://bioinformatics.burnham-inst.org/pdb_blast/) and using either 0.01 or 0.001 as the *E*-value cut-off below which a sequence is included in the next iteration. Ap-

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

PII: S0014-5793(02)02879-X

pearance of a member of a given GH family in the list of sequences resulting from a search using a different GH family was taken as an indication of possible common evolutionary origin for the two families. As input for the fold recognition and iterated database searches we used consensus sequences for families 27 and 31 (obtained from the CDD database [24]) and used *Escherichia coli* α-galactosidase (SwissProt code P16551) and *Streptococcus salivarius* dextranase (SwissProt code Q59979) as representatives of families 36 and 66, respectively. A model of the TIM barrel of the family 31 representative was constructed using MODELLER [25].

3. Results and discussion

Application of fold recognition methods was appropriate in the cases of GH families, since these methods are capable of assigning folds to sequences, even in the absence of significant sequence identity, through exploitation of derived characteristics such as predicted solvent exposure. Fold recognition is therefore capable of detecting cases of distant homology, where a structural similarity is maintained but sequence comparisons alone are incapable of demonstrating the evolutionary relationship.

Examination of the fold recognition results obtained for the consensus sequences of GH families 27 and 31 and the representative sequences of families 36 and 66 immediately revealed the presence of an α-amylase-like TIM barrel in each family. The consensus prediction scores for the four sequences were 59.9, 100.0, 49.9 and 69.7, respectively. These values are well in excess of the score of the highest false positive yet demonstrated for this method (40 – see http://bioinfo.pl/LiveBench/ [22]). In each case the structure to which the sequence matched was an α-amylase or isoamylase from GH family 13. The scores for the three individual methods contributing to the consensus predictions were generally also significant for GH family 13 proteins. GH family 13 proteins were the top hits except where noted. For FFAS [19] the scores for family 27 and 31 consensuses and representatives of families 36 and 66 were 11.2, 8.1, 7.0 and 5.15, respectively. By 3D-PSSM [20] the same sequences gave 0.019, 0.00074, 0.013 (second place) and 0.113 (second place), while for INBGU methods [21], the respective results were 8.3 (fourth place), 38.9, 44.9 and 27.3.

With relatively common protein folds such as the TIM barrel, there is occasionally doubt as to whether a resemblance between two given structures is the result of divergent evolution from a common ancestor or convergent evolution [26]. Since significant sequence similarity is indicative of the former, we carried out sensitive iterated sequence database searches to search for common evolutionary origins among these GH families 13, 27, 31, 36 and 66. Families 70 and 77 have already been shown to be related to family 13, forming clan H [3], and were not considered further.

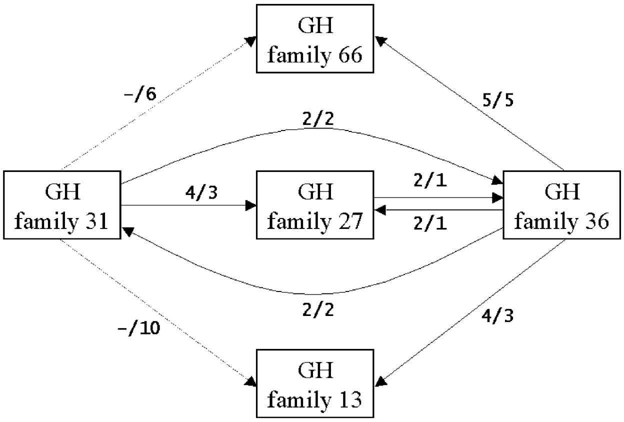


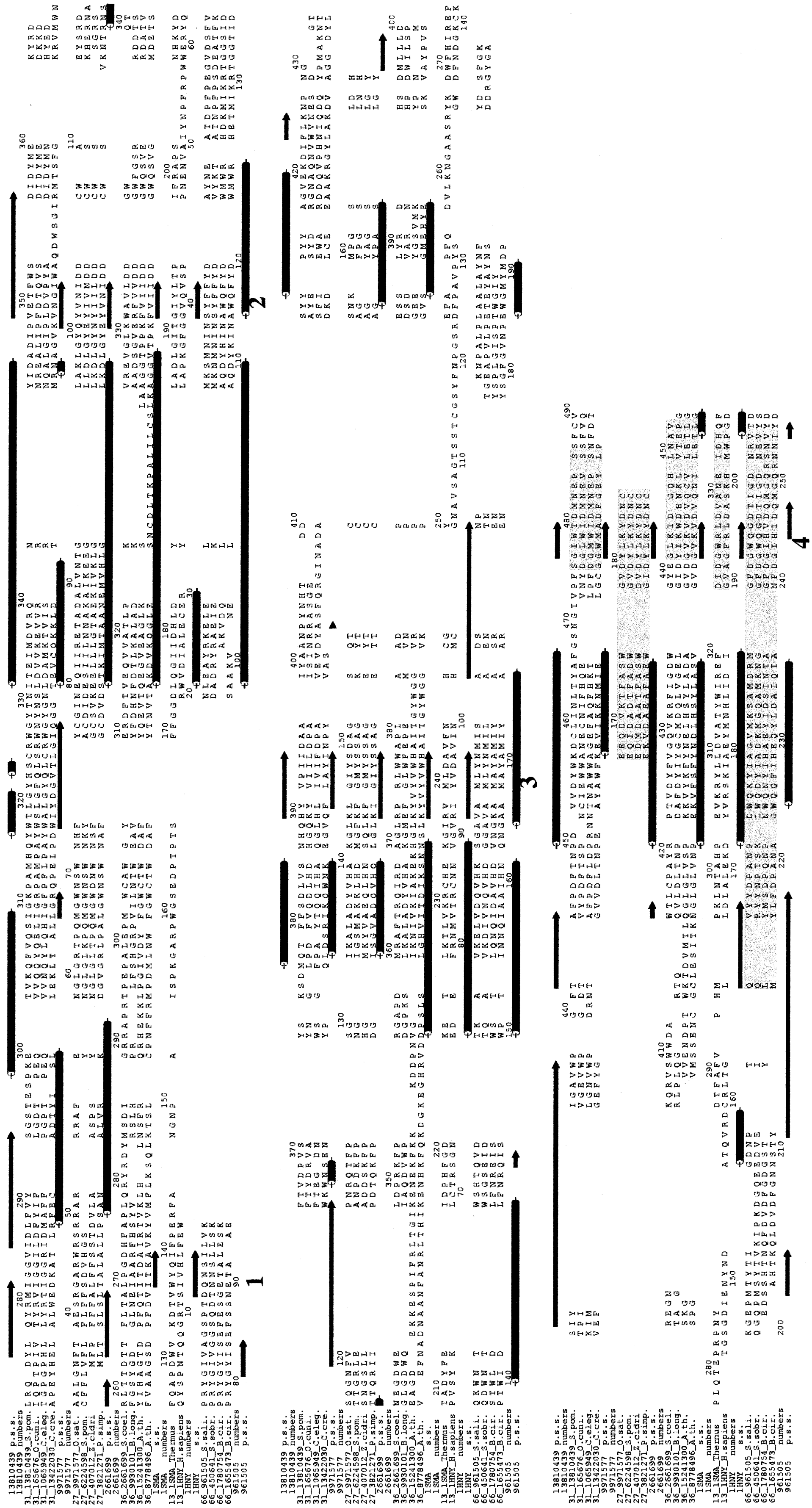
Fig. 1. Schematic representations of the evolutionary relationships established using PSI-BLAST. An arrow from family A to family B means that PSI-BLAST analysis of family A produced family B members among significant hits. The numbers *x*/*y* associated with each arrow are the number of iterations required to demonstrate each relationship using *E*-value cut-offs of 0.001 (*x*) or 0.01 (*y*). A dash in place of *x* signifies that the relationship was not apparent at the stricter *E*-value cut-off and these weaker relationships are shown as dotted lines. Note that not all relationships were demonstrable bidirectionally.

The results of PSI-BLAST analysis, represented schematically in Fig. 1, are indicative of a probable single evolutionary origin for these five GH families (from which, by implication, GH families 70 and 77 also evolved). Even using a conservative *E*-value cut-off value of 0.001, a clear network of links representing probable common evolutionary origin is formed. With the *E*-value cut-off of 0.01, typically used for the elucidation of distant evolutionary relationships [27], additional links are made. A single iteration (at *E*-value cut-off 0.01) is necessary to demonstrate the kinship of families 27 and 36, the only relationship in the diagram previously clearly demonstrated [12]. In contrast 10 iterations are required before family 13 sequences appear in the significant matches to family 31. Although weak local similarity between families 13 and 31 has been previously noted [28], they had generally been believed to exhibit no significant overall sequence similarity [29].

Among the sequences of known structure on the PSI-BLAST output, the most significant matches were obtained for *Thermus* maltogenic amylase (1SMA; [30]) in the cases of families 31 and 36. Family 13 proteins were not present in the results of PSI-BLAST analysis of family 27 or family 66 (Fig. 1) at *E*-value cut-offs of 0.01, but the other clear relationships elucidated (Fig. 1) enabled the structural correspondences to be made. An alignment of family 66 with human pancreatic amylase (1HNY; [31]) found using a slightly more relaxed *E*-value cut-off was also employed. Using these data, each family, represented by four maximally diverse members, was aligned with the α-amylases of known structure, previously structurally aligned, and the result carefully hand-modified. Excellent agreement between actual amylase secondary structure and the predicted secondary structures for families 27, 31, 36 and 66 was observed for the first half of the TIM barrel

→

Fig. 2. Portion of the alignment of GH families 31, 27, 36 and 66 with family 13. The region shown runs from the start of the TIM barrel to the catalytic nucleophile, thereby containing approximately the first half of the TIM barrel. The two family 13 sequences are shown in a structural alignment. Alignments of the other families to family 13 are from PSI-BLAST output. Four maximally diverse representatives from each of families 31, 27, 36 and 66 are shown and labelled with Genpept ID and abbreviated species name. Single representatives of each of these families (Table 1) are numbered and their secondary structure predictions shown. Motifs containing the nucleophilic Asp residue are shown as shaded regions. β-Strands of the family 13 TIM barrels are numbered beneath the alignment. The figure was produced with ALSCRIPT [39].


13810439 p.s.s.
13810439 numbers
31_13810439_S.pom.
31_1656476_O.cun.
31_1065949_C.eleg.
31_13422030_C.cre.
9971577 p.s.s.
9971577 numbers
27_9971577_O.sat.
27_6224598_S.pom.
27_4070121_Z.cldri
27_3821271_P.simp.
2661699 p.s.s.
2661699 numbers
36_2661699_S.coel.
36_9930101_B.long.
36_1624130_A.th.
36_8778496_A.th.
1SMA s.s.
1SMA numbers
13_1SMA_Thermus
13_1HNY_H.sapiens
1HNY numbers
1HNY s.s.
66_961505_S.sali.
66_4506411_S.sobr.
66_1780754_B.cir.
66_1655473_B.cir.
961505 numbers
961505 p.s.s.
1
2
3
4

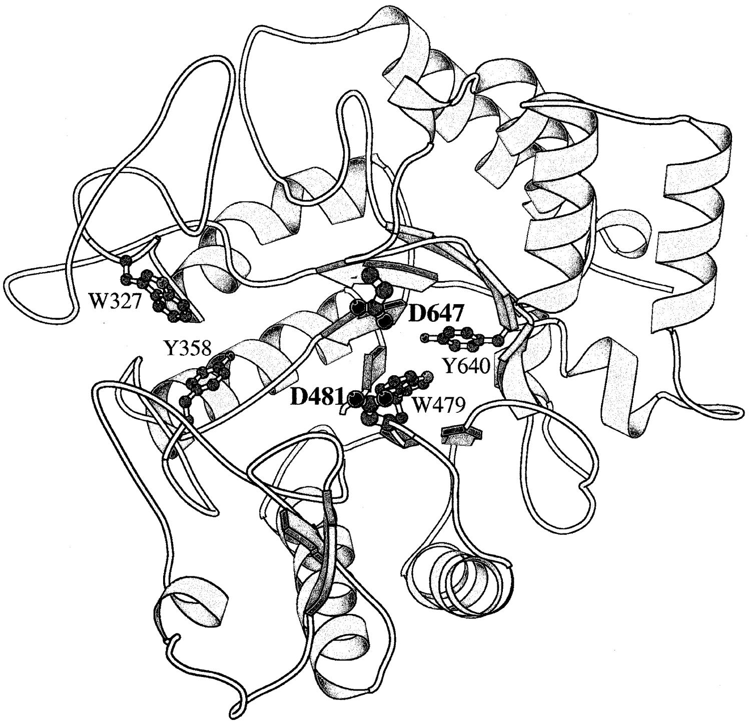


Fig. 3. MOLSCRIPT [40] figure of the model of the TIM barrel of family 31 representative *S. pombe* α-galactosidase. Putative catalytic (larger spheres, grey bonds) and substrate binding residues are shown using ball and stick representation and labelled.

(Fig. 2), worsening significantly in the second half. The contrast between the well-aligned first half of the TIM barrel and the more uncertain alignment in the second half (not shown) is interesting in the light of suggestions of half barrel-based modular evolution of TIM barrels [32]. Strikingly, a single position was entirely conserved in the alignment, Asp197 (1HNY numbering), previously determined experimentally to be the catalytic nucleophile in family 13 [28]. As shown in Fig. 2, the region surrounding this Asp is among the best conserved portions of the sequence, appearing in the top five MEME motifs in each family.

When the sequences shown in Fig. 2 were submitted as a group for motif discovery, a single significant motif common to all was obtained. It had the consensus sequence GFDGFKID, with the final Asp being the nucleophilic residue. The central Gly of the motif (193 in 1HNY numbering) is conserved with the exceptions of the four family 27 proteins and the Bifidobacterium representative of family 36 where it is replaced by Tyr. Examination of the 1HNY structure shows that this residue packs in the molecular core against position 94. Where a Gly is present at 193, position 94 contains a medium or large amino acid (Val, Ile, Leu, Arg, Met or Tyr; Fig. 2). However, in the five cases where a Tyr is present at 193, position 94 is a Gly (Fig. 2). For the preceding Asp of the motif, as similar situation is found. In 1SMA, Asp192 (maintaining 1HNY numbering) forms a buried electrostatic interaction with His87. Position 192 is occupied by an acidic residue with three exceptions, distributed among families 31, 36 and 13. A basic residue is present at position 87 in each protein, with the same three exceptions (Fig. 2). These clear

Table 1
Catalytic activities of GH families 27, 31, 36 and 66 and proposed catalytic residues

GH family number [3]	Catalytic activities	Representative used for numbering	Nucleophile	Candidate proton donors
27	α-galactosidase (EC 3.2.1.22) α-*N*-acetylgalactosaminidase (EC 3.2.1.49) isomalto-dextranase (EC 3.2.1.94)	*O. sativa* α-galactosidase (Genpept ID 9971577)	Asp185	Asp106, Asp107
31	α-glucosidase (EC 3.2.1.20) glucoamylase (EC 3.2.1.3) sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10) α-xylosidase (EC 3.2.1.-) α-glucan lyase (EC 4.2.2.13)	*S. pombe* α-galactosidase (Genpept ID 13810439)	Asp481 [29]	Asp 647 [29]
36	α-galactosidase (EC 3.2.1.22) stachyose synthase (EC 2.4.1.67) raffinose synthase (EC 2.4.1.82)	*Streptomyces coelicolor* hypothetical protein (Genpept ID 2661699)	Asp446	Asp336, Asp337
66	cycloisomaltooligosaccharide glucanotransferase (EC 2.4.1.-) dextranase (EC 3.2.1.11)	*S. salivarius* dextranase (Genpept ID 961505)	Asp247	Asp107, Asp378, Glu451

cases of compensatory mutations support the structural correspondence between these five families and their discovery is indicative of an accurate alignment.

With the nucleophilic residue clearly established for each of these families (Fig. 2), a search was made for possible catalytic proton donors (with the results summarised in Table 1). Clearly these will be among conserved acidic residues in multiple sequence alignments but, with the proven structural correspondence with α-amylases, these residues should be located in loops following the β-strands of the TIM barrel, on the side of the molecule in which catalytic sites are invariably located [33,34]. Predicted secondary structure was therefore analysed, although it is acknowledged to contain possible errors.

In family 31, the catalytic nucleophile has been identified as Asp481 (*Schizosaccharomyces pombe* α-glucosidase numbering) by chemical modification of human lysosomal α-glucosidase [35] and mutation of the corresponding residue, along with two others (Glu484 and Asp647), in *S. pombe* α-glucosidase [29] produced inactive enzymes. Each of these positions follows a predicted β-strand, supporting the rationale adopted above for seeking the proton donors of the families considered here. With hindsight, all of the other five mutated positions [29], whose mutation leads to less radical consequences for activity, would not have been favoured by our search criteria, lying either within predicted secondary structure elements or outside the predicted TIM barrel. Although position 484 is reported as a conserved Glu [29], alignment of CAZY members of family 31 reveals 15 sequences containing Asp, Thr, Val or Gln instead, including well-characterised algal enzymes [36]. This consideration strongly favours Asp647 as the proton donor in family 31. Initial PSI-BLAST and fold recognition alignments of families 13 and 31 do not align Asp647 with a conserved family 13 acidic residue. However, the predicted β-strand which Asp647 follows requires little adjustment to align with β-strand 7 of family 13. Thus family 31 enzymes may belong to the so-called 4/7 superfamily of TIM barrel glycosidases [37]. A model of the TIM barrel of the family 31 representative, *S. pombe* α-galactosidase, is shown in Fig. 3. Insertions and deletions relative to the template, *Thermus* maltogenic amylase (1SMA; [30]) were readily accommodated, but no attempt was made to model a 65 residue insertion comprising *S. pombe* α-galactosidase residues 491–555. Modelling highlights four highly conserved aromatic residues lining the substrate binding cleft. Interactions between aromatic residues and the hydrophobic faces of carbohydrate rings are commonly observed in carbohydrate binding proteins [38] so that these are possible substrate binding residues.

Application of our search criteria to the related families 27 and 36, combined with the assumption that catalytic machinery will be conserved between the two families, indicates two adjacent possible proton donors, Asps106 and Asp107 (numbering according to *Oryza sativa* α-galactosidase; Fig. 2). These immediately follow predicted β-strand 2 and are followed by conserved Gly–Trp and Cys–Trp in families 36 and 27, respectively. Again, the positioning of a conserved Trp near the catalytic site is strongly suggestive of a role in substrate binding [38]. Finally, in family 66, three suitably positioned conserved acidic residues could function as proton donors, Asp120 (numbering for *S. salivarius* dextranase) which located immediately after β-strand 2 and aligns with Asp107 in family 27. Alternatively, Asp378 and Glu451, conceivably positioned in the post β7 and post β8 loops, respectively, may have a catalytic role.

In summary, we have reason to believe that a common evolutionary origin is shared by two existing GH clans and two additional families. As a result, the nucleophilic Asp, uniquely conserved among all members of these families, is unambiguously identified. The assignment of a TIM barrel for families 27, 31, 36 and 66, along with their alignment to the well-understood family 13, enables a short list of possible proton donors to be produced, an advance on a simple search for conserved acidic residues in multiple sequence alignments. This work should encourage the search for common evolutionary origins among other GH families.

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] Bairoch, A. The ENZYME database in 2000, (2000) Nucleic Acids Res. 28, 304–305.

[5] Davies, G. and Henrissat, B. (1995) Structure 3, 853–859.

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

[7] Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.P. and Davies, G. (1995) Proc. Natl. Acad. Sci. USA 92, 7090–7094.

[8] Naumoff, D.G. (1999) FEBS Lett. 448, 177–179.

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

[10] Pons, T., Hernandez, L., Batista, F.R. and Chinea, G. (2000) Protein Sci. 9, 2285–2291.

[11] MacGregor, E.A., Janecek, S. and Svensson, B. (2001) Biochim. Biophys. Acta 1546, 1–20.

[12] Wang, A.M., Bishop, D.F. and Desnick, R.J. (1990) J. Biol. Chem. 265, 21859–21866.

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

[14] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.

[15] Bailey, T.L. and Elkan, C. (1994) in: Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28–36, AAAI Press, Menlo Park, CA.

[16] 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.

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

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

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

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

[21] Fischer, D. (2000) in: Pacific Symposium Biocomputing, Hawaii (Altman, R.B., Dunker, A.K., Hunter, L., Lauderdale, K. and Klein, T.E., Eds.), pp. 119–130, World Scientific, Singapore.

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

[23] 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.

[24] Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y. and Bryant, S.H. (2002) Nucleic Acids Res. 30, 281–283.

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

[26] Murzin, A.G. (1998) Curr. Opin. Struct. Biol. 8, 380–387.

[27] Aravind, L. and Koonin, E.V. (1999) J. Mol. Biol. 287, 1023–1040.

[28] McCarter, J.D. and Withers, S.G. (1996) J. Biol. Chem. 271, 6889–6894.

[29] Okuyama, M., Okuno, A., Shimizu, N., Mori, H., Kimura, A. and Chiba, S. (2001) Eur. J. Biochem. 268, 2270–2280.
[30] 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.
[31] Brayer, G.D., Luo, Y. and Withers, S.G. (1995) Protein Sci. 4, 1730–1742.
[32] Gerlt, J.A. and Babbitt, P.C. (2001) Nat. Struct. Biol. 8, 5–7.
[33] Wierenga, R.K. (2001) FEBS Lett. 492, 193–198.
[34] Nagano, N., Porter, C.T. and Thornton, J.M. (2001) Protein Eng. 14, 845–855.
[35] Hermans, M.M., Kroos, M.A., van Beeumen, J., Oostra, B.A. and Reuser, A.J. (1991) J. Biol. Chem. 266, 13507–13512.
[36] Bojsen, K., Yu, S., Kragh, K.M. and Marcussen, J. (1999) Biochim. Biophys. Acta 1430, 396–402.
[37] Jenkins, J., Lo Leggio, L., Harris, G. and Pickersgill, R. (1995) FEBS Lett. 362, 281–285.
[38] Quicho, F.A. and Vyas, N.K. (1999) in: Bioinorganic Chemistry: Carbohydrates, pp. 441–457, Oxford University Press, New York.
[39] Barton, G.J. (1993) Protein Eng. 6, 37–40.
[40] Kraulis, J. (1991) J. Appl. Cryst. 24, 946–950.