FEBS Letters 584 (2010) 4435–4441

journal homepage: www.FEBSLetters.org

Consolidation of glycosyl hydrolase family 30: A dual domain 4/7 hydrolase family consisting of two structurally distinct groups

Franz J. St John *, Javier M. González, Edwin Pozharski

University of Maryland, School of Pharmacy, Department of Pharmaceutical Sciences, Baltimore, MD, United States

ARTICLE INFO

Article history:
Received 12 July 2010
Revised 23 September 2010
Accepted 30 September 2010
Available online 8 October 2010

Edited by Takashi Gojobori

Keywords:
Glycosyl hydrolase family 30 (GH30)
Glycosyl hydrolase family 5 (GH5)
Glucuronoxylan xylanohydrolase
Endo-β-1,6-galactanase
Glucosylceramidase
Endo-β-1,6-glucanase

ABSTRACT

In this work glycosyl hydrolase (GH) family 30 (GH30) is analyzed and shown to consist of its currently classified member sequences as well as several homologous sequence groups currently assigned within family GH5. A large scale amino acid sequence alignment and a phylogenetic tree were generated and GH30 groups and subgroups were designated. A partial rearrangement in the GH30 defining side-associated β-domain contributes to the differentiation of two major groups that contain up to eight subgroups. For this CAZy family of Clan A enzymes the dual domain fold is conserved, suggesting that it may be a requirement for evolved function. This work redefines GH family 30 and serves as a guide for future efforts regarding enzymes classified within this family.

Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.

1. Introduction

Sequence based classification of proteins into families has been applied successfully for many years and has improved with the increasing number of new sequences made available through genome sequencing projects. The Carbohydrate-Active EnZymes database (CAZy) (http://www.cazy.org) [1–3] is the primary resource for sequence based categorization of enzymes that catalyze reactions involving the cleavage, creation or modification of glycosidic bonds. Recent biochemical characterization [4,5] and structural studies of the GH5 xylanase XynC from *Bacillus subtilis* [6] has directed attention to a previously noted inconsistency in the GH classification of this enzyme [7–11]. Early classification discussed the difficulty of making an unambiguous assignment based on hydrophobic cluster analysis comparisons with the XynC homolog, XynA from *Erwinia chrysanthemi* with GH family 5 and 30 $(\beta/\alpha)_8$ barrel domains [12,13]. Within the current, more populated database space these enzymes show similarity to GH30 enzymes.

Enzyme activities currently assigned within GH family 30 include glucosylceramidase (EC 3.2.1.45), β-glucosidase (EC 3.2.1.21), β-xylosidase (EC 3.2.1.37), and endo-β-1,6-glucanase (EC 3.2.1.75). Of these enzymes a crystal structure is available for human glucosylceramidase (Gba1) [14] and an uncharacterized protein, SrfJ, from *Salmonella enterica* [15]. The putative xylosidase activity was the subject of a single publication concerning *Bifidobacterium breve*. Interestingly, it was characterized to hydrolyze a single xylose moiety from the saponin, ginsenoside Ra1 derived from ginseng root [16]. The chemical characteristics of saponins may be considered analogous to glucosylceramides, both having glycosyl moieties O-linked to large hydrophobic molecules. A clear distinction is observed between these β-glucosidase/β-xylosidase like enzymes that have exo-function and the phylogenetically related endo-β-1,6-glucanase [17,18]. These later enzymes function in an endo-manner and have been characterized to degrade β-1,6-glucan, a polysaccharide component of fungal cell walls [19,20].

This work presents a phylogenetic analysis of sequences that show similarity to the originally classified GH30 family of enzymes. Amino acid and secondary structure alignments and available structure data are used to identify similarities that support combining these enzyme groups into a single family. From the analysis, it is clear that this family may have several additional

Abbreviations: GH, glycosyl hydrolase; GH5, glycosyl hydrolase family 5; GH30, glycosyl hydrolase family 30; ACC, UniProt accession number; PDB, PDB code; NJ, Neighbor-Joining; ML, maximum-likelihood

* Corresponding author. Present address: Institute for Microbial and Biochemical Technology, Forest Products Laboratory, US Forest Service, One Gifford Pinchot Dr., Bldg 34, Rm 123, Madison, WI 53726, United States. Fax: +1 608 231 9592.

E-mail address: fjstjohn@fs.fed.us (F.J. St John).

0014-5793/$36.00 Published by Elsevier B.V. on behalf of the Federation of European Biochemical Societies.
doi:10.1016/j.febslet.2010.09.051

enzymatic activities, including some that may show promise in the biodegradation of lignocellulosic polysaccharides. This report recommends the transfer of five GH5 protein subgroups representing approximately 140 amino acid sequences to GH30 and establishment of GH30 Group 2.

2. Methods

All sequences used in this work were obtained through the UniProt database [21]. An inclusive set of GH30 amino acid sequences with *E*-values as high as 10^{-20} were collected using *BLASTp* [22] with the amino acid sequences of the newly assigned GH30 glucuronoxylan xylanohydrolase XynC of *B. subtilis* (ACC Q45070) [4,6], a protein of unknown function also being reassigned to GH30 from *Bacteroides fragilis* (ACC Q5LF82), Gba1 (glucosylceramidase) from *H. sapiens* (ACC P04062) [14,23], a *B. breve* β-xylosidase (ACC B1P195) [16] and a *Trichoderma harzianum* endo-β-1,6-glucanase (ACC Q8J0I9) [17]. In each case, sequences having low amino acid sequence identity levels (∼27%) were used in secondary rounds of *BLASTp*. Similarity between any two sequences was verified with *PRSS* [24]. Protein modular composition was determined using *rpsBLAST*/CDD server [25]. Sequence trimming was performed using the alignment editor interface in *MEGA 4.0* [26] following detailed alignment analysis and comparison to the known protein structures. In this processing, secretion signal sequences, additionally appended function domains and in specific cases such as with glucosylceramidases, an additional N-terminus region was removed to result in a sequence dataset containing the core dual domain fold of GH30 enzymes. Final data was prepared from sequences trimmed to consist only of the GH30 $(\beta/\alpha)_8$ catalytic center, allowing for direct comparison to the GH5 $(\beta/\alpha)_8$ catalytic core domain. A group of GH5 enzymes were processed as the GH30 enzymes and was used to root the GH30 phylogenetic trees and verify family segregation. Sequence alignments were performed using *MAFFT* [27]. *MEGA 4.0* [26] was used to create phylogram trees using the Neighbor-Joining (NJ) method with bootstrap (1000 replicate) to infer evolutionary relationships [28] and also for tree image preparation. To verify initial findings, the maximum-likelihood (ML) based phylogeny program *PhyML* [29] was used with the LG substitution model and branch support analysis using the program *aLRT* with SH-like interpretation [30–32]. The original dataset sequence alignment was too large for ML analysis. To reduce the dataset complexity for ML analysis the dataset was reduced from 356 GH30 enzymes with 21 GH5 enzymes to 309 GH30 sequences and 9 GH5 sequences. The resulting dataset, used for ML analysis, is still largely representative of the breadth of GH30 enzymes. Other than the ML phylogenetic analysis used to support the initial NJ analysis all other studies (sequence logo analysis) used the original dataset or a smaller 50 sequence set selected to represent a diverse sampling of those sequences and those enzymes that have corresponding structural or biochemical data (Fig. S7, selected sequences denoted in Fig. 1 highlighted with an asterisk). The *DALI* server [33]

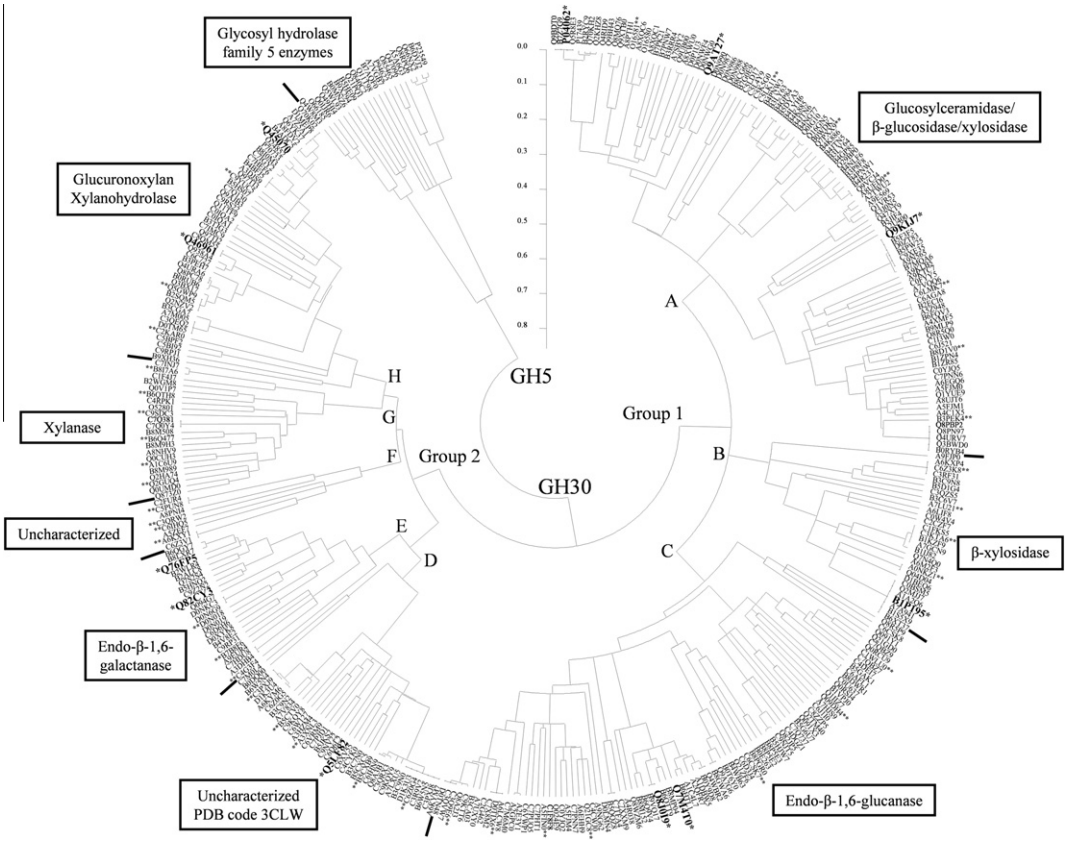

Fig. 1. A Neighbor-Joining bootstrap (1000 replicate) consensus phylogram (see Fig. S10 for visual details) displaying two major groups and up to eight subgroups of GH family 30 enzymes, rooted by the inclusion of a set of GH family 5 endoglucanase and mannanase enzymes. Data reflects the sequence of the $(\beta/\alpha)_8$ catalytic motif only. Group 1 includes GH30 subgroups A–C and includes the biochemically characterized enzymes glucosylceramidase (subgroup A), β-glucosidase/β-xylosidase (subgroup A), β-xylosidase (subgroup B) and an endo-β-1,6-glucanase (subgroup C). Group 2 include subgroups D–H which are represented by biochemically characterized endo-β-1,6-galactanase (subgroup E), an acidic β-1,4-xylanase (subgroup G) and β-1,4-glucuronoxylan xylanohydrolase (subgroup H) activities. Subgroups D and F have no biochemical data to define function although subgroup D has a crystal structure model available for comparison study. Approximately 140 sequences classify as Group 2 enzymes, all are thought to previously be classified as GH5 enzymes. Sequences that represent biochemically characterized enzymes or those that have structure models available are in bold face type. Sequences used for the *EXPRESSO* structure guided sequence alignment (Fig. S7) are highlighted with an asterisk(s).

was used to generate a 3D-structure alignment using the structure model of XynC from *B. subtilis* (PDB 3GTN) and the resulting secondary structure alignment of the top three structurally homologous proteins (PDB 1NOF, 3CLW and 2J25) were used as a template to extend secondary structure elements onto a larger structure guided alignment output from the program *EXPRESSO* [34]. A diverse set of GH5 enzyme sequences was collected and used to generate an *EXPRESSO* alignment for qualitative assessment of GH5 conservation patterns and comparison to the GH30 *EXPRESSO* alignment. *Weblogo* [35], *DIVAA* [36], *Coot* [37] and *PyMOL* [38] were used for analysis and figure preparation.

3. Results

3.1. One GH family with two groups

Phylogenetic analysis of known GH30 enzymes and those amino acid sequences similar to them revealed two major groups that can be divided into eight subgroups (Fig. 1, see Fig. S10 for visual details), distinguished by levels of sequence identity of around 30%. Subgroup identification was guided by the phylogenetic tree distribution. Glucosylceramidase, β-glucosidase, β-xylosidase and endo-β-1,6-glucanase activities were the original enzymatic activities characterizing the GH30 family and comprise subgroups A–C, respectively, of Group 1. The newly identified Group 2 consists of five subgroups, D–H that originate from the same primary phylogenetic branch. Subgroup D has no biochemically characterized members, although a structure model is available of a conserved exported protein from *B. fragilis* (PDB 3CLW). Subgroup E is identified by two biochemically characterized enzymes with endo-β-1,6-galactanase activity [39,40]. A recent report [41] assigned an acidic β-1,4-xylanase function to a fungal enzyme that shares approximately 50% identity with a subgroup G enzyme, also of fungal origin (ACC B8M989). No information is available for the sparsely populated subgroup F. Two subgroup H members are characterized having glucuronoxylan xylanohydrolase activity and structure models are available for these same enzymes [4–6,11,42].

Maximum likelihood phylogenetic analysis confirmed the initial findings with acceptable branch support values (Figs. S1 and S2). *PhyML* analysis produced optimized midpoint rooted trees [29,30]. The midpoint root in Fig. S1 clearly shows the evolutionary separation that exists between GH5 enzymes and the isolated $(\beta/\alpha)_8$ barrel of GH30 sequences and the differential, but close relationship between the two major GH30 groups (Fig. S2).

We performed additional phylogenetic analyses of the two primary GH30 Groups. Group 1 enzymes consisting of subgroups A, B and C branched as anticipated yielding high confidence branch support values (Fig. S4). Group 2 enzymes consisting of subgroups D, E, F, G and H grouped as expected (Fig. S3), except for three sequences that branched as subgroup G outliers in previous larger-dataset NJ or ML analyses. In this Group 2 tree they clustered with subgroup F with low branch support values suggesting that they may represent another subgroup which is underrepresented in the current dataset. Subgroups A and C were further analyzed and shown to consist of at least two well supported subfamilies (Figs. S5 and S6). While each of these subgroups contained sequences with identity levels over 40% neither had levels below 30%.

3.2. Side β-structure rearrangement

The overall structure of GH30 enzymes is unique in comparison to GH5 enzymes, being formed from the fusion of a common $(\beta/\alpha)_8$ barrel CAZy designated Clan A catalytic module (4/7 hydrolase, GH5 like) [2,43] with a side β-structure consisting of a 9-stranded aligned β-sandwich, immunoglobulin like fold. This family-wide structural addition is hinged with conserved sequence from the N- and C-terminus regions [6,11]. Group differentiation within the GH30 family at least partially results from a secondary structure rearrangement found within the side β-structure. This is represented by comparison of the structure models of Gba1 from *H. sapiens* (PDB 1OGS) and XynC from *B. subtilis* (PDB 3GTN). Fig. 2 depicts the secondary structure topology between these two homologous structures showing the alternative β-strand arrangement. For Gba1 (Group 1, Fig. 2a), three of the nine β-strands (βs1–βs3) of the side β-structure are coded for in the N-terminus and the

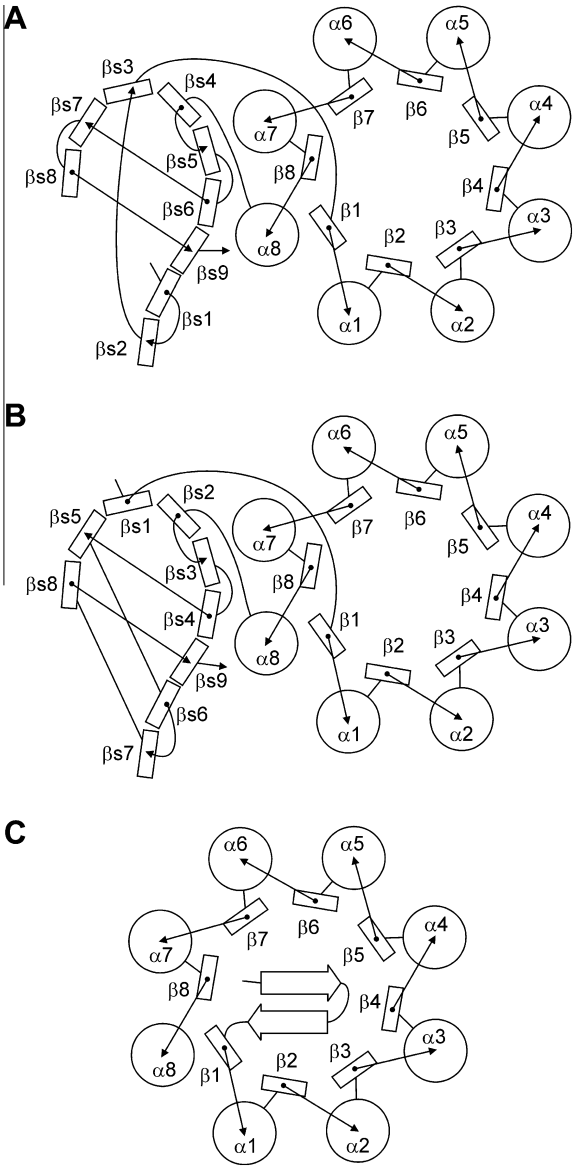

Fig. 2. Schematics of the side β-domain secondary structure rearrangement showing a difference that is thought to contribute to the distinction of the two observed GH30 groups. (A) Diagram showing the order of secondary structure elements in Gba1 (PDB 1OGS) which represents enzymes classified into GH30 Group 1. (B) Diagram showing the order of secondary structure elements in XynC (PDB 3GTN) which represents enzymes classified into GH30 Group 2. (C) Diagram showing the order of secondary structure elements as derived from the inspection of all (26 total) GH5 enzymes which have been structurally characterized. Only one of these structures did not have the depicted N-terminus cap structure (*Thermoascus aurantiacus* Eng1, PDB code 1GZJ). The two-stranded β-sheet is the most prevalent structure but there are examples of short α-helix structures and slightly larger more complex β-structures, all of which are in the same position appearing as a cap over the bottom end of the $(\beta/\alpha)_8$ barrel. βs# refers to a β-strand number of the side β-motif and β# refers to the β strand number of the $(\beta/\alpha)_8$ barrel.

remaining six in the C-terminus region (βs4–βs9). After the third β-strand the amino acid chain enters into the $(\beta/\alpha)_8$ barrel. Following this domain, the fourth β-strand of the side β-structure is formed (βs4). The remaining strands of this structure are then completed (βs5–βs9). In the case of XynC (Group 2, Fig. 2b), the first β-sheet of the mature secreted protein forms β-sheet one of the side β-structure (βs1). The amino acid chain then enters the $(\beta/\alpha)_8$ catalytic domain. Upon exiting this domain the second β-strand of the side β-structure is formed (βs2). The remaining β-strand order of XynC is partially different, however, the two domains overlay with a percentile-based spread [44] of just 1 Å indicating their close three-dimensional similarity. This secondary structure rearrangement can also be visualized in the *EXPRESSO* alignment presented in Fig. S7.

3.3. Group distinction within the $(\beta/\alpha)_8$ barrel

Although group differentiation was originally identified from the β-strand rearrangement in the side β-structure (Fig. 2), almost identical phylogenetic tree distribution is produced from alignments that include only the sequence of the $(\beta/\alpha)_8$ barrel rather than the complete dual domain structure. Several α-helical and loop regions within the $(\beta/\alpha)_8$ barrel contribute to this differentiation. From the *EXPRESSO* structure guided alignment (Fig. S7) specific differential regions include the α2 helix, α5 helix, α6 helix and the β8-α8 loop all of which appear more conserved among Group 1 sequences and less conserved in Group 2 sequences. The opposite is observed for the α7 helix which appears more conserved among Group 2 sequences. Together with numerous single amino acid differences these larger regions help segregate Group 1 from Group 2 sequences.

3.4. GH30 unique conserved regions

To emphasize the significance of the dual domain fold as a single functioning catalytic entity we analyzed regions of sequence conservation between GH30 and GH5 enzymes. From the *EXPRESSO* alignment it is found that GH30 enzymes have a distinctive sequence conservation pattern. These include the two linker regions which connect the side β-structure to the $(\beta/\alpha)_8$ catalytic center (Fig. 3a and d) and a continuous conserved region beginning with α-helix 3 and ending with the catalytic proton donor (Glu140 of XynC) positioned within β-strand 4 (Fig. 3b). Unexpectedly, the catalytic nucleophile region (Glu229 of XynC) positioned in β-strand 7 has considerable sequence diversity (Fig. 3c). In contrast, GH5 enzymes do not have the GH30 characteristic conserved hinge regions and the sequences in the β1 and α8 terminal barrel motifs do not share obvious sequence similarities with the same regions of GH30 sequences. We observe that the α8 region of GH5 $(\beta/\alpha)_8$ barrels show significant diversity (Fig. S8) within the GH5 family. The β7 region harboring the expected catalytic nucleophile appears much more conserved then in GH30 enzymes. GH5 enzymes are

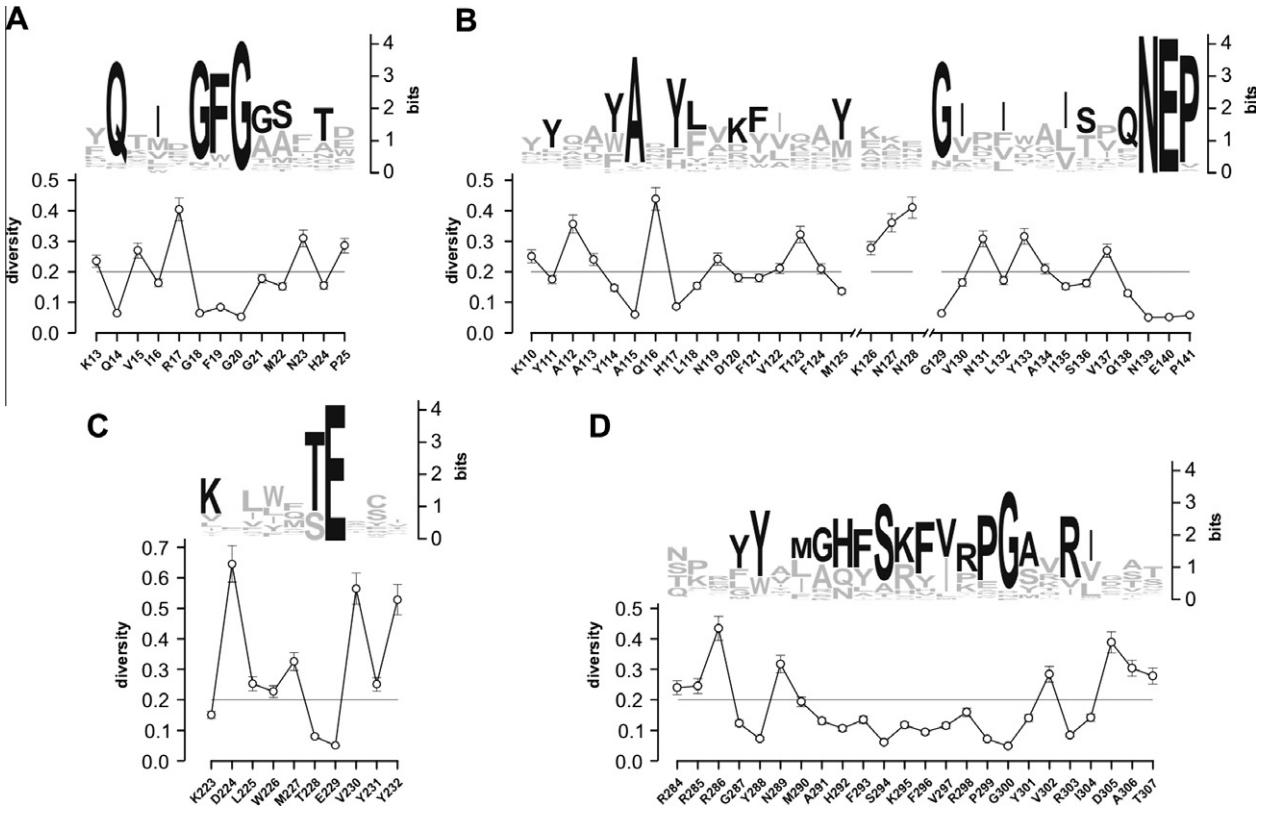

Fig. 3. Amino acid alignment derived, family wide similarities that support the grouping of several GH family 5 subfamilies into GH family 30. The programs *WebLogo* (top of each image) and *DIVAA* (bottom of each image) characterize the nature of amino acid conservation in several GH family 30 unique regions resulting from the full 356 sequence alignment (Fig. 1). *WebLogo* measures positional conservation and amino acid frequency as the column height (4.32 as the maximum) and the most represented amino acid (by the size of the top amino acid), respectively. *DIVAA* analyzes positional diversity ranging from a value of 0.05 to 1 for a single amino acid in a position to representation by all 20 amino acids. A line has been added at 0.2 of the *DIVAA* analysis to indicate amino acid positions that are represented by no more than four amino acids which correspond to an amino acid in bold print in the *WebLogo* image. (A) The N-terminus hinge region of GH30 enzymes which involves the βs1 strand of the side β-structure and β1 of the $(\beta/\alpha)_8$ barrel. (B) The α3–β4 region of the $(\beta/\alpha)_8$ barrel shows extended conservation due to a minimized connecting loop that introduces only a minor gap in the alignment between these two conserved regions. (C) The β7 strand of the $(\beta/\alpha)_8$ barrel which, by homology, includes the catalytic nucleophile appears to be uniquely diverse. (D) The C-terminus hinge region of GH30 enzymes which involves the α8 helix of the $(\beta/\alpha)_8$ barrel and βs2 of the side β-structure. Amino acid numbering is for XynC (PDB 3GTN).

also different having an N-terminus cap structure most often consisting of a small two stranded β-sheet (Fig. 2c). This is positioned on the bottom non-catalytic side of the $(\beta/\alpha)_8$ barrel and in no way resembles the initial N-terminus β-strands of the side β-structure of GH30 enzymes.

4. Discussion

In the CAZy database glycosyl hydrolase (GH) classification scheme, several biochemically different enzymes that are currently classified as members of GH family 5, are more similar to enzymes of GH family 30. With respect to the well-characterized, GH family 5 endoxylanase [4,5,11,13], several authors have previously noted this classification anomaly [7,9–11]. The findings presented in this work support a reclassification of these GH family 5 enzyme groups as GH30 enzymes.

4.1. Phylogenetic assessment

The initial dataset represented an intentionally inclusive, diverse sampling of sequences with similarity to the known GH30 enzymes as well as those with verified similarity from GH family 5. NJ phylogenetic analysis provided a bootstrap consensus tree with low branch support values (Fig. 1). It was considered that these results were primarily due to the breadth of the sequence sampling and the inclusion of many GH5 enzymes to root the tree. To confirm the findings of the initial NJ tree, ML analysis was performed. This analysis (Figs. S1–S6) fully supported the NJ results. From the midpoint root of the optimized *PhyML* tree it can be seen that GH5 enzymes cluster distinctly separate from GH30 enzymes (Fig. S1) and that GH30 sequences are represented by two major groupings (Figs. S1 and S2). Assessment of the structure guided *EXPRESSO* alignment (Fig. S7) suggests that distinguishing characteristics between these two major GH30 groups derive from loop region differences. An additional distinguishing characteristic not taken into consideration by the phylogenetic analysis of the $(\beta/\alpha)_8$ barrel involves the secondary structure rearrangement of the attached β-structure (Fig. 2).

The subgroup classification proposed in this work aligns phylogenetically similar sequences having levels of identity as low as 30% with biochemical and/or structurally characterized enzymes (Fig. 1). At most, only a single enzymatic activity could be assigned within a subgroup. Since identity levels are lower than the 40% needed for minimum confidence for enzymatic function annotation [45–47] and several of the subgroups have multiple well supported branches, the existence of subfamilies having different substrate specificity or enzymatic function is likely. This possibility is best represented by the large A and C subgroup ML analysis (Figs. S5 and S6) where each could be confidently separated into at least two different families with amino acid identities below 40% but higher than 30%. This level of distinction is not warranted without corresponding knowledge of functional differences. In contrast, subgroups G and H both have assigned xylanase activity. However, subgroup H xylanases are fundamentally unique in comparison to subgroup G having an enzymatic activity which relies upon recognition of an α-1,2-linked 4-O-methyl glucuronate moiety [4,5,41]. Noting the very similar enzymatic function between these two subgroups, albeit with alternative specificities, it is interesting to note that these enzymes all have comparable levels of identity below 30% as expected for differentiation between the GH30 subgroups.

4.2. Side β-structure rearrangement and group differentiation

For the purpose of this work the $(\beta/\alpha)_8$ barrel catalytic domain of GH30 enzymes was compared to that of GH5 enzymes. It was originally found that the full two-domain sequence of GH30 enzymes yielded a similar bipartite tree as reported in this work. The proposed group-wide β-structure rearrangement is supported by two crystal structures from Group 1 and three from Group 2 (Fig. 2a and b) and also by the structure guided *EXPRESSO* alignment (Fig. S7) that suggests this structural difference is consistent throughout the two GH30 groups. However, it should be noted that this region aligned poorly in *MAFFT* alignments of the full GH30 dataset. In the GH30 $(\beta/\alpha)_8$ barrel conserved catalytic domain motifs such as the regions that comprise the catalytic amino acids, function as an internal control to assess the alignment quality. In the side β-structure no such reference motif exists and the alignment in this region appears more variable. This variability was not apparent in the results of the smaller 50-sequence structure guided *EXPRESSO* alignment (Fig. S7) even with the selected sequences representing the breadth of the original NJ phylogram (Fig. 1, sequences highlighted with asterisks).

Without accounting for this rearrangement in side β-structure, phylogenetic analysis of sequence alignments resolved the GH30 subgroups into the two identically partitioned groups. Inspection of the *EXPRESSO* alignment revealed $(\beta/\alpha)_8$ barrel loop and α-helix region differences that may account for this. Even though, the β-structure rearrangement results in the same general protein fold, the possibility exists that the Group 1 β-structure arrangement supports the more extensive loop systems and exo-activity observed in Group 1 enzymes better than the Group 2 enzymes β-structure arrangement. Support for this view can be derived from the knowledge that many of the Group 1 enzymes (particularly of subgroup A) have additional N-terminus regions that were trimmed for this analysis. This additional sequence in Gba1 is positioned to the upper side of the side β-structure interfacing with outer edge of the β1-α1, β7-α7 and β8-α8 loops which are instrumental in substrate pocket formation. This additional structure establishes two disulphide bridges within its sequence and has an extended hydrophobic interface on the $(\beta/\alpha)_8$ barrel surface [14].

A *DALI* structure alignment identified three other CAZy families that share the two domain fold described here for GH30 enzymes. Assessment of GH families 39 [48], 44 [49] and 51 [50] reveals that their β-structure arrangement is identical to Group 2 enzymes. None are found to have significant levels of sequence identity to GH30 sequences just as with GH5 enzymes. All three of these families are characterized to degrade some component polysaccharide of lignocellulosic biomass (Fig. S9).

4.3. Clear differences between GH30 and GH5 enzymes

Two conserved motifs that help to define the GH30 family are derived from the highly conserved hinge regions that connect the catalytic $(\beta/\alpha)_8$ barrel with the apparent GH30 obligate side β-structure. Even with the group defining β-structure rearrangement (Fig. 2) these regions have characteristic sequences detectable through consensus sequence analysis (Fig. 3). GH5 enzymes, not having the side β-structure do not have the conserved adjoining β1 sheet and α8 helix which lead into this structure in GH30 enzymes. From our analysis, GH5 enzymes seem not to have the extended conserved α3-β4 region found in the GH30 enzymes (Fig. S8). Rather, high sequence identity is found confined to the β4 catalytic proton donor motif generally common among CAZy Clan A enzymes. GH5 enzymes also differ from GH30 by having a cap like structure on the bottom non-catalytic side of the $(\beta/\alpha)_8$ barrel.

4.4. Two domains may do what one cannot

Although the side β-structure is a characteristic feature of GH30 enzymes, its function remains unknown. However, it is clear that it is not simply a loosely associated β-structure that is common in

glycosyl hydrolases involved in lignocellulose degradation [51]. The side β-structure is fused with conserved sequence elements from the N- and C-terminus with a hinged linker and forms a tight hydrophobic interface with the outer surface of α-helix 7 and 8 of the $(\beta/\alpha)_8$ barrel surface. As evidenced from the family conserved hinge region connecting the dual domain fold (Fig. 3a and d), it is in question whether the two structures could be separated and enzymatic function maintained. The region of β-strand 7 that positions the catalytic nucleophile is proximal to α-helix 7 and 8 which are involved in the hydrophobic association of the side β-structure. The interface between these two domains likely strengthens and/or stiffens this region of the $(\beta/\alpha)_8$ barrel. It would also increase the order of variability in this region of the protein potentially allowing for a greater variety of evolutionarily selectable functions than what might be available in a simple $(\beta/\alpha)_8$ domain. The low similarity observed in the β-strand 7 catalytic nucleophile region may be attributable to this structural phenomenon. Since the dual domain fold is conserved and the linker and interface between the two domains is conserved, the entire three-dimensional region of the interface is conserved and likely supports an increased diversity in the amino acids surrounding the catalytic nucleophile. Or rather, due to the conserved surrounding structure, changes in β-strand 7 may be more easily tolerated.

4.5. Summary

This work serves to consolidate enzymes with similarity to the originally classified GH30 family of enzymes. In doing so, it was confirmed [7,9–11] that five groups of enzymes currently assigned in GH family 5 are more similar to GH30 enzymes. Subgroup E contains two similar sequences with endo-β-1,6-galactanase activity [39,40], subgroup G has a recently characterized xylanase activity [41] and subgroup H contains the reassigned glucuronoxylan xylanohydrolase enzymes [4,5,42]. Subgroups D and F contain no biochemically characterized enzymes. A commonality among these enzymes is a side β-structure connected through two conserved linkers involving both the N- and C-terminus that associates through a large hydrophobic interface on the side of the catalytic $(\beta/\alpha)_8$ barrel suggesting that these components may serve as a single functional unit. A protein structure alignment revealed that along with GH30, GH families 39, 44 and 51 share this type of dual domain fold. Due to the prevalence of this fold we propose the reference name $(\beta/\alpha)_8$ + β fold. This new description of the GH30 family of enzymes will facilitate more accurate sequence annotation and future enzyme function prediction.

Acknowledgments

We thank James F. Preston at the University of Florida, Department of Microbiology and Cell Science, Jason C. Hurlbert at Winthrop University, Department of Biochemistry and Phil Kersten at the Forest Products Laboratory for thoughtful review and comments regarding the manuscript. We also thank Bernard Henrissat for insightful discussion during manuscript preparation and the University of Maryland X-ray Crystallography Shared Service for providing computing resources.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.febslet.2010.09.051.

References

[1] Henrissat, B. (1991) A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem. J. 280, 309–316.
[2] Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J.P. and Davies, G. (1995) Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. Proc. Natl. Acad. Sci. USA 92, 7090–7094.
[3] Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V. and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res. 37, D233–D238.
[4] St John, F.J., Rice, J.D. and Preston, J.F. (2006) Characterization of XynC from *Bacillus subtilis* subsp. *subtilis* strain 168 and analysis of its role in depolymerization of glucuronoxylan. J. Bacteriol. 188, 8617–8626.
[5] Vrsanska, M., Kolenova, K., Puchart, V. and Biely, P. (2007) Mode of action of glycoside hydrolase family 5 glucuronoxylan xylanohydrolase from *Erwinia chrysanthemi*. FEBS J. 274, 1666–1677.
[6] St John, F.J., Godwin, D.K., Preston, J.F., Pozharski, E. and Hurlbert, J.C. (2009) Crystallization and crystallographic analysis of *Bacillus subtilis* xylanase C. Acta Crystallogr. F 65, 499–503.
[7] Mitreva-Dautova, M., Roze, E., Overmars, H., de Graaff, L., Schots, A., Helder, J., Goverse, A., Bakker, J. and Smant, G. (2006) A symbiont-independent endo-1,4-beta-xylanase from the plant-parasitic nematode *Meloidogyne incognita*. Mol. Plant Microbe Interact. 19, 521–529.
[8] Preston, J.F., Hurlbert, J.C., Rice, J.D., Ragunathan, A. and St. John, F.J. (2003) Microbial strategies for the depolymerization of glucuronoxylan: leads to biotechnological applications of endoxylanases (Mansfield, S.D. and Saddler, J.N., Eds.), Applications of Enzymes to Lignocellulosics, vol. 855, pp. 191–210, American Chemical Society, Washington DC.
[9] Brumshtein, B., Wormald, M.R., Silman, I., Futerman, A.H. and Sussman, J.L. (2006) Structural comparison of differently glycosylated forms of acid-beta-glucosidase, the defective enzyme in Gaucher disease. Acta Crystallogr. D 62, 1458–1465.
[10] Haegeman, A., Vanholme, B. and Gheysen, G. (2009) Characterization of a putative endoxylanase in the migratory plant-parasitic nematode *Radopholus similis*. Mol. Plant Pathol. 10, 389–401.
[11] Larson, S.B., Day, J., Barba de la Rosa, A.P., Keen, N.T. and McPherson, A. (2003) First crystallographic structure of a xylanase from glycoside hydrolase family 5: implications for catalysis. Biochemistry 42, 8411–8422.
[12] Durand, P., Lehn, P., Callebaut, I., Fabrega, S., Henrissat, B. and Mornon, J.P. (1997) Active-site motifs of lysosomal acid hydrolases: invariant features of clan GH-A glycosyl hydrolases deduced from hydrophobic cluster analysis. Glycobiology 7, 277–284.
[13] Keen, N.T., Boyd, C. and Henrissat, B. (1996) Cloning and characterization of a xylanase gene from corn strains of *Erwinia chrysanthemi*. Mol. Plant Microbe Interact. 9, 651–657.
[14] Dvir, H., Harel, M., McCarthy, A.A., Toker, L., Silman, I., Futerman, A.H. and Sussman, J.L. (2003) X-ray structure of human acid-beta-glucosidase, the defective enzyme in Gaucher disease. EMBO Rep. 4, 704–709.
[15] Kim, Y.G., Kim, J.H. and Kim, K.J. (2009) Crystal structure of the *Salmonella enterica* serovar *typhimurium* virulence factor SrfJ, a glycoside hydrolase family enzyme. J. Bacteriol. 191, 6550–6554.
[16] Shin, H.Y., Lee, J.H., Lee, J.Y., Han, Y.O., Han, M.J. and Kim, D.H. (2003) Purification and charaterization of ginsenoside Ra-hydrolyzing beta-D-Xylosidase from *Bifidobacterium breve* K-110, a Human Intestinal Anaerobic Bacterium. Biol. Pharm. Bull. 26, 1170–1173.
[17] Montero, M., Sanz, L., Rey, M., Monte, E. and Llobell, A. (2005) BGN16.3, a novel acidic beta-1, 6-glucanase from mycoparasitic fungus *Trichoderma harzianum* CECT 2413. FEBS J. 272, 3441–3448.
[18] Oyama, S., Inoue, H., Yamagata, Y., Nakajima, T. and Abe, K. (2006) Functional analysis of an endo-1,6-beta-D-glucanase gene (neg-1) from *Neurospora crassa*. Biosci. Biotechnol. Biochem. 70, 1773–1775.
[19] Manners, D.J., Masson, A.J., Patterson, J.C., Bjorndal, H. and Lindberg, B. (1973) The structure of a beta-(1–6)-D-glucan from yeast cell walls. Biochem. J. 135, 31–36.
[20] Narui, T., Sawada, K., Culberson, C.F., Culberson, W.L. and Shibata, S. (1999) Pustulan-type polysaccharides as a constant character of the Umbilicariaceae (lichenized Ascomycotina). Bryologist 102, 80–85.
[21] Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B.E., Martin, M.J., McGarvey, P. and Gasteiger, E. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinform. 10, 136.
[22] Johnson, M., Zaretskaya, I., Raytselis, Y., Merezhuk, Y., McGinnis, S. and Madden, T.L. (2008) NCBI BLAST: a better web interface. Nucleic Acids Res. 36, W5–W9.
[23] Lieberman, R.L., Wustman, B.A., Huertas, P., Powe Jr., A.C., Pine, C.W., Khanna, R., Schlossmacher, M.G., Ringe, D. and Petsko, G.A. (2007) Structure of acid beta-glucosidase with pharmacological chaperone provides insight into Gaucher disease. Nat. Chem. Biol. 3, 101–107.
[24] Pearson, W.R. (1998) Empirical statistical estimates for sequence similarity searches. J. Mol. Biol. 276, 71–84.
[25] Marchler-Bauer, A., Anderson, J.B., Chitsaz, F., Derbyshire, M.K., DeWeese-Scott, C., Fong, J.H., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z., Lanczycki, C.J., Liebert, C.A., Liu, C., Lu, F., Lu, S., Marchler, G.H., Mullokandov, M., Song, J.S., Tasneem, A., Thanki, N., Yamashita, R.A., Zhang, D., Zhang, N. and Bryant, S.H. (2009) CDD: specific functional annotation with the Conserved Domain Database. Nucleic Acids Res. 37, D205–D210.
[26] Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. 9, 299–306.

[27] Katoh, K., Misawa, K., Kuma, K. and Miyata, T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30, 3059–3066.
[28] Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
[29] Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52, 696–704.
[30] Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59, 307–321.
[31] Anisimova, M. and Gascuel, O. (2006) Approximate likelihood-ratio test for branches: a fast, accurate, and powerful alternative. Syst. Biol. 55, 539–552.
[32] Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., Dufayard, J.F., Guindon, S., Lefort, V., Lescot, M., Claverie, J.M. and Gascuel, O. (2008) Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 36, W465–W469.
[33] Holm, L., Kaariainen, S., Rosenstrom, P. and Schenkel, A. (2008) Searching protein structure databases with DaliLite v.3. Bioinformatics 24, 2780–2781.
[34] Armougom, F., Moretti, S., Poirot, O., Audic, S., Dumas, P., Schaeli, B., Keduas, V. and Notredame, C. (2006) Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 34, W604–W608.
[35] Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004) WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190.
[36] Rodi, D.J., Mandava, S. and Makowski, L. (2004) DIVAA: analysis of amino acid diversity in multiple aligned protein sequences. Bioinformatics 20, 3481–3489.
[37] Emsley, P. and Cowtan, K. (2004) Coot: model-building tools for molecular graphics. Acta Crystallogr. D 60, 2126–2132.
[38] DeLano, W.L. (2002) The PyMOL User's Manual, DeLano Scientific, Palo Alto.
[39] Ichinose, H., Kotake, T., Tsumuraya, Y. and Kaneko, S. (2008) Characterization of an endo-beta-1,6-Galactanase from *Streptomyces avermitilis* NBRC14893. Appl. Environ. Microbiol. 74, 2379–2383.
[40] Kotake, T., Kaneko, S., Kubomoto, A., Haque, M.A., Kobayashi, H. and Tsumuraya, Y. (2004) Molecular cloning and expression in *Escherichia coli* of a *Trichoderma viride* endo-beta-(1→6)-galactanase gene. Biochem. J. 377, 749–755.
[41] Luo, H., Yang, J., Li, J., Shi, P., Huang, H., Bai, Y., Fan, Y. and Yao, B. (2010) Molecular cloning and characterization of the novel acidic xylanase XYLD from *Bispora* sp. MEY-1 that is homologous to family 30 glycosyl hydrolases. Appl. Microbiol. Biotechnol. 86, 1829–1839.
[42] Hurlbert, J.C. and Preston 3rd, J.F. (2001) Functional characterization of a novel xylanase from a corn strain of *Erwinia chrysanthemi*. J. Bacteriol. 183, 2093–2100.
[43] Nagano, N., Porter, C.T. and Thornton, J.M. (2001) The (betaalpha)(8) glycosidases: sequence and structure analyses suggest distant evolutionary relationships. Protein Eng. 14, 845–855.
[44] Pozharski, E. (2010) Percentile-based spread: a more accurate way to compare crystallographic models. Acta Crystallogr. D Biol. Crystallogr. 66, 970–978.
[45] Devos, D. and Valencia, A. (2000) Practical limits of function prediction. Proteins 41, 98–107.
[46] Rost, B. (1999) Twilight zone of protein sequence alignments. Protein. Eng. 12, 85–94.
[47] Whisstock, J.C. and Lesk, A.M. (2003) Prediction of protein function from protein sequence and structure. Q. Rev. Biophys. 36, 307–340.
[48] Yang, J.K., Yoon, H.J., Ahn, H.J., Lee, B.I., Pedelacq, J.D., Liong, E.C., Berendzen, J., Laivenieks, M., Vieille, C., Zeikus, G.J., Vocadlo, D.J., Withers, S.G. and Suh, S.W. (2004) Crystal structure of beta-D-xylosidase from *Thermoanaerobacterium saccharolyticum*, a family 39 glycoside hydrolase. J. Mol. Biol. 335, 155–165.
[49] Kitago, Y., Karita, S., Watanabe, N., Kamiya, M., Aizawa, T., Sakka, K. and Tanaka, I. (2007) Crystal structure of Cel44A, a glycoside hydrolase family 44 endoglucanase from *Clostridium thermocellum*. J. Biol. Chem. 282, 35703–35711.
[50] Taylor, E.J., Smith, N.L., Turkenburg, J.P., D'Souza, S., Gilbert, H.J. and Davies, G.J. (2006) Structural insight into the ligand specificity of a thermostable family 51 arabinofuranosidase, Araf51, from *Clostridium thermocellum*. Biochem. J. 395, 31–37.
[51] Boraston, A.B., Bolam, D.N., Gilbert, H.J. and Davies, G.J. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem. J. 382, 769–781.