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]

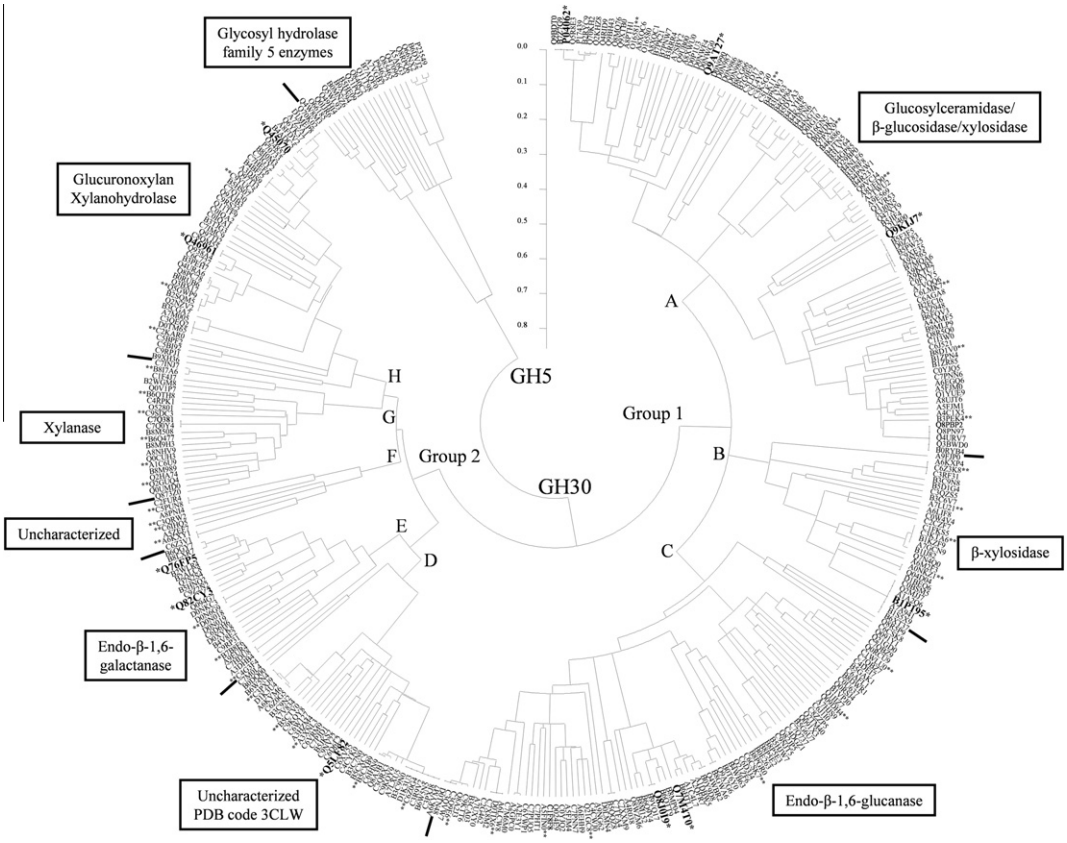

Fig. 1. A Neighbor-Joining bootstrap (1000 replicate) consensus phylogram (see Fig. S10 for visual details) displaying two major groups and up to eight subgroups of GH family 30 enzymes, rooted by the inclusion of a set of GH family 5 endoglucanase and mannanase enzymes. Data reflects the sequence of 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).