**Review**

# A sequence-based protein feature survey on Glycoside Hydrolases Family 28 in Aspergillus

**Master's thesis**

**Jer-gung Chang**

**Affiliation: Utrecht University, Molecular and Cellular Life Science**

**Date: June 2013**

**Supervisor: Dr. Miaomiao Zhou[1]**

**Dr. Ronald de Vries[1]**

**Affiliation supervisors: Fungal Physiology Group, Centre of fungal biodiversity, KNAW, Utrecht. Uppsalalaan 8, 3584 CT Utrecht, the Netherlands**

**Content**


- **Summary** **3**
- **Abstract** **4**
- **Introduction** **4**
  - **Plant biomass** **4**
  - **Pectin** **7**
  - **Glycoside Hydrolases family 28** **8**
  - **Polygalacturonase** **8**
  - **Rhamnogalacturonase** **11**
  - **Xylogalacturonan hydrolase** **12**
  - **GH28 family is abundant in *Aspergilli*** **13**
  - **Comparison of GH28 protein in A. niger (CBS 513.88)** **13**
- **Analysis** **14**
  - **Sequence collection** **14**
  - **Sequence analysis** **19**
  - **Sequence conservation in active site of GH28** **27**
  - **Hidden Markov Model for active site of GH28** **29**
  - **Structure comparison of active site** **29**
- **Discussion and Conclusion** **31**
- **Reference** **32**
- **Supplementary** **36**
  - **Sequence collection** **36**
  - **Sequence analysis** **36**
  - **Hidden Markov Model (HMMdata)** **36**

## ➢ Summary

Biofuel has received attention in recent years because of its sustainability and less eco-toxicity. Plant is an important raw material for biofuel production. However, the sugars embedded in plant cell wall greatly hinder the accessibility of carbon source. Pectin is a component of the cell walls that is composed of acidic sugar-containing backbones with neutral sugar-containing side chains. Glycoside hydrolases family 28 (GH28) is an enzyme family which contribute in pectin hydrolysis and it is classified into endopolygalacturonase, exopolygalacturonase, endorhamnogalacturonase, exorhamnogalacturonase, and xylogalacturonan hydrolase according to their substrate specificities. GH28 is widely present in *Aspergilli* species. *Aspergilli* are known to be suitable for industrial usage because of their well established molecular technique and long history of classical genetic and biochemical study. The Carbohydrate-Active enZYmes Database (CAZy) and PROSITE have provided solid sequence models to identify GH28 proteins from protein sequences even in genome scale. However, by far no detailed sub-classification on enzyme specificity of GH28 members is available.

In this study, sequence analysis on (putative) GH28 enzymes in selected ten *Aspergilli* species were carried out. Sequence features close to the putative catalytic site of each protein were extracted and aligned. In order to know the evolutionary relationships among different enzyme subfamilies within GH28, a phylogenic tree was generated. According to this tree, GH28 enzymes could be clustered into distinct clades which contain members with same substrate specificities. Every group of those enzymes contains distinctive features around the putative active sites. Among all the five groups in GH28, endorhamnogalacturonase is the most distinctive one which has a similarity identified Histidine active site situated outside the catalytic site cleft. 3D structures of representative sequences from each group were extracted or created by homology. Comparison of 3D models was combined with the sequence analysis results. The final outcome implies that the substrate specificities of GH28 enzymes might be strongly affected by the active site sequence composition and their subsequent structures. For each group, manually aligned active site sequences were used to generate a Hidden Markov Model, which could be used to (sub-)classify GH28 enzymes. Furthermore, this study indicates the possible mutation site that may have crucial role(s) in enzyme activities, thus can be used as guides for future experimental validations/applications.

## ➢ Abstract

GH 28 is a large protein family which contributes to pectin hydrolysis. This enzyme family is widely exists in *Aspergilli* species, which are known to be suitable for industrial usage for biomass conversion to produce biofuel. GH28 enzymes were separated into endopolygalacturonase, exopolygalacturonase, endorhamnogalacturonase, exorhamnogalacturonase and xylogalacturonan hydrolase based on their enzymatic specificities. The CAZy Database and PROSITE have provided solid sequence models to identify GH28 proteins from protein sequences, even genome scale. However, by far no detailed classification on enzyme activity of GH28 members is available.

In this research, sequence analysis on enriched GH28 enzymes in *Aspergilli* sp. were carried out and the sequences near known enzymatic active site were aligned. The phylogenetic tree generated by aligning active sites revealed that GH28 enzymes could be clustered into distinct clade according to specific enzyme activities. For each group of enzyme specificity, unique conserved sequence features could be detected. Among all groups, Endo-rhamnogalacturonase was the most distinct group from the GH28 due to the fact that it has an active site identified by similarity located outside the active site cleft. This indicated that the enzyme substrate specificity is strongly affected by the active site structure and therefore, the amino acid sequence composition of the protein. For each group, the manually curated active site alignments were used to generate Hidden Markov Models. These models can be served as methods classify newly identified (putative) *Aspergillus* GH28 enzymes for their functions. Moreover, this report points out the possible mutation sites and effect of mutations that might affect the enzymatic activities of GH28 members, therefore can be used as supports for future experimental targets.


## ➢ Introduction

The limited reservation of fossil fuels accelerates the necessity of development on renewable energy resources [1]. Plants are increasingly used as raw materials in the production of ethanol and other liquid biofuels. But those sugars are embedded in plant cell walls which greatly hinder the accessibility of carbon sources. This is the major challenge to economically viable implementation of these technologies [2].

### ■ Plant biomass

Plant contains unlimited source of renewable energy for mankind and is an important raw material construction materials, textiles, pulp, and paper, as well as many other products [3]. Advent of biotechnology for short-rotation forestry and advances in enzyme technology will allow innovative fiber engineering to alter the

structure, composition, and properties of the raw material [3].

Plant comprise of primary and secondary cell walls, both of which are fortified by cellulose microfibrils. Primary cell walls typically contain cellulose, hemicellulose (xyloglucans), pectin and proteins. In grasses, glucuronoarabinoxylan is cross-linked by diferulate, substitutes for the pectin. Cellulose microfibers are crisscrossed within the cell wall with closer alignment and spacing in primary cell walls than in secondary cell walls. Secondary cell walls are composed of cellulose, hemicellulose and lignin, which constitute the majority of the cell wall mass; for example, 70-80% weight of corn stover is present in secondary cell walls [4].

Pectin plays an important role in efficiency of biofuels production process from raw biomass to feedstocks because pectin can affect the accessibility of other cell wall components to enzymatic degradation and sugars contained in pectin itself represent captured photosynthetic energy [5]. In biomass processing methods, biomass is first treated to disrupt the cell wall structure then saccharified by enzymatic, chemical, or thermal treatment. Nevertheless, the structural properties of cell walls, which has been proposed to be a cellulose –hemicellulose network embedded in pectin matrix [6], (Fig 1.) imply that pectin might mask cellulose and/or hemicelluloses [7], blocking their exposure to degradative enzymes. In fiber hemp processing, pectinase treatment can increase cellulose surface expose to the cellulose degradative enzymes [8].

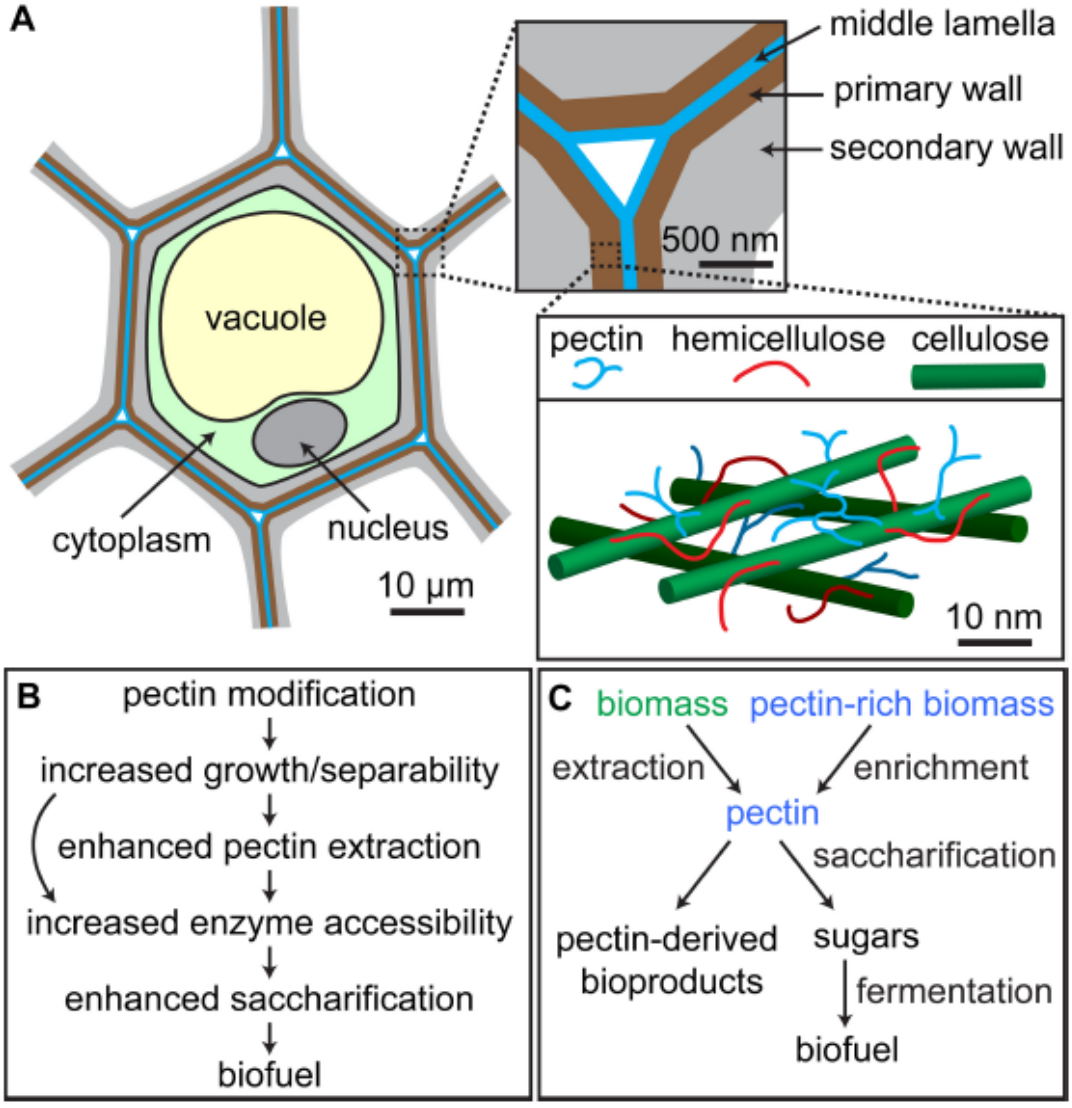


Figure 1. Location and roles of pectin in biomass. A. Schematic of plant cell showing arrangement of cell walls: pectin is abundant in the primary cell wall synthesized by growing cell (brown) in the middle lamella that adhere the cells (blue), but is also present in lower amounts in secondary walls produced after the cessation of growth (grey). Inset at lower right is a simplified model of the primary cell wall showing one possible arrangement of cellulose microfibrils (green), hemicelluloses (red), and pectin (blue). B. Pectin rich biomass can be derived from liginocellulosic feedstocks or naturally pectin-rich plant material, after which it can be processed into pectin derived high-value bioproducts and/or saccharified and fermented into biofuel. C. Promising positive impact of pectin modification in bioenergy crop plants on biomass processing. In some cases, pectin modification might allow for the elimination of processing steps, such as pectin extraction(Adapted from *Xiao, C. et al., 2013* [5]).

## ■ Pectin

Pectin is a complex heteropolysaccharide that hydrates and further cements the primary cell wall matrix. This biopolymer accounts for 30-40% of non-cellulosic polysaccharides in the primary cell walls of herbaceous dicotyledons and non-graminaceous monocots with significantly lesser amounts found in grasses, woody tissue, and secondary cell walls. Pectin consists of long homogalacturonan chains of α-(1-4)-linked D-galacturonic acid and is often esterified with methyl or acetyl groups. Homogalacturonan (HG) is interspersed with the branched polysaccharides, rhamnogalacturonan I (primarily), rhamnogalacturonan II (RGI and II) and xylogalacturonan (XGA) [2, 9]. (Schematic structure of pectin shown in Figure 2) It is desirable to hydrolyse pectin because it blocks cellulase, xylanase and xylan-debranching enzymes from reaching their substrates. In addition, pectin is an important aspect in the conversion of citrus waste and sugar beet pulp into ethanol, where the polysaccharides are abundant [4].

Biodegradation of complex and heterogeneous structure of pectin requires many different enzymatic activities. Exo and endo-polygalacturonan hydrolases (exo and endo-PGs), pectin lyases and pectate lyases degrade HG. XGA can be degraded by XGA hydrolases (endo-XGHs) and exo-PGs, whereas RG hydrolases (RGHs) and RG lyases degrade RG I. Furthermore, the complete enzymatic depolymerization of pectin requires the presence of different type of esterase activities [10-12].

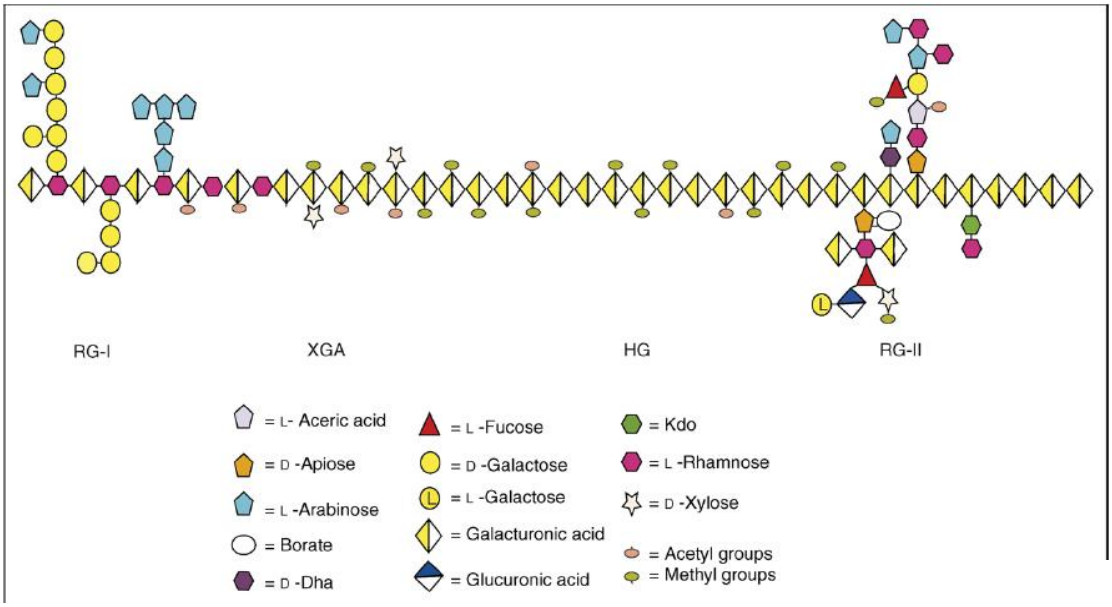


Figure 2. Schematic structure of pectin showing the four pectic polysaccharides homogalacturonan (HG), xylogalacturonan (XGA), rhamnogalacturonan I (RG-I) and rhamnogalacturonan II (RG-II) linked to each other (Adapted from *Mohnen, D. et al.*,

*2008*) [2].

- **Glycoside Hydrolases family 28**

Based on the sequence similarities the glycoside hydrolases (GH) degrading pectin have been classified into the family 28 [13]. GH28 is a set of structural related enzymes that hydrolyze homogalacturonan and rhamnogalacturonan components of pectin, and are important extracellular enzymes found in organisms across the plant, fungal and bacterial kingdoms [14]. This protein family has interesting functional diversity and is variable in copy number among organisms, making this family a likely candidate for birth-and-death evolution. They are involved in diverse biological functions such as fruit ripening, biomass recycling and plant pathogenesis [14]. Some of the previous researches reported that catalytic reaction of the enzymes occur via single displacement inverting mechanism results in fully saturated products with an altered stereochemistry around the anomeric carbon. It is mechanically different from other major class of pectinase like pectate and pectin lyases, which cleave glycosidic linkages by β-elimination, resulting in products with a 4,5-unsaturation at their non-reducing end [14]. GH28 enzymes are categorized into polygalacturonases (PG), which hydrolyze GalA-GalA linkages (E.C.'s 3.2.1.15 [endo-PG] and 3.2.1.67 [exo-PG]), rhamnogalacturonases (RG), which hydrolyze GalA-rhamnose bonds (E.C. 3.2.1.-), and xylogalacturonases (XG), which hydrolyze GalA-xylose bonds (E.C. 3.2.1.-) [10, 13].

The sequence analysis of GH28 has been done 12 year ago by *Markovic O., et al, 2001* [13]. They analyzed the GH28 enzyme sequences cover bacteria, fungi, plants and insect. At that time, GH28 were classified into (1) polygalacturonase (3.2.1.15), (2) exo-polygalacturonase (3.2.1.67), (3) exo-poly-α-galacturonosidase (3.2.1.82), (4) rhamnogalacturonase (3.2.1.-) and (5) endo-xylogalacturonan hydrolase (3.2.1.-). Their sequence analysis result showed that sequence of GH28 diverse between different organisms [13]. The differences between this study and *Markovic O., et al, 2001*are: exo-poly-α-galacturonosidase was included into the exo-polygalacturonase group, and rhamnogalacturonase was separated into endo-RG and exo-RG. In addition, we focus mainly on the *Aspergillus* sp. and those enzymes with crystal structures.

- **Polygalacturonase**

Polygalacturonase (PG) hydrolyze the 1,4-α-D-galactosiduronic linkage in smooth region of pectin, have been biochemically studied which include endo-PG(I, II)and exo-PG (A, B, C, X). *A. niger* PGII is the best characterized PGs [10]. The structure of *A. niger* PGII (N400) has been resolved by crystallography with 1.68Å resolution [15]. The overall structure of *A. niger* PGII folds into a right-handed

parallel β-sheet structure composing 10 complete turns with overall dimensions of approximately 65 Å×35 Å×35 Å. Site-directed mutagenesis studies comparing with the available polygalacturonase sequences identified some highly conserved residues on PGII which includes Asp180, Asp201, Asp202, His223, Gly224, Arg256, and Lys258 [16]. Eight conserved residues form a predominantly negatively charged patch in the cleft. Three conserved Aspartate residues (Asp180, Asp201 and Asp202) appeared critical for catalysis. Asp180 with the assistance of Asp202, was proposed to act as a base to activate the bound water molecule, and Asp201 was tentatively identified as the general acid that protonates the leaving group. His223 was shown to be involved in substrate binding [16].

PG I, A, C and D of *A. niger* have processive behavior which means it do not release the polymer substrate after the hydrolysis reaction, but feed it through the active site cleft for the next cleavage event. *A. niger* PGI has 60% sequence similarity with *A. niger* PGII. Site-directed mutagenesis experiment revealed that Asp153, Asp173 and Asp174 in *A. niger* PGI are involved in its catalytic function, whereas His195, Arg226, Lys228, and Tyr262 function primarily in substrate recognition [17, 18]. The structure of PGI from *A. niger* shows that Arg96 has a crucial role in the processive behavior, is flexible and able to bind oxygen-containing molecules in several well-defined conformations [19]. This may reflect the role of Arg96 in binding the polygalacturonic acid substrate, preventing its release, but at the same time being flexible enough to guide the substrate towards the active site [20].

The crystal structure of *Chondrostereum purpureum* endo-PGI complexed with galacturonic acid allows us to identify some structure features that are involved in substrate recognition [17]. Bond frequency study shows that some endo-PG only hydrolyze the first glycosidic bond from the reducing end of tri and tetragalacturonate, which indicate that form only one productive enzyme-substrate with these substrates and most likely recognize the reducing end galactopyranuronic acid (GalpA) residue at the +1 subsite [21-23]. The endo-PGII mutagenesis identified residues which are His195, Arg226, Lys228, and Tyr262 in endo-PGI can increase the $K_m$ by 10-fold [16, 18]. These residues are within the hydrogen binding distances to the carboxy group of the GalpA which is correlated with the mutagenesis result in endo-PGII [17]. (See Figure 3.) In contrast, the replacement of corresponding residue Asp173, only caused a 2-fold increase $K_m$ value, but greatly decrease $K_{cat}$ value [16]. Asp173 is at a distance to make a hydrogen bond with O4 of the bound GalpA. From this structure, they postulated that Asp173 work as a general acid catalyst that donates a proton to the glycosidic oxygen [24].

In four known endo-PG structures, nonpropyl cis-peptide bond between Gly200 and Ser201 and the Lys228 residues are conserved which postulated to be recognized

by GalfA [15, 25, 26]. The cis-peptide bond and Lys228 are also conserved in the structure of *A. aculeatus* endo-RGA [27]. The structure conservation of three residues implies that the cis-peptide bond and Lys228 residues possibly form a carboxy group recognition motif in the -1 subsite of both endo-PG and endo-RG [17]. GalfA and GalpA structural model were constructed with a substrate molecule bond in both -1 and +1 subsite of both across from the catalytic residues. (Shown in Figure 3.)

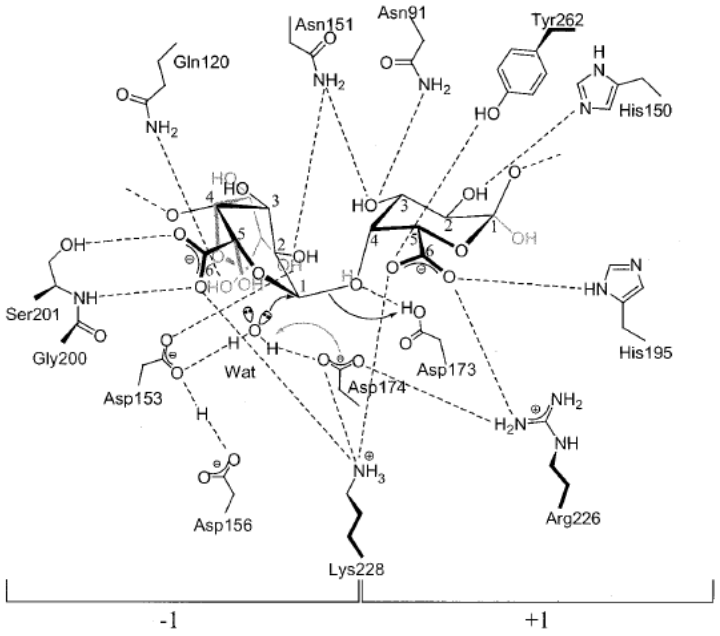


Figure 3. Schematic drawing of the proposed structure for substrate binding in the -1 and +1 subsites. The proposed substrate (black) was modeled based on structure of GalpA and GalfA (galactofuranuronic acid) molecules (grey) The nuleophilic water is in the preferred position to attack the C1 atom of the GalpA unit in the -1 subsite [17]. (Adapted from *Pages S., et al., 2000*)

Previous published sequence analysis of GH28 enzymes have shown that fungi endo-PGs have highly conserved Cysteine, furthermore, the conservation of Cysteine is different between organisms and protein groups. These enzymes use disulfide bonds to stabilize its molecules structures and the position of those Cysteines reflects the taxonomy of the enzymes is very similar to each other [13].

Exo-PG structure has been resolved in *Yersinia enterocolitica* by *Abbott D. W., et al., 2007* [14]. Its exo-enzyme activity is caused by insertion stretches of amino acid residues that transformed the active site from the open-ended channel observed in the endo-PG to close the pocket that limit the enzyme to the exclusive attack of the non-reducing end of oligogalacturonide substrates, however, endo-PG also has this sequence of loop but it oriented to different direction without blocking the active site [14]. By analyzing the interaction of digalacturonic acid (diGalUA) with *Y.*

*enterocolitica* exo-PG active site reveal that there are some basic amino acid residues participate in stabilizing the residue in -2 subsite, which also function to enclose the non-reducing end of the active site and position the substrate for hydrolysis (Fig. 4) [14].

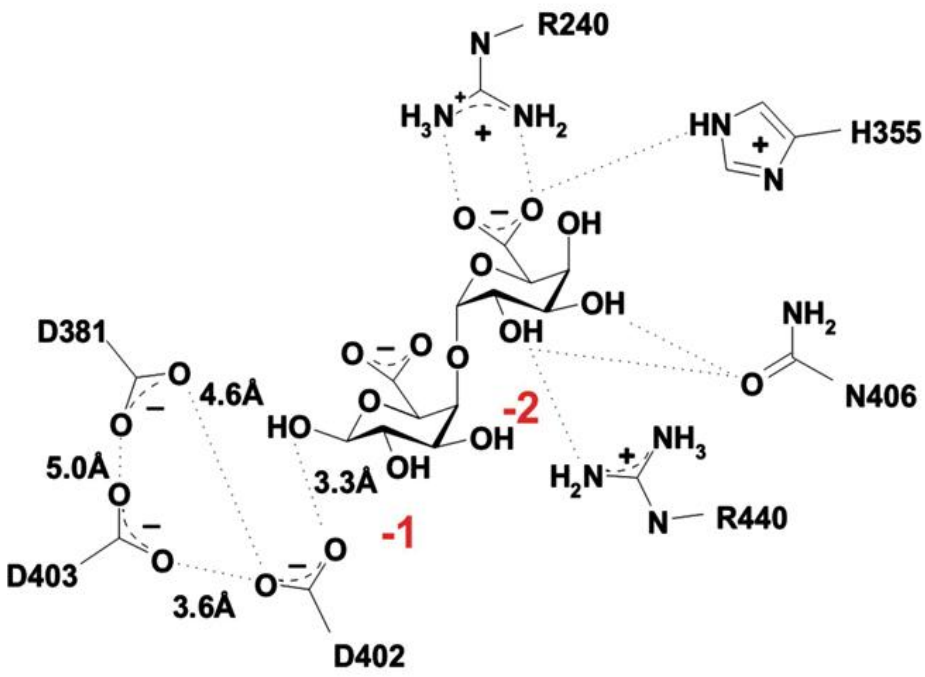


Figure 4. Schematic representation of *Y. enterocolitica* exo-PG active site -1 and -2 interaction with diGalUA. Adapted from *Abbott D. W., et al., 2007* [14].

- **Rhamnogalacturonase**

Rhamnogalacturonase (RG) includes endo-RG and exo-RG hydrolyze α-D-galacturonopyranosyl-(1, 4)- α-L-rhamnopyranosyl linkages in the backbone of hairy region in pectin [11]. Earlier sequence analysis done by *Markovic, O. 2001* [13] shows that amino acid sequences of endo-RG are highly conserved among different organisms, and ten Cysteins were found to be conserved in alignment of endo-RG in their studies [13]. Furthermore, there are 13 Tyrosines, 6 Phenylalanines and 4 Tryptophans invariantly conserved in the four endo-RG they analyzed, Trp200 and Trp302 being found present in the active site of the *A. aculeatus* endo-RG [13]. The active site of exo-RG is in a geometry that is similar to PG which has Aspartates near residue 200 and a Histidine after the Aspartate active site. Endo-RG has longer distance between the active site Aspartate and Histidine which indicates Histidine might locate at different site compare to other GH28 enzyme. The structure of endo-PG from *Pectobacterium carotovora* ssp. *carotovora* was compared with endo-RGA from *A. aculeatus* which identified two conserved Aspartates participate in

the catalytic activities. In addition, Phe175 (151), Lys229 (203), and Lys282 (253) are conserved (Shown in Figure 5) [28]. In both *P. carotovora* endo-PG and *A. aculeatus* endo-RGA a water molecule is hydrogen bounded between Asp202 (177) and Asp223 (197). A second conserved water molecule forms hydrogen bonds to Asp202, (Asp/Glu)224 (198) and Lys229 (203) [28]. The *A. aculeatus* endo-RG is lack of Histidine at the active site cleft [28].

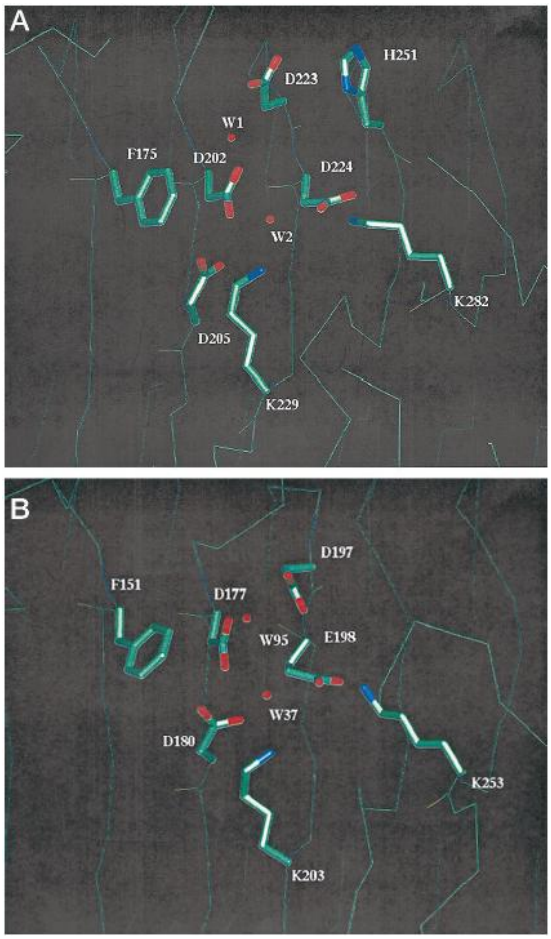


Figure 5A. Active site cleft of *P. carotovora* endo-PG. 5B. Active site cleft of *A. aculeatus* endo-RGA. Figure was produced by using QUANTA after superimposition of the two structure using O. Asp202 (177) and 223 (197) are conserved between two enzymes and they are therefore proposed to be the catalytic Aspartates. His251 is only found in active site of *P. carotovora* endo-PG, however, Phe175 (151), Lys229 (203), and Lys282 (253) are common in both enzymes. Water molecules W1 (95) and W2 (37) are conserved between two enzymes [28]. (Adapt from *Pickersgill, R., et al., 1998*)

- **Xylogalacturonan hydrolase**

Xylogalacturonan hydrolase (XGH) hydrolyzes α-D-galacturonopyranosyl-(1, 4)-β-D-xylosyl linkages [12]. XGH acted from non-reducing end towards the reducing end of the substrate xylogalacturonan (XGA) processively. The endo-XGH from *A. tubingensis* was shown to have both endo-XGH and exo-XGH activity from the XGA degradation assay [12]. Previous BLAST search done by *Markovic, O. 2001* shows XGH from *A. tubigensis* has high similarity with exo-PG from *Cochliobolus carbonum* and two sequences have 39.9% identity and 55.4% similarity [13]. In spite of this high sequence similarity to fungal exo-PG, the XGH sequence does not contain most of the conserved regions characteristic of fungal exo-PG, thus implying its

enzymatic uniqueness [13]. However, it has been reported that *Aspergillus* sp. exo-PG could degrade XGA [12]. Moreover, the active site of XGH is also similar to exo-PG which includes Aspartate at around residue 200 and Histidine after the Aspartate active site [12, 13].

Though XGH is recognized as an endo-enzyme, it mainly behaves in an exo-lytic way during degradation of XGA [12]. XGH works on $Gal_4Xyl_3$ from the non-reducing end towards the reducing end, which implies that it is an exo-acting enzyme [29, 30]. This kind of exo-acting character correlates with the high sequence similarity with exo-PG while comparing with endo-PG [12]. Stepwise release of the GalAXyl from $Gal_4Xyl_3$ indicates that XGH has processive behavior [31].

- **GH28 family is abundant in *Aspergilli***

*Aspergillus niger* is an excellent producer in the industry for pectinolytic enzyme production because comparing with other produced by many other fungi in same genus, it has a GRAS (general mark as safe) status. Due to its important industrial usage, *Aspergilli* have been studied a lot and many GH28 in *Aspergilli* have been identified in different strains. Various strains have been sequenced allow us to compare the GH28 within *Aspergilli*. Here by we compared the GH28 in *A. niger* and 9 other GH28 in *Aspergillus* genus including some pathogenic species to generate a phylogeny tree. A further attempt was carried out to find out which residue is important for the enzymatic functionality of the protein.

- **Comparison of GH28 protein in *A.niger* (CBS 513.88)**

Previous study done by *Elena S. M. et al, 2006* used sequence alignment of complete *A. niger* family GH28 analyzed by using program T-coffee and manually curation, then create corresponding sequence distance dendrogram by using Neighbor-Joining Method (Figure 6A) [32]. Based on the sequence similarity three major groups were observed. The first group was endo-PG. The second group contained exo-PG and exo-RG. The most diverged sequence was exo-PGC (PGXC in figure). The last group was endo-RG which has no active site Histidine. At last XGH, did not group together with any other sequences [32].

By in depth inspection of the enzyme sequence (Figure 6B) revealed that previous identified catalytic Aspartates residues are conserved in *A. niger* GH28 except for alignment position 362 where endo-RG are replaced with Glutamate residue. The substrate binding residue Lysine at alignment position 428 conserved throughout the *A. niger* GH28, while second substrate binding residue Arginine 426 only present in endo-PG and other four exo-group enzyme includes exo-PGA, exo-PGX, exo-PGB, exo-PGC. Their alignment also shows that Histidine 386, Serine

389, and Tyrosine 479 which considered carry out substrate recognition are conserved within the same set of enzyme groups [32].

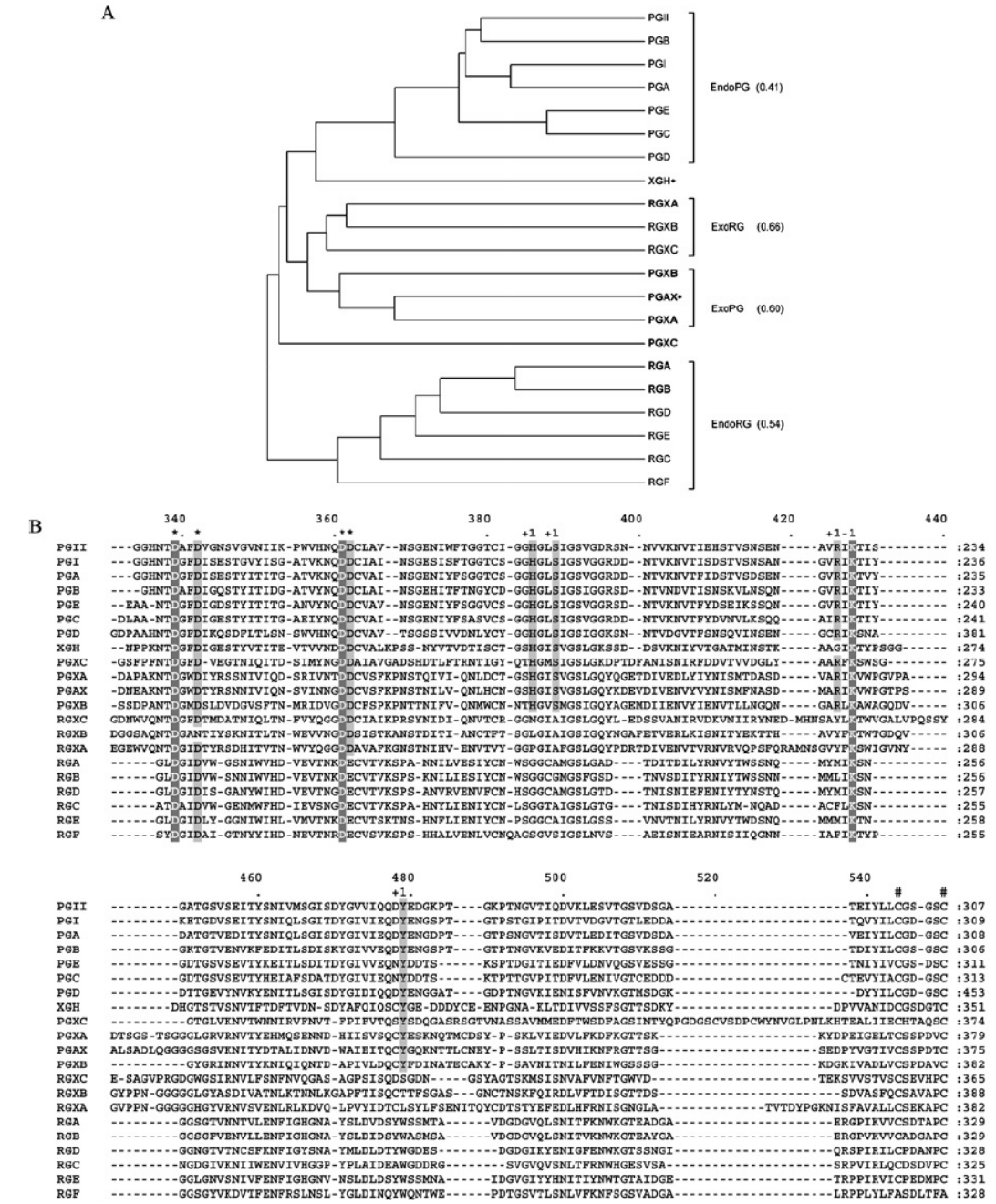


Figure 6. A. Phylogeny tree of *A. niger* GH28. B. Region of multiple alignment of *A. niger* family GH28, adapted from *Elena S. M. et al, 2006* [32].

- **Analysis**

- **Sequence collection**

The (putative) family 28 glycoside hydrolases was obtained from the Carbohydrate-Active enZYmes Database (CAZy) [33]. Corresponding protein

sequences were retrieved from UniProt (www.uniprot.org), GenBank, or AspGD (www.aspgd.org) databases. This dataset included GH28 from *Aspergillus niger* (strain CBS.120.49/N400, CBS 513.88, M1, RH5344), *A. nidulans* (FGSC A4), *A. aculeatus* (KSM 510, CBS 115.80), *A. oryzae* (RIB 40), *Gibberella moniliformis* (FC-10), *Chondrostereum purpureum*, *Botryotinia fuckeliana* (WS38), *Colletotrichum lupini* var. setosum (SHK788), *Pectobacterium carotovorum* (SCC3193), *Thermotoga maritima* (MSB8), *Yersinia enterocolitica* (ATCC 9610D), *Medicago sativa*, *Juniperus ashei* Several of the proteins with available structural information were obtained from Protein Data Base (PDB) for further structural analysis and comparison (table 1).

The dataset was further enriched by adding orthologous gene sequences acquired from 10 *Aspergilli* genomes. The COG (Cluster of Orthologous Groups of proteins) of 10 *Aspergilli* genomes were performed by using the program orthoMCL (parameter: e value1e-5, inflation 1, coverage 50) with the completed genomes of *A. clavatus* (NRRL 1), *A. flavus* (NRRL 3357), *A. fumigates* (A1163), *A. niger* (ATCC 1015), *A. terreus* (NIH 2624), *A. fischeri* (NRRL 181), *A. niger* (CBS 513.88), *A. fumigates* (A1163), *A. oryzae* (RIB 40), *A. nidulans* (FGSC A4).
Collected sequences are indicated in Table 1.

Table 1

| Organism | Gene name | PDB | Protein | key residues | PDB residues number | mutation | effect | reference |
|---|---|---|---|---|---|---|---|---|
| *A. niger* (N400) | P26214 | 1CZF | Endo-PGII (3,2,1,15) | D180, D201, D202, H223 | D180, D201, D202 | Y326L, N186E, E252A | Equal or increased specific activity compare to wild type | [18] |
| | | | | | | D183N, Y291F, Y291L | decreased specific activity | |
| | | | | | | D180E, D201E, D202E, D202N, H223A | decrease activity significantly without effect its Km toward substrates | |
| | | | | | | R256N, K258N | decrease activity significantly and reduce Km for substrate 10-fold | |
| *A. niger* (N400) | P26213 | 1NHC | Endo-PGI (3.2.1.15) | R96, D186,D207, D208, H229 | R96 | R96S | loss enzyme processivity | [20, 34] |
| *A. niger* (CBS 513.88) | An01g11520 | | Endo-PGI (3.2.1.15) | D207, H229 | | | | [35] |
| | An15g05370 | | Endo-PGII(3,2,1,15) | D201, H223 | | | | |
| | An16g06990 | | Endo-PGA(3.2.1.15) | D207, H229 | | | | |
| | An02g04900 | | Endo-PGB(3.2.1.15) | D199, H221 | | | | |
| | An05g02440 | | Endo-PGC(3.2.1.15) | D222, H244 | | | | |
| | An09g03260 | | Endo-PGD(3.2.1.15) | D337, H359 | | | | |
| | An01g14670 | | Endo-PGE(3.2.1.15) | D219, H241 | | | | |
| | An11g04040 | | Exo-PGA (3.2.1.67) | D246, H269 | | | | |
| | An03g0674 | | Exo-PGB (3.2.1.67) | D255, H278 | | | | |
| | An02g12450 | | Exo-PGC (3.2.1.67) | D229, H253 | | | | |
| | An12g07500 | | Exo-PGX (3.2.1.67) | D244, H267 | | | | |
| | An01g14650 | | Exo-RGA (3.2.1.67) | D237 | | | | [32, 35] |
| | An03g02080 | | Exo-RGB (3.2.1.40) | D231 | | | | |
| | An18g0410 | | Exo-RGC (3.2.1.67) | D229 | | | | |
| | An04g09700 | | Endo-XGHA (3.2.1.-) | D228, H251 | | | | |
| | An12g00950 | | Endo-RGA (3.2.1.171) | D216, H291 | | | | [35] |
| | An14g04200 | | Endo-RGB (3.2.1.171) | D219, H294 | | | | |
| | An06g02070 | | Endo-RGC (3.2.1.-) | D216, H290 | | | | [32, 35] |
| | An11g06320 | | Endo-RGD (3.2.1.-) | D215 | | | | |
| | An11g08700 | | Endo-RGE (3.2.1.-) | D221, H296 | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | An07g01000 | | Endo-RGF (3.2.1.-) | | | | | [32] |
| *A. niger* (N400) | Q27UB0 | | Exo-PGC (3.2.1.67) | D229, H253 | | | | |
| | Q2EQQ2 | | Exo-RGA (3.2.1.171) | D237 | | | | |
| | Q1ZZM4 | | Endo-RGF (3.2.1.-) | | | | | |
| | Q1ZZM3 | | Endo-RGE (3.2.1.-) | D222, H297 | | | | |
| | Q9P4W2 | | Endo-PGD (3.2.1.15) | D337, H359 | | | | [36] |
| | Q9P4W4 | | Endo-PGA (3.2.1.15) | D207, H229 | | | | [37] |
| | Q9P4W3 | | Endo-PGB (3.2.1.15) | D199, H221 | | | | |
| | O42809 | | Endo-PGE (3.2.1.15) | D219, H241 | | | | [38] |
| | Q12554 | | Endo-PGC (3.2.1.15) | D221, H243 | | | | [39] |
| | P87160 | | Endo-RGA (3.2.1.171) | D216 | | | | [40] |
| | P87161 | | Endo-RGB (3.2.1.171) | D219, H294 | | | | |
| *A. niger* (M1) | C6KLC9 | | Endo-PGI (3.2.1.15) | | | | | |
| *A. niger* (RH5344) | 1568342 | | Endo-PG (3.2.1.15) | | | | | |
| *A. aculeatus* (KSM 510) | O74213 | 1IA5, 1IB4 | Endo-PGI(3.2.1.15) | D219, H241 | D180, H202 | | | [25] |
| | Q00001 | 1RMG | Endo-RGA (3.2.1.171) | D215, H290 | D197, H272 | | | [11, 24, 26, 27, 41] |
| *A. nidulans* FGSC A4 | AN8891.2 | | Exo-PGB (3.2.1.67) | D260, H283 | | | | [42, 43] |
| | AN3389.2 | | Endo-XGHA (3.2.1.-) | D222, H245 | | | | |
| | AN8761.2 | | Exo-PGX-1 (3.2.1.67) | D245, H268 | | | | [42-44] |
| | AN10274.3 | | Endo-PGA (3.2.1.15) | | | | | |
| | AN9134.2 | | Endo-RGA (3.2.1.171) | D215, H290 | | | | |
| | AN4372.2 | | Endo-PGB (3.2.1.15) | D203, H225 | | | | |
| | AN9045.2 | | Exo-PGX-2 (3.2.1.67) | D254, H277 | | | | |
| | AN8327.2 | | Endo-PG (3.2.1.15) | D215, H237 | | | | |
| | AN6656.2 | | Endo-PGD (3.2.1.15) | D356,H378 | | | | [42, 43] |
| *A. oryzae* KBN616 | gi\|404092 | | Endo-PG (3.2.1.15) | | | | | [31] |
| | Q2UHL4 | | Endo-PGI (3.2.1.15) | D206, H228 | | | | [42, 45] |
| *A. oryzae* RIB40 | AO090009000470 | | Exo-RGB (3.2.1.171) | | | | | [42] |
| | AO090005000067 | | Endo-RGA (3.2.1.171) | | | | | |
| | AO090005000186 | | Endo-PGD(3.2.1.15) | D334, H356 | | | | |
| | AO090005001400 | | Exo-PGC (3.2.1.67) | D229, H253 | | | | |
| | AO090001000133 | | Exo-PGA (3.2.1.67) | D236 | | | | |
| | AO090003000524 | | Endo-RGA (3.2.1.171) | | | | | |
| | AO090023000161 | | Endo-PGB (3.2.1.15) | D202, H224 | | | | [31] |
| | AO090023000401 | | Endo-PGI(3.2.1.15) | D206, H228 | | | | [42, 45] |
| | AO090026000120 | | Endo-XGH (3.2.1.-) | | | | | [42] |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | AO090026000252 | | Endo-RGC (3.2.1.-) | D217, H291 | | | | |
| | AO090026000784 | | Exo-PG (3.2.1.15) | | | | | |
| | AO090102000011 | | Endo-XGHA (3.2.1.-) | D228, H251 | | | | |
| | AO090102000139 | | Exo-RGB (3.2.1.-) | | | | | |
| | AO090124000009 | | Endo-RGE (3.2.1.-) | D221, H296 | | | | |
| | AO090113000199 | | Exo-RGB (3.2.1.-) | | | | | |
| | AO090038000552 | | Endo-RGA (3.2.1.171) | D217, H292 | | | | |
| | AO090010000484 | | Endo-RGB (3.2.1.171) | D219, H294 | | | | |
| | AO090010000753 | | Exo-PGB (3.2.1.67) | D254, H277 | | | | |
| *Medicago sativa* | Q40312 | | PG (3.2.1.15) | D219, H242 | | | | |
| *Chondrostereum purpureum* | Q9P8M3 | | Endo-PGC (3.2.1.15) | | | | | [46] |
| | P79074 | 1K5C, 1KCC, 1KCD | Endo-PGI (3.2.1.15) | D153, D173, D174 | +24 | | | [17, 47, 48] |
| *Botryotinia fuckeliana* (WS38) | A4VB48 | | Endo-PGII (3.2.1.15) | | | | | [49, 50] |
| *Juniperus ashei* | Q9FY19 | | Endo-PG (3.2.1.15) | D256, H279 | | | | [51] |
| *Gibberella moniliformis* (FC-10) | Q07181 | 1HG8 | PG (3.2.1.15) | D212, H234 | D212, H234 | | | [52, 53] |
| *Pectobacterium carotovorum* SCC3193 | P26509 | 1BHE | Endo-PG (3.2.1.15) | D228, D249, D250, H277 | D202, D223, D224 H251 | | | [28, 54, 55] |
| *Thermotoga maritima* (MSB8) | Q9WYR8 | 3JUR | Exo-PG (3.2.1.67) | | | | | [56, 57] |
| *Yersinia enterocolitica* (ATCC 9610D) | O68975 | 2UVE, 2UVF | Exo-PG (3.2.1.82) | | | | | [14, 58] |
| *Colletotrichum lupini* var. setosum SHK788 | A1E266 | 2IQ7 | Endo-PG (3.2.1.15) | | +23 | | | [59] |

Table 1. Abbreviations: polygalacturonase: PG, rhamnogalacturonase: RG, xylogalacturonase hydrolase: XGH. Other orthologous gene sequences from the *Aspergillus* were also included in the sequence analysis are shown in the supplementary data (Aspergillus_sequences.FASTA).

- **Sequence analysis**

Sequence alignment of protein sequences were performed by using MAFFT [60] to align the amino acids obtained from data collection. Then, G-Blockss [61] was used to retrieve conserved sites of GH28. Following this step, both results were visualized by Jalview [62] in order to perform manual curation on the sequence alignment [63]. The phylogeny tree was generated by average distance using BLOSUM62 matrix in Jalview.

MAFFT is a kind of high performance multiple sequence alignment (MSA) program based on sequence similarity. It could be useful for evolutionary information because the sequence to be aligned were generated from a common ancestor in the course of revolution [60]. MAFFT assumes that the input sequences are all homologous which are descended from a common ancestor. Therefore, all the letters in the input data are aligned. Genomic rearrangement or domain shuffling is not assumed, thus MAFFT preserved all the order of letters in the sequence [60].

The GH28 protein sequences were analyzed by online version of MAFFT (http://mafft.cbrc.jp/alignment/server/). The output of MAFFT result was visualized by the Jalview (version 2.8) and generated the phylogeny tree by average distance using BLOSUM62 matrix. The phylogeny tree is shown in Figure 7. In the figure we can see that the MAFFT separated most of the GH28 enzymes into different groups and the topology of the phylogeny tree is very similar to the one created by *Elena S. M., et al, 2006* (Figure 6).

The group endo-RG is the most distant from the rest of the GH28 family which has different active site than the others which was mentioned before in the RG introduction. The endo-RG is separated into a distinct group in the bottom. XGH and exo-PG were on the same clade which means they have high sequence similarity compare to other group, which may be relate to the fact that exo-PG can also carry out XHG activities [12]. Exo-enzymes exo-RG and exo-PG are on the same clade just like the phylogeny tree from *Elena S. M., et al, 2006*. The endo-PG is separated as a clade on the top. The subgroups were not separated very well might because the difference between the species along is more than subgroups, however, some subgroup still clustered together.

Some of the GH28 enzymes are not grouped at the correct clade, such as some enzyme from outside of the genus *Aspergillus*, these protein sequences may be diverse due to the fact that they are from entirely different organism or the alignment is simply not good enough.

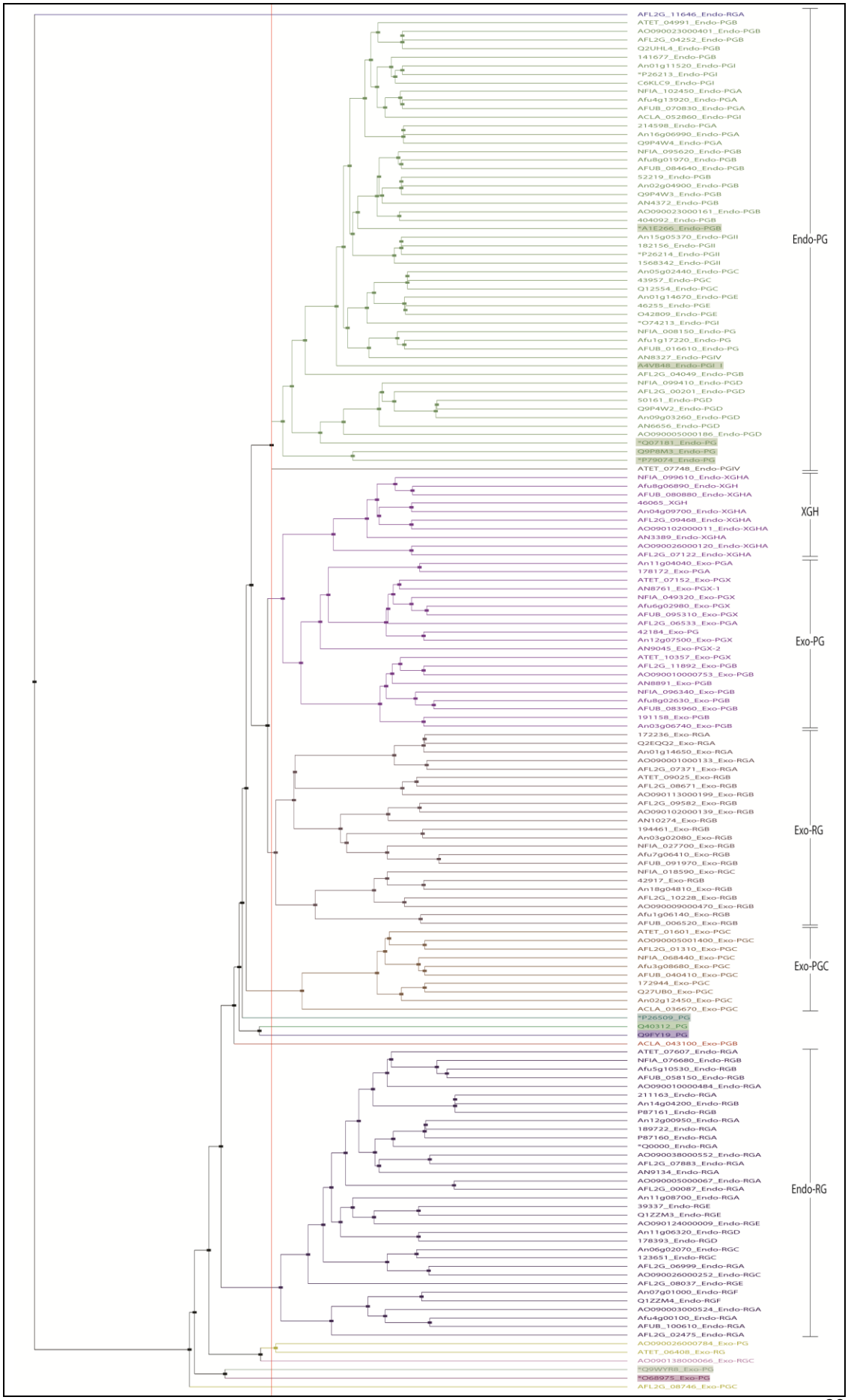
AFL2G_11646_Endo-RGA
ATET_04991_Endo-PGB
AO090023000401_Endo-PGB
AFL2G_04252_Endo-PGB
Q2UHL4_Endo-PGB
141677_Endo-PGB
An01g11520_Endo-PGI
*P26213_Endo-PGI
C6KLC9_Endo-PGI
NFIA_102450_Endo-PGA
Afu4g13920_Endo-PGA
AFUB_070830_Endo-PGA
ACLA_052860_Endo-PGI
214598_Endo-PGA
An16g06990_Endo-PGA
Q9P4W4_Endo-PGA
NFIA_095620_Endo-PGB
Afu8g01970_Endo-PGB
AFUB_084640_Endo-PGB
52219_Endo-PGB
An02g04900_Endo-PGB
Q9P4W3_Endo-PGB
AN4372_Endo-PGB
AO090023000161_Endo-PGB
404092_Endo-PGB
*A1E266_Endo-PGB
An15g05370_Endo-PGII
182156_Endo-PGII
*P26214_Endo-PGII
1568342_Endo-PGII
An05g02440_Endo-PGC
43957_Endo-PGC
Q12554_Endo-PGC
An01g14670_Endo-PGE
46255_Endo-PGE
O42809_Endo-PGE
*O74213_Endo-PGI
NFIA_008150_Endo-PG
Afu1g17220_Endo-PG
AFUB_016610_Endo-PG
AN8327_Endo-PGIV
A4VB48_Endo-PGI_I
AFL2G_04049_Endo-PGB
NFIA_099410_Endo-PGD
AFL2G_00201_Endo-PGD
50161_Endo-PGD
Q9P4W2_Endo-PGD
An09g03260_Endo-PGD
AN6656_Endo-PGD
AO090005000186_Endo-PGD
*Q07181_Endo-PG
Q9P8M3_Endo-PG
*P79074_Endo-PG
ATET_07748_Endo-PGIV
NFIA_099610_Endo-XGHA
Afu8g06890_Endo-XGH
AFUB_080880_Endo-XGHA
46065_XGH
An04g09700_Endo-XGHA
AFL2G_09468_Endo-XGHA
AO090102000011_Endo-XGHA
AN3389_Endo-XGHA
AO090026000120_Endo-XGHA
AFL2G_07122_Endo-XGHA
An11g04040_Exo-PGA
178172_Exo-PGA
ATET_07152_Exo-PGX
AN8761_Exo-PGX-1
NFIA_049320_Exo-PGX
Afu6g02980_Exo-PGX
AFUB_095310_Exo-PGX
AFL2G_06533_Exo-PGA
42184_Exo-PG
An12g07500_Exo-PGX
AN9045_Exo-PGX-2
ATET_10357_Exo-PGX
AFL2G_11892_Exo-PGB
AO090010000753_Exo-PGB
AN8891_Exo-PGB
NFIA_096340_Exo-PGB
Afu8g02630_Exo-PGB
AFUB_083960_Exo-PGB
191158_Exo-PGB
An03g06740_Exo-PGB
172236_Exo-RGA
Q2EQQ2_Exo-RGA
An01g14650_Exo-RGA
AO090001000133_Exo-RGA
AFL2G_07371_Exo-RGA
ATET_09025_Exo-RGB
AFL2G_08671_Exo-RGB
AO090113000199_Exo-RGB
AFL2G_09582_Exo-RGB
AO090102000139_Exo-RGB
AN10274_Exo-RGB
194461_Exo-RGB
An03g02080_Exo-RGB
NFIA_027700_Exo-RGB
Afu7g06410_Exo-RGB
AFUB_091970_Exo-RGB
NFIA_018590_Exo-RGC
42917_Exo-RGB
An18g04810_Exo-RGB
AFL2G_10228_Exo-RGB
AO090009000470_Exo-RGB
Afu1g06140_Exo-RGB
AFUB_006520_Exo-RGB
ATET_01601_Exo-PGC
AO090005001400_Exo-PGC
AFL2G_01310_Exo-PGC
NFIA_068440_Exo-PGC
Afu3g08680_Exo-PGC
AFUB_040410_Exo-PGC
172944_Exo-PGC
Q27UB0_Exo-PGC
An02g12450_Exo-PGC
ACLA_036670_Exo-PGC
*P26509_PG
Q40312_PG
Q9FY19_PG
ACLA_043100_Exo-PGB
ATET_07607_Endo-RGA
NFIA_076680_Endo-RGB
Afu5g10530_Endo-RGB
AFUB_058150_Endo-RGB
AO090010000484_Endo-RGA
211163_Endo-RGA
An14g04200_Endo-RGB
P87161_Endo-RGB
An12g00950_Endo-RGA
189722_Endo-RGA
P87160_Endo-RGA
*Q9000_Endo-RGA
AO090038000552_Endo-RGA
AFL2G_07883_Endo-RGA
AN9134_Endo-RGA
AO090005000067_Endo-RGA
AFL2G_00087_Endo-RGA
An11g08700_Endo-RGA
39337_Endo-RGE
Q1ZZM3_Endo-RGE
AO090124000009_Endo-RGE
An11g06320_Endo-RGD
178393_Endo-RGD
An06g02070_Endo-RGC
123651_Endo-RGC
AFL2G_06999_Endo-RGA
AO090026000252_Endo-RGC
AFL2G_08037_Endo-RGE
An07g01000_Endo-RGF
Q1ZZM4_Endo-RGF
AO090003000524_Endo-RGA
Afu4g00100_Endo-RGA
AFUB_100610_Endo-RGA
AFL2G_02475_Endo-RGA
AO090026000784_Exo-PG
ATET_06408_Exo-RG
AO090138000066_Exo-RGC
*Q9WYR8_Exo-PG
*O68975_Exo-PG
AFL2G_08746_Exo-PGC
Endo-PG
XGH
Exo-PG
Exo-RG
Exo-PGC
Endo-RG

Figure 7. Phylogeny tree of sequence alignment generated by MAFFT. "*" indicates that the protein has a crystalized structure (PDB file available). The shaded proteins are outside of genus *Aspergillus*. On the right indicates the different group named by its enzyme activity.

The sequences were not really well aligned because active site Histidine is separated in endo-RG. Manual curation by Jalview is needed to have a better alignment. During the manual curation, active site sequence near Histidine was aligned together and gaps were removed. In the total alignments, very long gaps created by less than 10% of the proteins were as well removed in order to keep only conserved sequence features. A few sequences acquired from the CAZy only contain fragment of the enzymes, these kind of incomplete protein sequences were removed. The excluded sequences include AO090138000067_exorhamnogalacturonase_C (lack C-terminal part of protein sequence), AO090138000066_exorhamnogalacturonase_C (lack N-terminal part of protein sequence), and ATET_07748_endopolygalacturonase_IV (lack C-terminal part of protein sequence).

After the active site Aspartate and Histidine were aligned, a large gap between active site Aspartate and Histidine except the endo-RG was created. Serine 389 mentioned by *Elena S. M. et al, 2006* were also aligned in the endo-RG with active site Histidine. Serine 389 form *Elena S. M. et al, 2006* could not be found back in endo-RG after aligned the active site Histidine, however, Serine 392 could be aligned in some endo-RG enzymes. Region of multiple alignment of *A. niger* and enzymes with 3D structure GH28 enzymes is shown in Figure 8.

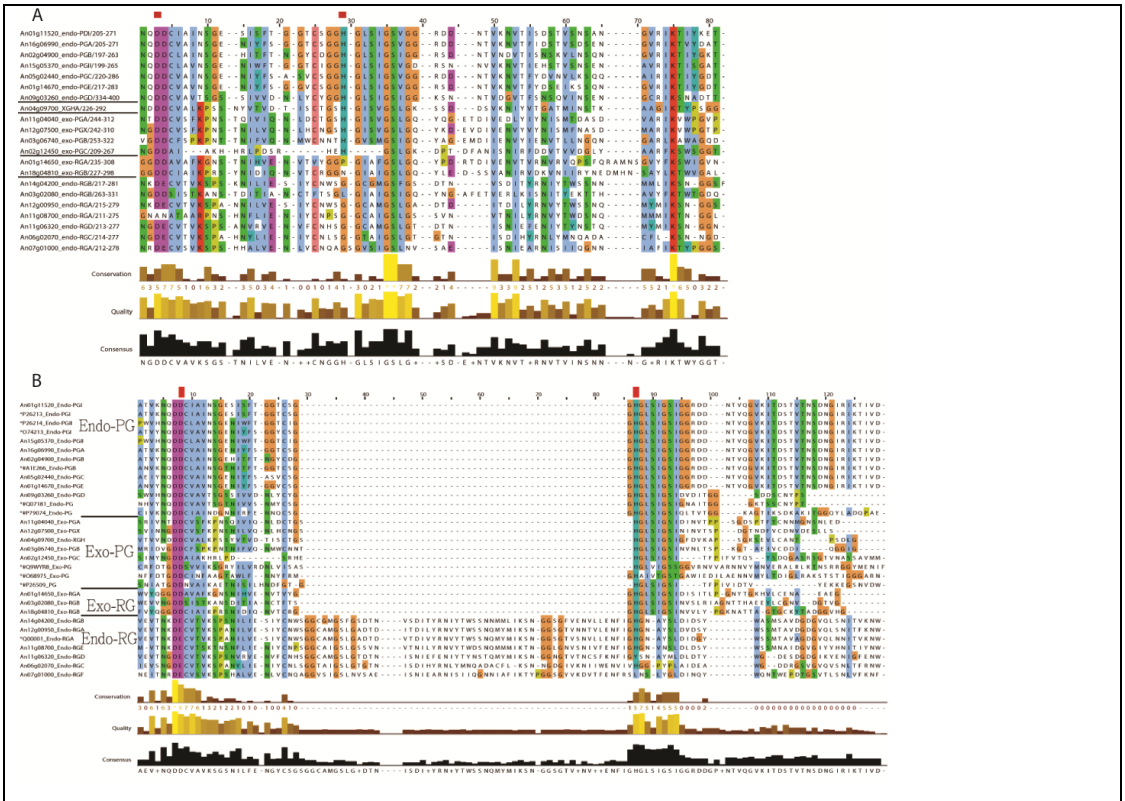


Figure 8. Multiple alignment of active site GH28 enzymes. A. Before manual curation. B. After manual curation. The alignment is visualized by Jalview. Red boxes indicate the position of active site. "*" indicates the enzyme with 3D structure. "#" signifies the enzyme outside the genus *Aspergillus*.

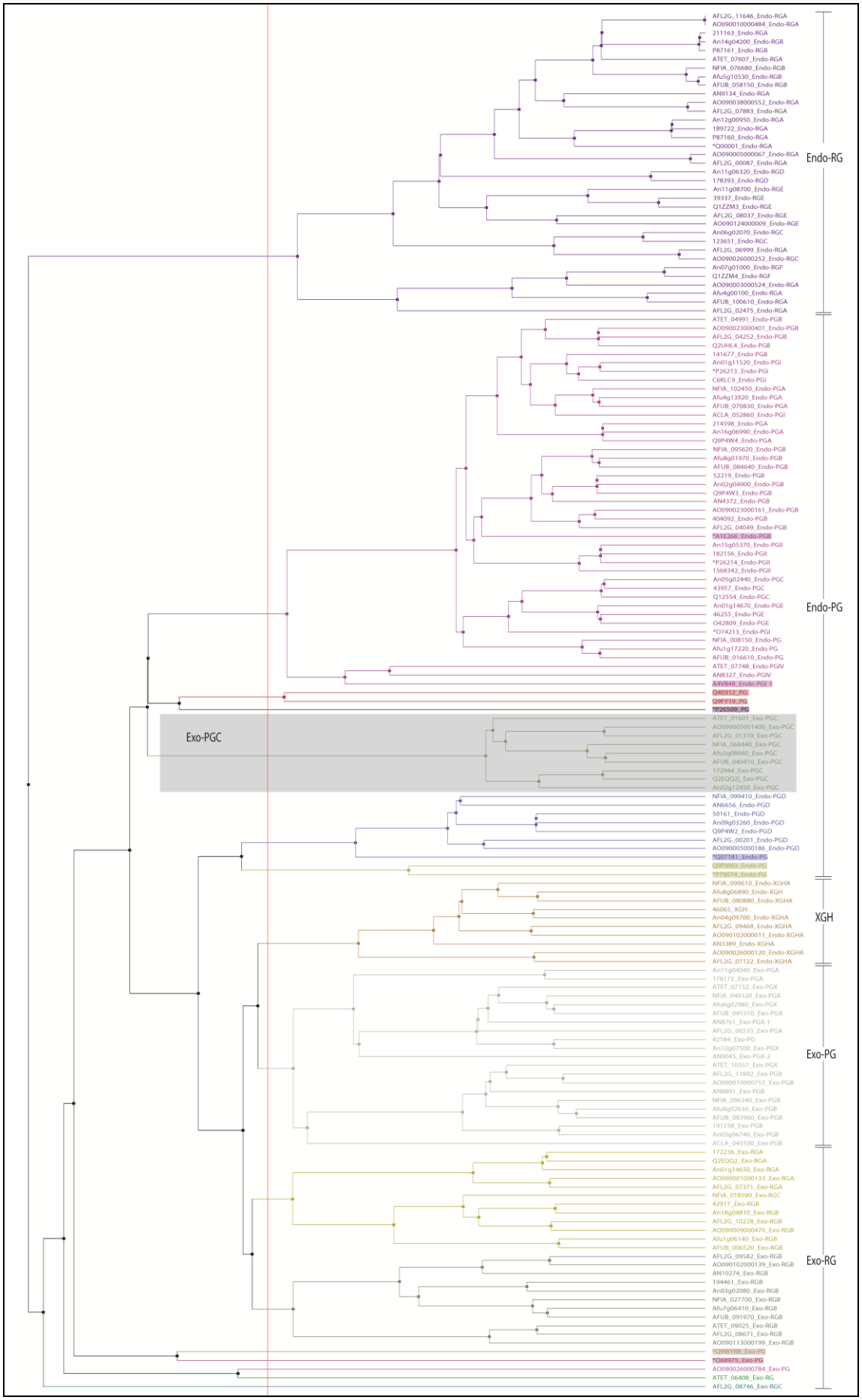
AFL2G_11646_Endo-RGA
AO090010000484_Endo-RGA
211163_Endo-RGA
An14g04200_Endo-RGB
P87161_Endo-RGB
ATET_07607_Endo-RGA
NFIA_076680_Endo-RGB
Afu5g10530_Endo-RGB
AFUB_058150_Endo-RGB
AN9134_Endo-RGA
AO090038000552_Endo-RGA
AFL2G_07883_Endo-RGA
An12g00950_Endo-RGA
189722_Endo-RGA
P87160_Endo-RGA
*Q00001_Endo-RGA
AO090005000067_Endo-RGA
AFL2G_00087_Endo-RGA
An11g06320_Endo-RGD
178393_Endo-RGD
An11g08700_Endo-RGE
39337_Endo-RGE
Q1ZZM3_Endo-RGE
AFL2G_08037_Endo-RGE
AO090124000009_Endo-RGE
An06g02070_Endo-RGC
123651_Endo-RGC
AFL2G_06999_Endo-RGA
AO090026000252_Endo-RGC
An07g01000_Endo-RGF
Q1ZZM4_Endo-RGF
AO090003000524_Endo-RGA
Afu4g00100_Endo-RGA
AFUB_100610_Endo-RGA
AFL2G_02475_Endo-RGA
ATET_04991_Endo-PGB
AO090023000401_Endo-PGB
AFL2G_04252_Endo-PGB
Q2UHL4_Endo-PGB
141677_Endo-PGB
An01g11520_Endo-PGI
*P26213_Endo-PGI
C6KLC9_Endo-PGI
NFIA_102450_Endo-PGA
Afu4g13920_Endo-PGA
AFUB_070830_Endo-PGA
ACLA_052860_Endo-PGI
214598_Endo-PGA
An16g06990_Endo-PGA
Q9P4W4_Endo-PGA
NFIA_095620_Endo-PGB
Afu8g01970_Endo-PGB
AFUB_084640_Endo-PGB
52219_Endo-PGB
An02g04900_Endo-PGB
Q9P4W3_Endo-PGB
AN4372_Endo-PGB
AO090023000161_Endo-PGB
404092_Endo-PGB
AFL2G_04049_Endo-PGB
*A1E266_Endo-PGB
An15g05370_Endo-PGII
182156_Endo-PGII
*P26214_Endo-PGII
1568342_Endo-PGII
An05g02440_Endo-PGC
43957_Endo-PGC
Q12554_Endo-PGC
An01g14670_Endo-PGE
46255_Endo-PGE
O42809_Endo-PGE
*O74213_Endo-PGI
NFIA_008150_Endo-PG
Afu1g17220_Endo-PG
AFUB_016610_Endo-PG
ATET_07748_Endo-PGIV
AN8327_Endo-PGIV
A4VB48_Endo-PGI 1
Q40312_PG
Q9P719_PG
*P26509_PG
Exo-PGC
ATET_01601_Exo-PGC
AO090005001400_Exo-PGC
AFL2G_01310_Exo-PGC
NFIA_068440_Exo-PGC
Afu3g08680_Exo-PGC
AFUB_040410_Exo-PGC
172944_Exo-PGC
Q2EQQ2|_Exo-PGC
An02g12450_Exo-PGC
NFIA_099410_Endo-PGD
AN6656_Endo-PGD
50161_Endo-PGD
An09g03260_Endo-PGD
Q9P4W2_Endo-PGD
AFL2G_00201_Endo-PGD
AO090005000186_Endo-PGD
*Q07181_Endo-PG
Q9P8M3_Endo-PG
*P79074_Endo-PG
NFIA_099610_Endo-XGHA
Afu8g06890_Endo-XGH
AFUB_080880_Endo-XGHA
46065_XGH
An04g09700_Endo-XGHA
AFL2G_09468_Endo-XGHA
AO090102000011_Endo-XGHA
AN3389_Endo-XGHA
AO090026000120_Endo-XGHA
AFL2G_07122_Endo-XGHA
An11g04040_Exo-PGA
178172_Exo-PGA
ATET_07152_Exo-PGX
NFIA_049320_Exo-PGX
Afu6g02980_Exo-PGX
AFUB_095310_Exo-PGX
AN8761_Exo-PGX-1
AFL2G_06533_Exo-PGA
42184_Exo-PG
An12g07500_Exo-PGX
AN9045_Exo-PGX-2
ATET_10357_Exo-PGX
AFL2G_11892_Exo-PGB
AO090010000753_Exo-PGB
AN8891_Exo-PGB
NFIA_096340_Exo-PGB
Afu8g02630_Exo-PGB
AFUB_083960_Exo-PGB
191158_Exo-PGB
An03g06740_Exo-PGB
ACLA_043100_Exo-PGB
172236_Exo-RGA
Q2EQQ2_Exo-RGA
An01g14650_Exo-RGA
AO090001000133_Exo-RGA
AFL2G_07371_Exo-RGA
NFIA_018590_Exo-RGC
42917_Exo-RGB
An18g04810_Exo-RGB
AFL2G_10228_Exo-RGB
AO090009000470_Exo-RGB
Afu1g06140_Exo-RGB
AFUB_006520_Exo-RGB
AFL2G_09582_Exo-RGB
AO090102000139_Exo-RGB
AN10274_Exo-RGB
194461_Exo-RGB
An03g02080_Exo-RGB
NFIA_027700_Exo-RGB
Afu7g06410_Exo-RGB
AFUB_091970_Exo-RGB
ATET_09025_Exo-RGB
AFL2G_08671_Exo-RGB
AO090113000199_Exo-RGB
*Q09761_Exo-PG
*O68975_Exo-PG
AO090026000784_Exo-PG
ATET_06408_Exo-RG
AFL2G_08746_Exo-RGC
Endo-RG
Endo-PG
XGH
Exo-PG
Exo-RG

Figure 9. Phylogeny tree after manual curation. "*" indicates the protein has PDB 3D structure. The shaded proteins are outside of genus *Aspergillus*. Exo-PGC is shaded in grey which is not included in the endo-PG group. On the right indicates the different group named by its enzyme activity.

The overall phylogeny tree after manual curation is very similar to the tree after MAFFT alignment, which means aligning active site together didn't deteriorate the phylogenetic relationship of proteins. (See Figure 9.) It is clear that proteins with same functionality could be clustered in the same clade. Endo-PGC is also isolated as a distinct group rather than exo-PG group. Endo-PG and endo-RG were separated into different clades. XGH is also in the same clade with exo-PG correspond to their common catalytic activity [12]. The major different is endo-PGD is grouped with exo-RG and exo-PG clade instead of on the clade of the rest of endo-PG. This observation indicated that endo-PGD is distant from the other endo-PG enzymes which is also the case in the Figure 6A.

Some GH28 enzymes are still not grouped in the correct site after the manual curation such as the Q9WYR8 and O68975 which are not within the genus *Aspergillus*. These sequences from other organisms may have less similarity compare with other enzyme from *Aspergillus*. This is compatible to what described by *Markovic O., et al, 2001* 12 years ago shows that different organism has different sequence preferences within certain GH28 group (such as endo-PG).

In order to know whether the active site or the rest of the sequence contribute to the enzyme activity we also generate phylogeny tree of sequence alignment of only active site and sequence alignment without the active site. The range of active site includes active site Asp362 -6 till active site His386 (annotation from *Elena S. M. et al, 2006*) +42. All the phylogeny trees are generated by average distance using BLOSUM62 in Jalview. Figure 10 shows the phylogeny tree of only active site sequence alignment.

In general, the phylogeny tree of active site after manual curation is quite similar to the phylogeny tree after manual curation. Most of the correlated GH28 enzymes clustered into different groups. Endo-RG is separated into one clade on the top which is the sequence has large insertion between the active sties and endo-RG is the most distant from the rest of GH28 family protein. Endo-PG is also clustered together but the endo-PGD still on the different clade from the endo-PG. However, XGH AO090026000120 and AFL2G_07122 were grouped into clade exo-PG. Since XGH and exo-PG share common enzyme activity, they may also have similar active site sequences. Exo-PGC is also separated into a distinct group rather than within exo-PG shows that the active site of exo-PGC is different with the rest of the exo-PG. This result implied that the difference within the active site is enough to divide various groups in GH28.

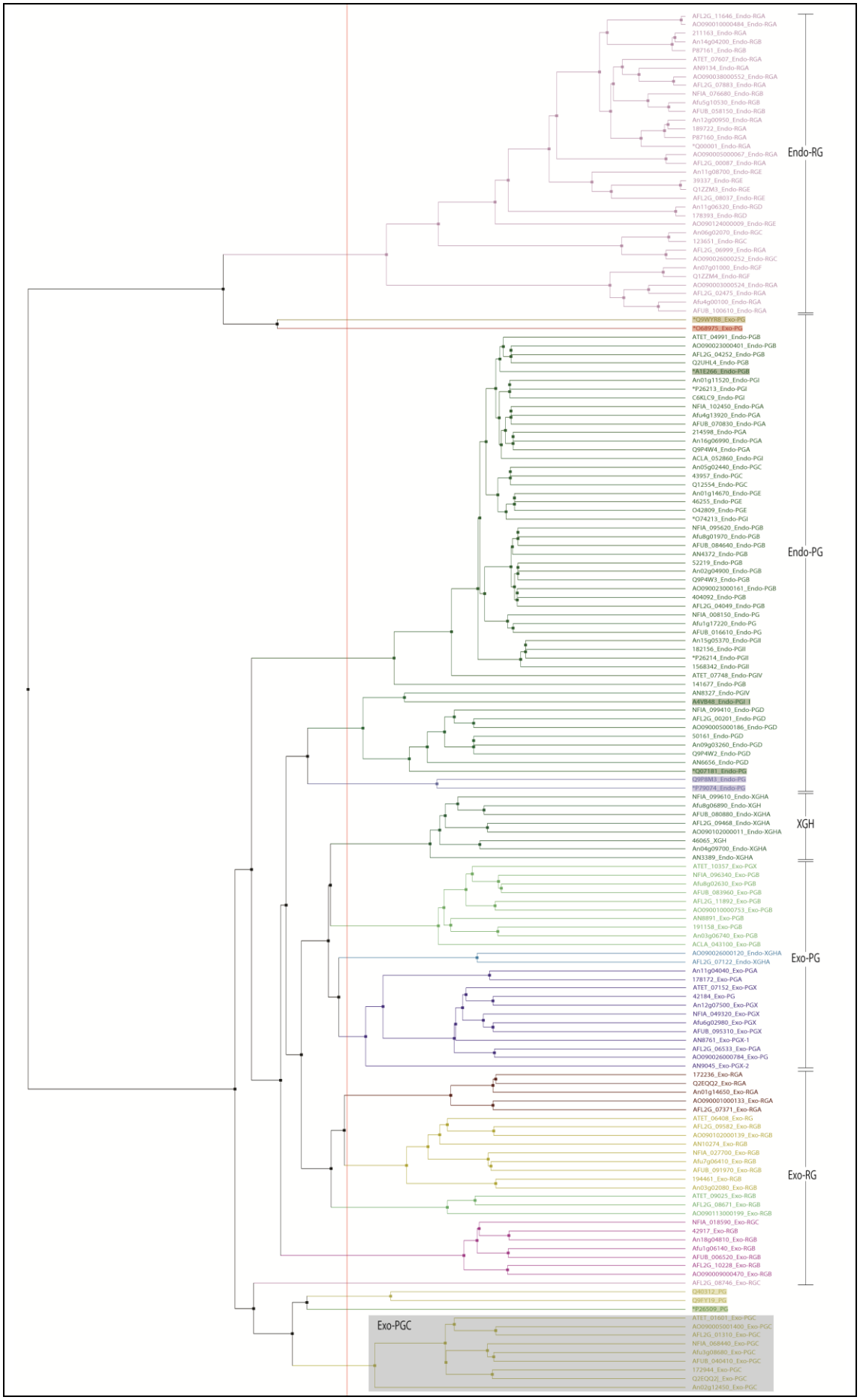
AFL2G_11646_Endo-RGA
AO090010000484_Endo-RGA
211163_Endo-RGA
An14g04200_Endo-RGB
P87161_Endo-RGB
ATET_07607_Endo-RGA
AN9134_Endo-RGA
AO090038000552_Endo-RGA
AFL2G_07883_Endo-RGA
NFIA_076680_Endo-RGB
Afu5g10530_Endo-RGB
AFUB_058150_Endo-RGB
An12g00950_Endo-RGA
189722_Endo-RGA
P87160_Endo-RGA
*Q00001_Endo-RGA
AO090005000067_Endo-RGA
AFL2G_00087_Endo-RGA
An11g08700_Endo-RGE
39337_Endo-RGE
Q1ZZM3_Endo-RGE
AFL2G_08037_Endo-RGE
An11g06320_Endo-RGD
178391_Endo-RGD
AO090124000009_Endo-RGE
An06g02070_Endo-RGC
123651_Endo-RGC
AFL2G_06999_Endo-RGA
AO090026000252_Endo-RGC
An07g01000_Endo-RGF
Q1ZZM4_Endo-RGF
AO090003000524_Endo-RGA
AFL2G_02475_Endo-RGA
Afu4g00100_Endo-RGA
AFUB_100610_Endo-RGA
*Q96WY8_Exo-PG
*O60975_Exo-PG
ATET_04991_Endo-PGB
AO090023000401_Endo-PGB
AFL2G_04252_Endo-PGB
Q2UHL4_Endo-PGB
*A1L266_Endo-PGB
An01g11520_Endo-PGI
*P26213_Endo-PGI
C6KLC9_Endo-PGI
NFIA_102450_Endo-PGA
Afu4g13920_Endo-PGA
AFUB_070830_Endo-PGA
214598_Endo-PGA
An16g06990_Endo-PGA
Q9P4W4_Endo-PGA
ACLA_052860_Endo-PGI
An05g02440_Endo-PGC
43957_Endo-PGC
Q12554_Endo-PGC
An01g14670_Endo-PGE
46255_Endo-PGE
O42809_Endo-PGE
*O74213_Endo-PGI
NFIA_095620_Endo-PGB
Afu8g01970_Endo-PGB
AFUB_084640_Endo-PGB
AN4372_Endo-PGB
52219_Endo-PGB
An02g04900_Endo-PGB
Q9P4W3_Endo-PGB
AO090023000161_Endo-PGB
404092_Endo-PGB
AFL2G_04049_Endo-PGB
NFIA_008150_Endo-PG
Afu1g17220_Endo-PG
AFUB_016610_Endo-PG
An15g05370_Endo-PGII
182156_Endo-PGII
*P26214_Endo-PGII
1568342_Endo-PGII
ATET_07748_Endo-PGIV
141677_Endo-PGB
AN8327_Endo-PGIV
A4VB48_Endo-PGII
NFIA_099410_Endo-PGD
AFL2G_00201_Endo-PGD
AO090005000186_Endo-PGD
50161_Endo-PGD
An09g03260_Endo-PGD
Q9P4W2_Endo-PGD
AN6656_Endo-PGD
*Q07181_Endo-PG
Q9P8M3_Endo-PG
*P79074_Endo-PG
NFIA_099610_Endo-XGHA
Afu8g06890_Endo-XGH
AFUB_080880_Endo-XGHA
AFL2G_09468_Endo-XGHA
AO090102000011_Endo-XGHA
46065_XGH
An04g09700_Endo-XGHA
AN3389_Endo-XGHA
ATET_10357_Exo-PGX
NFIA_096340_Exo-PGB
Afu8g02630_Exo-PGB
AFUB_083960_Exo-PGB
AFL2G_11892_Exo-PGB
AO090010000753_Exo-PGB
AN8891_Exo-PGB
191158_Exo-PGB
An03g06740_Exo-PGB
ACLA_043100_Exo-PGB
AO090026000120_Endo-XGHA
AFL2G_07122_Endo-XGHA
An11g04040_Exo-PGA
178172_Exo-PGA
ATET_07152_Exo-PGX
42184_Exo-PG
An12g07500_Exo-PGX
NFIA_049320_Exo-PGX
Afu6g02980_Exo-PGX
AFUB_095310_Exo-PGX
AN8761_Exo-PGX-1
AFL2G_06533_Exo-PGA
AO090026000784_Exo-PG
AN9045_Exo-PGX-2
172236_Exo-RGA
Q2EQQ2_Exo-RGA
An01g14650_Exo-RGA
AO090001000133_Exo-RGA
AFL2G_07371_Exo-RGA
ATET_06408_Exo-RG
AFL2G_09582_Exo-RGB
AO090102000139_Exo-RGB
AN10274_Exo-RGB
NFIA_027700_Exo-RGB
Afu7g06410_Exo-RGB
AFUB_091970_Exo-RGB
194461_Exo-RGB
An03g02080_Exo-RGB
ATET_09025_Exo-RGB
AFL2G_08671_Exo-RGB
AO090113000199_Exo-RGB
NFIA_018590_Exo-RGC
42917_Exo-RGB
An18g04810_Exo-RGB
Afu1g06140_Exo-RGB
AFUB_006520_Exo-RGB
AFL2G_10228_Exo-RGB
AO090009000470_Exo-RGB
AFL2G_08746_Exo-RGC
Q40312_PG
Q9FYI9_PG
*P26509_PG
ATET_01601_Exo-PGC
AO090005001400_Exo-PGC
AFL2G_01310_Exo-PGC
NFIA_068440_Exo-PGC
Afu3g08680_Exo-PGC
AFUB_040410_Exo-PGC
172944_Exo-PGC
Q2EQQ2|_Exo-PGC
An02g12450_Exo-PGC
Endo-RG
Endo-PG
XGH
Exo-PG
Exo-RG
Exo-PGC

Figure 10. Phylogeny tree of sequence alignment contains only active site. "*" indicates the protein has PDB 3D structure. The shaded proteins are outside of genus *Aspergillus*. Exo-PGC clade is shaded in grey. On the right indicates the different group named by its enzyme activity.

Figure 11 shows the phylogeny tree of sequence alignment excluding the active site. The GH28 enzymes still cluster into distinct groups similar to the phylogeny tree of sequence contain only active site. Endo-PG cluster into a clade which is on the top. Exo-PGC also clusters into a distinct group rather than within exo-PG. Endo-PGD is in the same clade as exo-PG and exo-RG. XGH is also in the same clade with exo-PG like the phylogeny tree of sequence only active site. Exo-enzyme exo-PG and exo-RG also cluster in the same clade. Endo-RG is also cluster into a distinct group on the bottom.

High similarity between the phylogeny tree of active site and without active site indicate that aligning the active site together during the manual curation didn't deteriorate the overall sequence alignment and the sequence differences is not just within the active site.

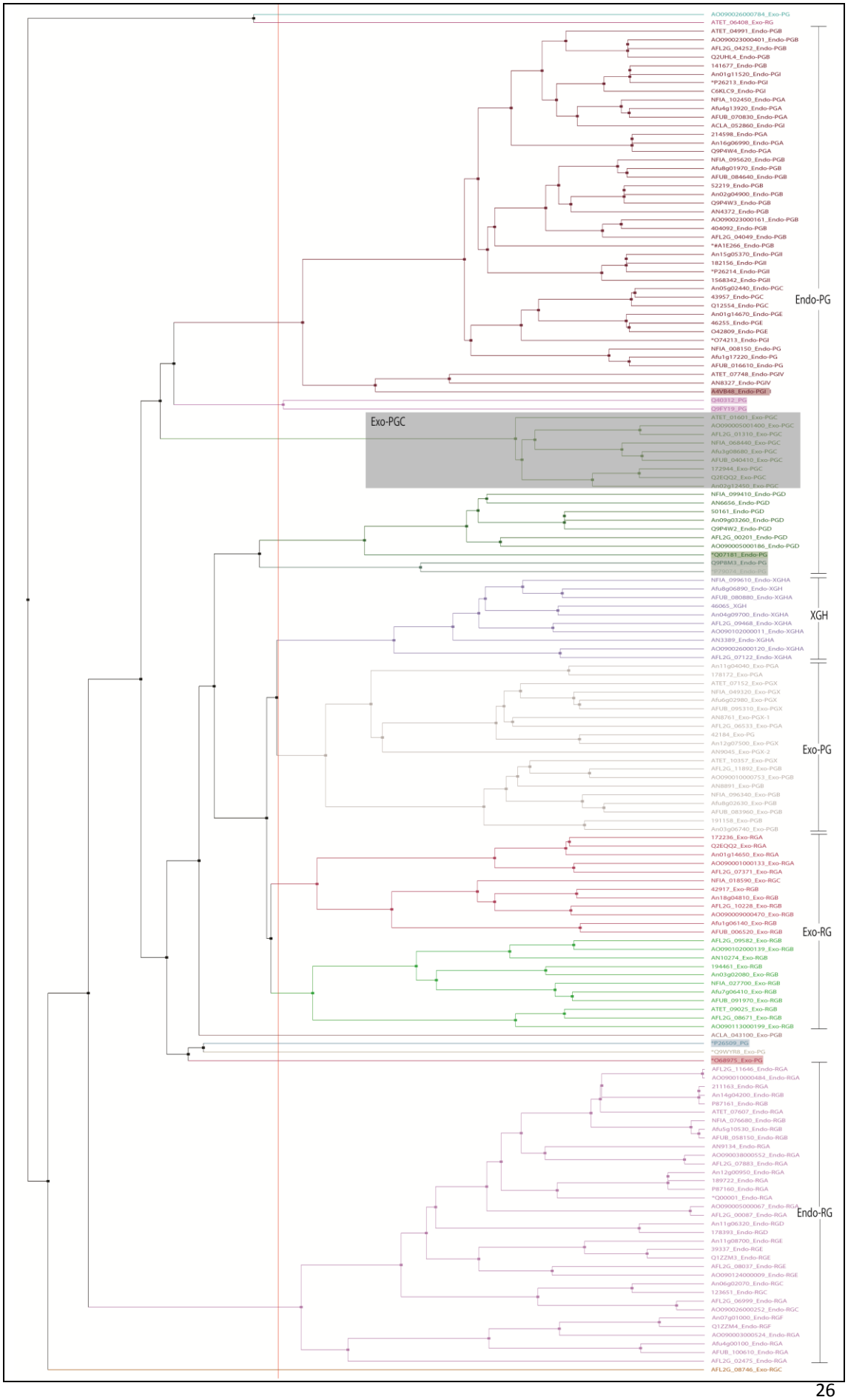
AO090026000784_Exo-PG
ATET_06408_Exo-RG
ATET_04991_Endo-PGB
AO090023000401_Endo-PGB
AFL2G_04252_Endo-PGB
Q2URL4_Endo-PGB
141677_Endo-PGB
An01g11520_Endo-PGI
*P26213_Endo-PGI
C6KLC9_Endo-PGI
NFIA_102450_Endo-PGA
Afu4g13920_Endo-PGA
AFUB_070830_Endo-PGA
ACLA_052860_Endo-PGI
214598_Endo-PGA
An16g06990_Endo-PGA
Q9P4W4_Endo-PGA
NFIA_095620_Endo-PGB
Afu8g01970_Endo-PGB
AFUB_084640_Endo-PGB
52219_Endo-PGB
An02g04900_Endo-PGB
Q9P4W3_Endo-PGB
AN4372_Endo-PGB
AO090023000161_Endo-PGB
404092_Endo-PGB
AFL2G_04049_Endo-PGB
*#A1E266_Endo-PGB
An15g05370_Endo-PGII
182156_Endo-PGII
*P26214_Endo-PGII
1568342_Endo-PGII
An05g02440_Endo-PGC
43957_Endo-PGC
Q12554_Endo-PGC
An01g14670_Endo-PGE
46255_Endo-PGE
O42809_Endo-PGE
*O74213_Endo-PGI
NFIA_008150_Endo-PG
Afu1g17220_Endo-PG
AFUB_016610_Endo-PG
ATET_07748_Endo-PGIV
AN8327_Endo-PGIV
A4VB48_Endo-PGI
Q40312_PG
Q9FY19_PG
Exo-PGC
ATET_01601_Exo-PGC
AO090005001400_Exo-PGC
AFL2G_01310_Exo-PGC
NFIA_068440_Exo-PGC
Afu3g08080_Exo-PGC
AFUB_040410_Exo-PGC
172944_Exo-PGC
Q2EQQ2_Exo-PGC
An02g12450_Exo-PGC
NFIA_099410_Endo-PGD
AN6656_Endo-PGD
50161_Endo-PGD
An09g03260_Endo-PGD
Q9P4W2_Endo-PGD
AFL2G_00201_Endo-PGD
AO090005000186_Endo-PGD
*Q07181_Endo-PG
Q9P8M3_Endo-PG
*P79074_Endo-PG
Endo-PG
NFIA_099610_Endo-XGHA
Afu8g06890_Endo-XGH
AFUB_080880_Endo-XGHA
46065_XGH
An04g09700_Endo-XGHA
AFL2G_09468_Endo-XGHA
AO090102000011_Endo-XGHA
AN3389_Endo-XGHA
AO090026000120_Endo-XGHA
AFL2G_07122_Endo-XGHA
XGH
An11g04040_Exo-PGA
178172_Exo-PGA
ATET_07152_Exo-PGX
NFIA_049320_Exo-PGX
Afu6g02980_Exo-PGX
AFUB_095310_Exo-PGX
AN8761_Exo-PGX-1
AFL2G_06533_Exo-PGA
42184_Exo-PG
An12g07500_Exo-PGX
AN9045_Exo-PGX-2
ATET_10157_Exo-PGX
AFL2G_11892_Exo-PGB
AO090010000753_Exo-PGB
AN8891_Exo-PGB
NFIA_096340_Exo-PGB
Afu8g02630_Exo-PGB
AFUB_083960_Exo-PGB
191158_Exo-PGB
An03g06740_Exo-PGB
Exo-PG
172236_Exo-RGA
Q2EQQ2_Exo-RGA
An01g14650_Exo-RGA
AO090010000133_Exo-RGA
AFL2G_07371_Exo-RGA
NFIA_018590_Exo-RGC
42917_Exo-RGB
An18g04810_Exo-RGB
AFL2G_10228_Exo-RGB
AO090009000470_Exo-RGB
Afu1g06140_Exo-RGB
AFUB_006520_Exo-RGB
AFL2G_09582_Exo-RGB
AO090102000139_Exo-RGB
AN10274_Exo-RGB
194461_Exo-RGB
An03g02080_Exo-RGB
NFIA_027700_Exo-RGB
Afu7g06410_Exo-RGB
AFUB_091970_Exo-RGB
ATET_09025_Exo-RGB
AFL2G_08671_Exo-RGB
AO090113000199_Exo-RGB
Exo-RG
ACLA_043100_Exo-PGB
*P26509_PG
*Q9WV98_Exo-PG
*O68975_Exo-PG
AFL2G_11646_Endo-RGA
AO090010000484_Endo-RGA
211163_Endo-RGA
An14g04200_Endo-RGB
P87161_Endo-RGB
ATET_07607_Endo-RGA
NFIA_076680_Endo-RGB
Afu5g10530_Endo-RGB
AFUB_058150_Endo-RGB
AN9134_Endo-RGA
AO090038000552_Endo-RGA
AFL2G_07883_Endo-RGA
An12g00950_Endo-RGA
189722_Endo-RGA
P87160_Endo-RGA
*Q00001_Endo-RGA
AO090005000867_Endo-RGA
AFL2G_00087_Endo-RGA
An11g06320_Endo-RGD
178393_Endo-RGD
An11g08700_Endo-RGE
39337_Endo-RGE
Q1ZZM3_Endo-RGE
AFL2G_08037_Endo-RGE
AO090124000009_Endo-RGE
An06g02070_Endo-RGC
123651_Endo-RGC
AFL2G_06999_Endo-RGA
AO090026000252_Endo-RGC
An07g01000_Endo-RGF
Q1ZZM4_Endo-RGF
AO090003000524_Endo-RGA
Afu4g00100_Endo-RGA
AFUB_100610_Endo-RGA
AFL2G_02475_Endo-RGA
Endo-RG
AFL2G_08746_Exo-RGC

Figure 11. Phylogeny tree of sequence alignment without the active site. "*" indicates the protein has PDB 3D structure. The shaded proteins are outside of genus *Aspergillus*. On the right indicates the different group named by its enzyme activity.

■ **Sequence conservation in active site of GH28**

Here by we compared the enzyme sequence near the active site through analyze the multiple sequence alignment from Jalview by WebLogo [64]. WebLogo is a sequence logo generator that can help us to visualize the sequence conservation. More specifically, WebLogo analyze the sequence alignment and output a sequence of logo, and each logo contains a stack of letters with different height indicates the sequence conservation in that position [64].

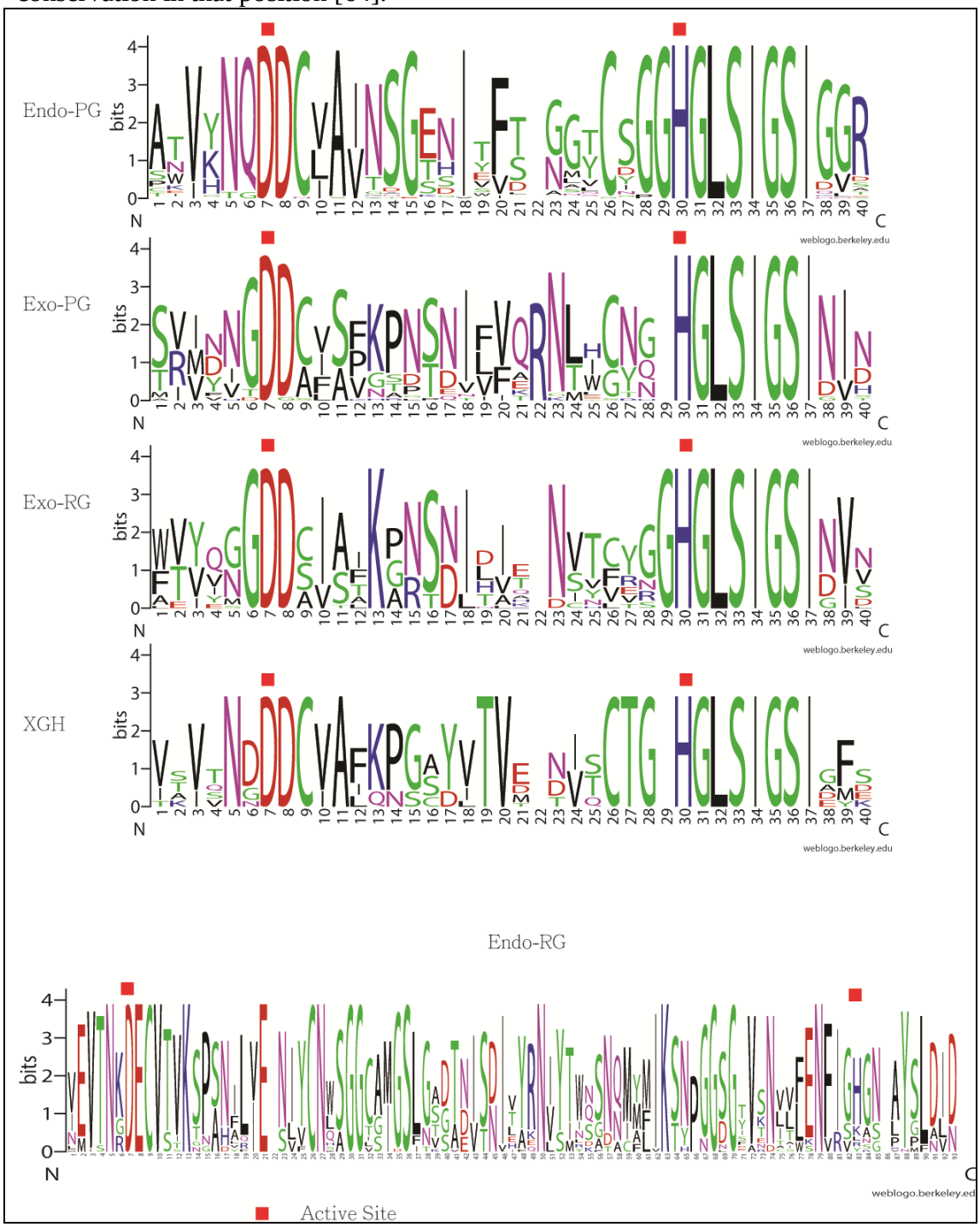

Figure 12. GH28 enzymes active site conservation. The graph was generated by WebLogo [64] to help us to visualize the sequence conservation of the GH28 enzyme. The overall height of each stack indicates the sequence conservation at that position. Red box indicate the position of active site. Abbreviations: PG, endo-ploygalacturonase; exo-PG, exo-polygalacturonase; endo-RG, endo-rhamnogalacturonase; exo-RG, exo-rhamnogalacturonase; XGH, xylogalacturonan hydrolase.

In Figure 12, it shows that GH28 enzymes have conserved active site Aspartate and Histidine. The amino acid sequences between active sites Aspartate and Histidine is connected by β-sheets and have sequence differences within the GH28 protein family but the sequence near the active site Aspartate and Histidine are very similar to each other. Active site Aspartate has one Aspartate/Glutamate on the next amino acid. Active site Histidine is surrounded by Glycine or some non-charged amino acids. Near the active site Aspartate has more negatively charged amino acid side chains like Aspartate and Glutamate. On the other hand, near active site Histidine are mostly non-polar side-chains amino acids mean it has different roles with Aspartate on the catalysis. There are conserved Asparagines in the exo-PG and exo-RG instead of endo-enzymes, which may be crucial in the substrate specificity.

Due to the fact that exo-PGC and endo-PGD were separated from the original group in the phylogeny tree, they are also separately analyzed by the WebLogo to see whether their sequences are similar to original group exo-PG and endo-PG. (Figure 9)

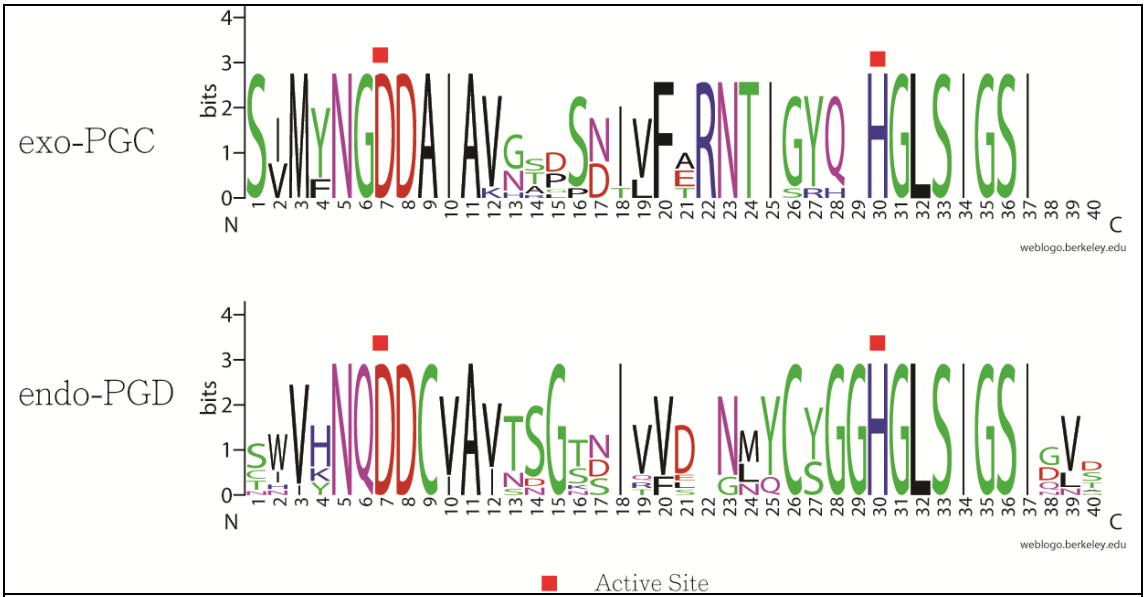


Figure 13. WebLogo of exo-PGC and endo-PGD. The graph was generated by WebLogo [64] to help us to visualize the sequence conservation of the GH28 enzyme. The overall height of each stack indicates the sequence conservation at that position. Red box indicate the position of active site.

From the Figure 13 we can see that the overall sequence conservation of exo-PGC is very similar to the exo-PG (in Figure 9) but exo-PGC is lack of negatively charged amino acid residue after active site Histidine at the position 38 and 40 in the WebLogo. Following the active site Histidine are amino acids with hydrophobic side chain. In the endo-PGD also has very similar overall sequence conservation with endo-PG but actually it is more similar to the exo-RG and exo-PG rather than the

endo-PG.

The hydropathy of the active site of GH28 proteins were calculated by GRAVY (grand average of hydropathy) from Sequence Manipulation Suite online (http://www.bioinformatics.org/sms2/protein_gravy.html). The range of active site includes active site Asp362 -6 till active site His386 (annotation from *Elena S. M. et al, 2006*) +42. The GRAVY value is calculated by adding the hydropathy of each amino acid residue and dividing by the length of the sequence [65]. Then each group's active site hydropathy average was calculated in order to understand the physiochemical differences between the groups. We found that there are differences of hydropathy between groups and also within group such as exo-PGC and endo-PGD which has less sequence similarity in the original group. However, the value of standard deviation is much more than 15% of average which means it is not consistent of hydropathy on active regions within the group.

- **Hidden Markov Model for active site of GH28**

Hidden Markov models (HMMs) are a formal foundation for making probabilistic models of linear sequence 'labeling' problems [66]. HMMs can help us to recognize specific pattern such as speech, handwriting, etc. In our case HMMs was used to "learn" the protein sequences similarity. We used the current different groups of sequences to generate all the HMMs in the GH28. The program HMMER [67] (http://hmmer.janelia.org/software) was used under Linus operation system to generate HMMs [67]. Those generated HMMs can help us to classify those putative GH28 family sequences, which its enzymatic activity has not been characterized. The HMMs of GH28 are in the supplementary data. (HMM models are available as supplementary files)

- **Structure comparison of Active site**

The protein structures of GH28 were compared to know whether there are any differences between the structures within the GH28 family protein that may contribute to the substrate specificity. The crystal structures of proteins are from the Protein Data Base (PDB) and the one without structure were generated by SWISSMODEL [68] with the alignment of protein sequences of closest known crystal structure. The structure of exo-RG and XGH were generated by SWISSMODEL with the sequence of *A. niger* exo-RG (An01g14650) and XGH (An04g09700) with known structure 1CZF and 1NHC to generate the homology model.

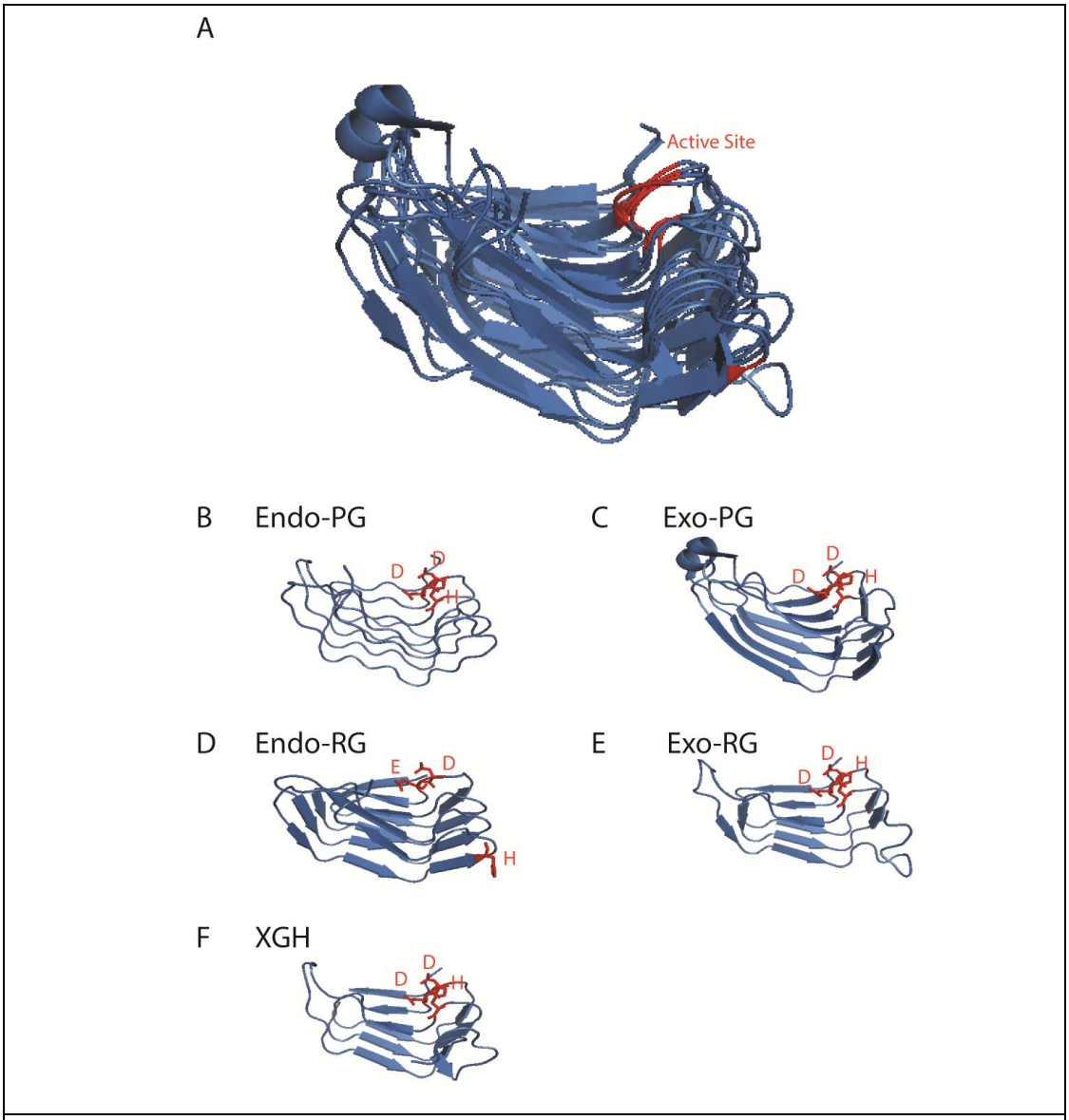


Figure 13. Structure comparison of GH28. A. Different group of GH28 structure near active site were superimposed together by program Pymol and the active sites are marked in red. Only the structures near the active site were shown. B. Structure of endo-PG 1CZF (PDB) from *A. niger* (N400). C. Structure of exo-PG 3JUR from *Thermotoga maritime*. D. Structure of endo-RG 1RMG from *A. aculeatus*. E. Structure of exo-RG modeled by SWISSMODEL from the sequence An01g14650 aligned with known structure 1CZF. F. Structure of XGH modeled by SWISSMODEL from sequence An04g09700 aligned with known structure 1NHC.

In Figure 13 shows the structures of the different sub-group in GH28 family near the active site. Figure 12A superimpose all the structures together indicate that the structures near the active site are very similar and the locations of active sites are highly conserved except the endo-RG which has Histidine located at the opposite of the Aspartate active site which may involve in the substrate specificity of the enzyme. While we separate all the protein shows that there are still some strucutral differences between groups. This is correlated with study done by *Markovic O., et al, 2001* 12 years ago claims that different enzyme group in GH28 has different sequence specificities.

## ➢ Discussion and Conclusion

Among the different groups in GH28 there are some highly conserved amino acid residues in the active site which is considered to contribute to the catalytic activity of the enzyme. In the sequence manual curation process, we specifically aligned two active sites Aspartate and Histidine together throughout the sequences instead of aligning all the residues that are involved in the catalytic activity. However, it is very hard to really align all the residues involved in the enzyme activity because the active site region of the protein is quite diverse within the GH28 family. Previously it was suggested that there is no Histidine at the active site of the endo-RG [28] , we found out there is actually a Histidine residue with similar neighbor on the position after the active site and it turn out to be on the opposite site of the active site and is identified as a active site due to its sequence similarity with an enriched GH28 enzyme dataset. This kind of sequence may be evolved from insertion of transposon in DNA sequence between the active site Aspartate and Histidine form other GH28 enzyme. Moreover, this kind of longer sequence between the Aspartate and Histidine is conserved within the endo-RG only. This may mean that it contributes to its substrate specificity. Thus it is possible to verify whether the unique properties of endo-RG active site play an important role in substrate specificities by replacing another GH28 family enzyme active site with endo-RG active site.

Majority of endo-RG have highly conserved Histidine after the active site, however, in 189722_endorhamnogalacturonase_A and P87160_endorhamnogalacturonase_A, the active site Histidine was replaced with Arginine which is also positively charged may be able to carry out same function in endo-RGA. This variable characteristic implies that Histidine in endo-RG may play a different role compare to the rest of the GH28 family enzymes Histidine at the active site.

While the active site alignment and without the active site alignment are compared. They are still able to distinguish between groups with their sequences. This means the manual curation didn't disrupt the overall alignment and the sequence outside the active site range we selected may also contribute to the substrate specificity, such as the loop that block the active site of *Y. enterocolitica* exo-PG.

In the WebLogo from the active site sequence alignment we found that different enzyme has its sequence composition preference. Those differences of amino acids residue conservation could affect the chemical environment like hydropathy and charge of the active site and therefore affects the substrate specificity. The hydropathy of different GH28 enzymes active site are calculated but they diverse within the groups, indicating it may not be significantly contribute to the substrate specificity. The substrate may be contributed by the structure of the active site. Furthermore, the active site sequences were summarized into HMMs to help identify those unknown protein sequences, which will be very useful when new fungus genomes become available. We didn't go into the subgroups of GH28 due to time limitation, however similar method can as well be applied to the subgroups and bring sub-classification in higher resolution. . The future perspective of this research might require experimental validation to support GH28 subgroups classification.

Currently there are crystal structure of endo-PG, exo-PG, and endo-RG available for us to compare the working mechanism of different GH28 family enzymes. In PG, whether endo or exo-enzyme is decided by the loop insertion in the active site. With loop insertion blocking the accessibility of the substrate so the enzyme works as exo-PG. The crystal structure of exo-RG and XGH were not available yet but we generated homology model from SWISSMODEL in order to compare the structural

differences. The cleft structures of generated model are quite similar to the rest of structures but there are still some differences which may affect the substrate specificity, e.g. extended loop near the active site. The extended loops near the active site are rather diverse within the GH28 and we are not sure whether the homology model can really reflect the real scenario. They may contribute to the substrate specificity, thus more crystal structures are needed to verify whether those loops are relevant to the substrate specificity.

## ➢ References


1. Lee, S.J. and D.W. Lee, *Design and development of synthetic microbial platform cells for bioenergy.* Front Microbiol, 2013. **4**(92).
2. Mohnen, D., *Pectin structure and biosynthesis.* Curr Opin Plant Biol, 2008. **11**(3): p. 266-77.
3. Aspeborg, H., et al., *Carbohydrate-active enzymes involved in the secondary cell wall biogenesis in hybrid aspen.* Plant Physiol, 2005. **137**(3): p. 983-97.
4. Jordan, D.B., et al., *Plant cell walls to ethanol.* Biochem J, 2012. **442**(2): p. 241-52.
5. Xiao, C. and C.T. Anderson, *Roles of pectin in biomass yield and processing for biofuels.* Front Plant Sci, 2013. **4**(67): p. 27.
6. Cosgrove, D.J., *Loosening of plant cell walls by expansins.* Nature, 2000. **407**(6802): p. 321-6.
7. Marcus, S.E., et al., *Pectic homogalacturonan masks abundant sets of xyloglucan epitopes in plant cell walls.* BMC Plant Biol, 2008. **8**(60): p. 1471-2229.
8. Pakarinen, A., et al., *Enzymatic accessibility of fiber hemp is enhanced by enzymatic or chemical removal of pectin.* Bioresour Technol, 2012. **107**: p. 275-81.
9. de Vries, R.P. and J. Visser, *Aspergillus enzymes involved in degradation of plant cell wall polysaccharides.* Microbiol Mol Biol Rev, 2001. **65**(4): p. 497-522.
10. Sprockett, D.D., H. Piontkivska, and C.B. Blackwood, *Evolutionary analysis of glycosyl hydrolase family 28 (GH28) suggests lineage-specific expansions in necrotrophic fungal pathogens.* Gene, 2011. **479**(1-2): p. 29-36.
11. Suykerbuyk, M.E., et al., *Cloning, sequence and expression of the gene coding for rhamnogalacturonase of Aspergillus aculeatus; a novel pectinolytic enzyme.* Appl Microbiol Biotechnol, 1995. **43**(5): p. 861-70.
12. Zandleven, J., et al., *Mode of action of xylogalacturonan hydrolase towards xylogalacturonan and xylogalacturonan oligosaccharides.* Biochem J, 2005. **387**(Pt 3): p. 719-25.
13. Markovic, O. and S. Janecek, *Pectin degrading glycoside hydrolases of family 28: sequence-structural features, specificities and evolution.* Protein Eng, 2001. **14**(9): p. 615-31.
14. Abbott, D.W. and A.B. Boraston, *The structural basis for exopolygalacturonase activity in a family 28 glycoside hydrolase.* J Mol Biol, 2007. **368**(5): p. 1215-22.
15. van Santen, Y., et al., *1.68-A crystal structure of endopolygalacturonase II from Aspergillus niger and identification of active site residues by site-directed mutagenesis.* J Biol Chem, 1999. **274**(43): p. 30474-80.

16. Armand, S., et al., *The active site topology of Aspergillus niger endopolygalacturonase II as studied by site-directed mutagenesis.* J Biol Chem, 2000. **275**(1): p. 691-6.
17. Shimizu, T., et al., *Active-site architecture of endopolygalacturonase I from Stereum purpureum revealed by crystal structures in native and ligand-bound forms at atomic resolution.* Biochemistry, 2002. **41**(21): p. 6651-9.
18. Pages, S., et al., *Subsite mapping of Aspergillus niger endopolygalacturonase II by site-directed mutagenesis.* J Biol Chem, 2000. **275**(38): p. 29348-53.
19. Pages, S., et al., *Changing a single amino acid residue switches processive and non-processive behavior of Aspergillus niger endopolygalacturonase I and II.* J Biol Chem, 2001. **276**(36): p. 33652-6.
20. van Pouderoyen, G., et al., *Structural insights into the processivity of endopolygalacturonase I from Aspergillus niger.* FEBS Lett, 2003. **554**(3): p. 462-6.
21. Hasui, Y., et al., *Isolation, characterization, and sugar chain structure of endoPG Ia, Ib and Ic from Stereum purpureum.* Biosci Biotechnol Biochem, 1998. **62**(5): p. 852-7.
22. Benen, J.A., H.C. Kester, and J. Visser, *Kinetic characterization of Aspergillus niger N400 endopolygalacturonases I, II and C.* Eur J Biochem, 1999. **259**(3): p. 577-85.
23. Bonnin, E., et al., *Study of the mode of action of endopolygalacturonase from Fusarium moniliforme.* Biochim Biophys Acta, 2001. **15**(3): p. 301-9.
24. Kofod, L.V., et al., *Cloning and characterization of two structurally and functionally divergent rhamnogalacturonases from Aspergillus aculeatus.* J Biol Chem, 1994. **269**(46): p. 29182-9.
25. Cho, S.W., S. Lee, and W. Shin, *The X-ray structure of Aspergillus aculeatus polygalacturonase and a modeled structure of the polygalacturonase-octagalacturonate complex.* J Mol Biol, 2001. **311**(4): p. 863-78.
26. Pitson, S.M., et al., *Stereochemical course of hydrolysis catalysed by alpha-L-rhamnosyl and alpha-D-galacturonosyl hydrolases from Aspergillus aculeatus.* Biochem Biophys Res Commun, 1998. **242**(3): p. 552-9.
27. Petersen, T.N., S. Kauppinen, and S. Larsen, *The crystal structure of rhamnogalacturonase A from Aspergillus aculeatus: a right-handed parallel beta helix.* Structure, 1997. **5**(4): p. 533-44.
28. Pickersgill, R., et al., *Crystal structure of polygalacturonase from Erwinia carotovora ssp. carotovora.* J Biol Chem, 1998. **273**(38): p. 24660-4.
29. Sakamoto, T., et al., *Purification and characterisation of two exo-polygalacturonases from Aspergillus niger able to degrade xylogalacturonan and acetylated homogalacturonan.* Biochim Biophys Acta, 2002. **15**(1): p. 10-8.
30. dos Santos Cunha Chellegatti, M.A., M.J. Fonseca, and S. Said, *Purification and partial characterization of exopolygalacturonase I from Penicillium frequentans.* Microbiol Res, 2002. **157**(1): p. 19-24.
31. Kitamoto, N., et al., *Structural features of a polygalacturonase gene cloned from Aspergillus oryzae KBN616.* FEMS Microbiol Lett, 1993. **111**(1): p. 37-41.
32. Martens-Uzunova, E.S., et al., *A new group of exo-acting family 28 glycoside hydrolases of Aspergillus niger that are involved in pectin degradation.* Biochem J, 2006. **400**(1): p. 43-52.

33. Cantarel, B.L., et al., *The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.* Nucleic Acids Res, 2009. **37**(Database issue): p. 5.
34. Bussink, H.J., et al., *Identification and characterization of a second polygalacturonase gene of Aspergillus niger.* Curr Genet, 1991. **20**(4): p. 301-7.
35. Pel, H.J., et al., *Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88.* Nat Biotechnol, 2007. **25**(2): p. 221-31.
36. Parenicova, L., et al., *Characterization of a novel endopolygalacturonase from Aspergillus niger with unique kinetic properties.* FEBS Lett, 2000. **467**(2-3): p. 333-6.
37. Parenicova, L., et al., *pgaA and pgaB encode two constitutively expressed endopolygalacturonases of Aspergillus niger.* Biochem J, 2000. **3**: p. 637-44.
38. Parenicova, L., et al., *pgaE encodes a fourth member of the endopolygalacturonase gene family from Aspergillus niger.* Eur J Biochem, 1998. **251**(1-2): p. 72-80.
39. Bussink, H.J., et al., *The polygalacturonases of Aspergillus niger are encoded by a family of diverged genes.* Eur J Biochem, 1992. **208**(1): p. 83-90.
40. Suykerbuyk, M.E., et al., *Cloning and characterization of two rhamnogalacturonan hydrolase genes from Aspergillus niger.* Appl Environ Microbiol, 1997. **63**(7): p. 2507-15.
41. Azadi, P., et al., *The backbone of the pectic polysaccharide rhamnogalacturonan I is cleaved by an endohydrolase and an endolyase.* Glycobiology, 1995. **5**(8): p. 783-9.
42. Galagan, J.E., et al., *Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae.* Nature, 2005. **438**(7071): p. 1105-15.
43. Wortman, J.R., et al., *The 2008 update of the Aspergillus nidulans genome annotation: a community effort.* Fungal Genet Biol, 2009. **46**(1): p. 25.
44. Bauer, S., et al., *Development and application of a suite of polysaccharide-degrading enzymes for analyzing plant cell walls.* Proc Natl Acad Sci U S A, 2006. **103**(30): p. 11417-22.
45. Kitamoto, N., et al., *Utilization of the TEF1-alpha gene (TEF1) promoter for expression of polygalacturonase genes, pgaA and pgaB, in Aspergillus oryzae.* Appl Microbiol Biotechnol, 1998. **50**(1): p. 85-92.
46. Williams, H.L., Y. Tang, and W.E. Hintz, *Endopolygalacturonase is encoded by a multigene family in the basidiomycete Chondrostereum purpureum.* Fungal Genet Biol, 2002. **36**(1): p. 71-83.
47. Miyairi, K., et al., *Cloning and sequence analysis of cDNA encoding endopolygalacturonase I from Stereum purpureum.* Biosci Biotechnol Biochem, 1997. **61**(4): p. 655-9.
48. Senda, M., et al., *Characterization of an endopolygalacturonase Gene cppg1 from Phytopathogenic Fungus Chondrostereum purpureum.* Journal of General Plant Pathology, 2001. **67**(1): p. 41-44.
49. Rowe, H.C. and D.J. Kliebenstein, *Elevated genetic variation within virulence-associated Botrytis cinerea polygalacturonase loci.* Mol Plant Microbe Interact, 2007. **20**(9): p. 1126-37.
50. Cettul, E., et al., *Evolutionary analysis of endopolygalacturonase-encoding genes of Botrytis cinerea.* Mol Plant Pathol, 2008. **9**(5): p. 675-85.
51. Yokoyama, M., et al., *Purification, identification, and cDNA cloning of Jun a*

*2, the second major allergen of mountain cedar pollen.* Biochem Biophys Res Commun, 2000. **275**(1): p. 195-202.
52. Caprari, C., et al., *Cloning and characterization of a gene encoding the endopolygalacturonase of Fusarium moniliforme.* Mycological Research, 1993. **97**(4): p. 497-505.
53. Federici, L., et al., *Structural requirements of endopolygalacturonase for the interaction with PGIP (polygalacturonase-inhibiting protein).* Proc Natl Acad Sci U S A, 2001. **98**(23): p. 13425-30.
54. Saarilahti, H.T., et al., *Structural analysis of the pehA gene and characterization of its protein product, endopolygalacturonase, of Erwinia carotovora subspecies carotovora.* Mol Microbiol, 1990. **4**(6): p. 1037-44.
55. Koskinen, J.P., et al., *Genome sequence of Pectobacterium sp. strain SCC3193.* J Bacteriol, 2012. **194**(21): p. 00681-12.
56. Nelson, K.E., et al., *Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima.* Nature, 1999. **399**(6734): p. 323-9.
57. Pijning, T., et al., *The crystal structure of a hyperthermoactive exopolygalacturonase from Thermotoga maritima reveals a unique tetramer.* FEBS Lett, 2009. **583**(22): p. 3665-70.
58. Liao, C.H., et al., *Genetic and biochemical characterization of an exopolygalacturonase and a pectate lyase from Yersinia enterocolitica.* Can J Microbiol, 1999. **45**(5): p. 396-403.
59. Bonivento, D., et al., *Crystal structure of the endopolygalacturonase from the phytopathogenic fungus Colletotrichum lupini and its interaction with polygalacturonase-inhibiting proteins.* Proteins, 2008. **70**(1): p. 294-9.
60. Katoh, K. and D.M. Standley, *MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.* Molecular Biology and Evolution, 2013. **30**(4): p. 772-780.
61. Castresana, J., *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Mol Biol Evol, 2000. **17**(4): p. 540-52.
62. Waterhouse, A.M., et al., *Jalview Version 2--a multiple sequence alignment editor and analysis workbench.* Bioinformatics, 2009. **25**(9): p. 1189-91.
63. Dereeper, A., et al., *Phylogeny.fr: robust phylogenetic analysis for the non-specialist.* Nucleic Acids Res, 2008. **36**(Web Server issue): p. 19.
64. Crooks, G.E., et al., *WebLogo: a sequence logo generator.* Genome Res, 2004. **14**(6): p. 1188-90.
65. Kyte, J. and R.F. Doolittle, *A simple method for displaying the hydropathic character of a protein.* J Mol Biol, 1982. **157**(1): p. 105-32.
66. Eddy, S.R., *What is a hidden Markov model?* Nat Biotechnol, 2004. **22**(10): p. 1315-6.
67. Eddy, S.R., *Accelerated Profile HMM Searches.* PLoS Comput Biol, 2011. **7**(10): p. 20.
68. Arnold, K., et al., *The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling.* Bioinformatics, 2006. **22**(2): p. 195-201.

- **Supplementary Data**

- **Sequence collection**

  Supplementary data 1. Aspergillus_sequences.FASTA (including data from CAZy and its orthologs from COG)
  Supplementary data 2. Aspergilla_COG.xls (COG file for *Aspergilli*)

- **Sequence analysis**

  Supplementary data 3. After MAFFT (MAFFT_Af.FASTA)
  After manual curation
  Supplementary data 4. Full sequence (MAFFT_MC.FASTA)
  Supplementary data 5. Only active site (MAFFT_MCActiveSite.FASTA)
  Supplementary data 6. Excluding active site (MAFFT_MCNoActiveSite.FASTA)

- **Hidden Markov Models (HMMdata)**

  Supplementary data 7. Endo-PG.hmm
  Supplementary data 8. Exo-PG.hmm
  Supplementary data 9. Endo-RG.hmm
  Supplementary data 10. Exo-RG.hmm
  Supplementary data 11. XGH.hmm