Comprehensive Bioinformatic Analysis of Glycoside Hydrolase Family 10 Proteins

Sherry Wu

A Thesis in

The Department of

Biology

Presented in Partial Fulfillment of the Requirements for

the Degree of Master of Science (Biology) at Concordia

University

Montreal, Quebec, Canada

January 2015

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared:

By: Sherry Wu

Entitled: Comprehensive Bioinformatic Analysis of Glycoside Hydrolase Family 10 Proteins

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Biology)

complies with the regulation of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

Dr. Vincent Martin _____ Chair

Dr. David Walsh _____ External examiner

Dr. Selvadurai Dayanandan _____ Examiner

Dr. Paul Joyce _____ Examiner

Dr. Adrian Tsang _____ Supervisor

Approved by _____

Chair of Department or Graduate Program Director

Date _____

Dean of Faculty

ABSTRACT

Comprehensive Bioinformatic Analysis of Glycoside Hydrolase Family 10 Proteins

Sherry Wu

Glycoside Hydrolase Family 10 (GH10) contains endo-1, 4-β-xylanase which catalyzes the hydrolysis of xylan, the most abundant hemicellulose in lignocellulosic biomass. In this study, different bioinformatic approaches were used to comprehensively analyze the distribution, the phylogeny, the function and the evolutionary origin of a large GH10 protein dataset. The goal was to explore the correlation between sequence similarity and function of GH10 proteins to better understand xylan utilization pattern within the family.

Predicted glycoside hydrolase family 10 sequences from fungal, bacterial, archaeal, and non-fungal eukaryotic genomes as well as biochemically characterized proteins were used to perform a phylogenetic analysis. Based on the tree topology, 626 GH10 sequences were classified into 50 well-supported subfamilies. Among the analyzed sequences, 42 remained unclustered. The complex topology of the family tree suggests multiple duplication events followed by lineage specific gene loss during evolution. In addition, the Maximum Likelihood phylogeny of GH10 proteins does not mirror the previously established species taxonomic tree, suggesting that the divergence of the GH10 family ancestral gene preceded the appearance of the eukaryotic lineages.

A set of non-fungal GH10 proteins were manually curated employing criteria used in *myco*CLAP, a database for biochemically characterized fungal lignocellulose active enzymes. Experimental data of biochemically characterized GH10 proteins were mapped onto the phylogenetic tree to establish relationships, if any, between biochemical properties and sequence similarity. Only 24 subfamilies contain members with characterization, demonstrating that 26 phylogenetically diverse subfamilies remain uncharacterized. Among the subfamilies with experimental data, a distantly related subfamily with tomatinase activity was identified. By comparing the tertiary structures of well-characterized subfamilies, I have identified subfamilies that display different xylan substrate preferences and hydrolysis patterns. Correlations were also observed between sequence similarity and the pH and/or temperature optimum in the GH10 family. The accumulation of mutations within subfamilies reflects how they have diverged over time. Subfamily discriminating residue analyses were performed to identify subfamily-specific polymorphisms. Detailed lists of subfamily discriminating residues are provided. The majority of these residues are involved in secondary structure formation based on alignment to 3D structures, suggesting they might be functionally and structurally important.

ACKNOLEDGEMENTS

I would like to express my deepest gratitude to my supervisor Dr. Adrian Tsang, who gave me the chance to work on this exciting project. I highly appreciate the immense knowledge, the encouragement and the support he provided throughout the year. The completion of this thesis would not have been possible without his guidance. I would like to thank Dr. Selvadurai Dayanandan and Dr. Paul Joyce for being my committee members as well as Dr. David Walsh for being my external examiner.

I would like to express my sincere thanks to my dear friends Thi Truc Minh Nguyen and Caitlin Murphy who helped me in so many areas. In my daily work, I have been blessed with a supportive group of friends: Min Wu, Nadeeza Ishmael, Annie Bellemare, Erin McDonell, Dr. Wendy Findley, Dr. Ingo Morgenstern, and Carol Nyaga. I wish to express appreciation for their support, encouragement, and advices over the years.

In addition, I would like to thank the Department of Biology at Concordia University for giving me the opportunity to complete my Master degree and I would like to acknowledge people in the Center of Structural and Functional Genomics for their help.

Lastly, I would like to thank my family and friends for always being there for me and believing in me. I must express my sincere thanks to my best friend Carlos Molina who provided me with unconditional love, support, and care over the years.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Araf	α-L-arabinofuranoside
BLAST	Basic Local Alignment Search Tool
CAZy	Carbohydrate Active Enzymes
CE	Carbohydrate Esterase
EC	Enzyme Commission
GA	α-D-glucuronic acid
GH	Glycoside Hydrolase
IMG	Integrated Microbial Genomes
IUBMB	International Union of Biochemistry and Molecular Biology
JGI	Joint Genome Institute
MAFFT	Multiple Alignment using Fast Fourier Transform
MeGA	4-O-methyl-D-glucuronic acid
MP	Maximum Parsimony
ML	Maximum Likelihood
MSA	Multiple Sequence Alignment
MUSCLE	MUltiple Sequence Comparison by Log-Expectation
*myco*CLAP	Characterized Lignocellulose-Active Proteins (of fungal origin)
NCBI	National Center for Biotechnology Information
NJ	Neighbor-Joining
OUT	Operational Taxonomic Unit
PDB	Protein Data Bank
PL	Polysaccharide Lyase
PMID	PubMed Identifier
T-Coffee	Tree-based Consistency Objective Function for alignment Evaluation
TIM	Triosephosphate isomerase
UPGAM	Unweighted Pair Group method with Arithmetic Mean

CHAPTER ONE: INTRODUCTION

1.1 Lignocellulosic residues, a sustainable alternative of fossil fuels

1.1.1 First generation vs second generation biofuels

For over a hundred years, fossil fuels have been used as the primary source of transportation fuels and chemicals. Adverse environmental impact along with the finite nature of this energy source has prompted an intense search for more sustainable alternatives. Biofuels are considered the most promising alternative as they are produced from renewable biosources. In addition, the use of biofuels instead of fossil fuels decreases the net emission of greenhouse gases, which has been directly linked to global warming [1,2].

Biofuels can be classified into first and second generations. First generation biofuels are produced from sugar, starch, vegetable oils, and animal fat. On the other hand, second generation biofuels are generated from lignocellulosic materials. First-generation bioethanol has been in commercial production since the 1970s because the technologies for the conversion of sugar to alcohol are well understood. However, the use of food crops as the source of feedstocks has caused concerns such as the increase in food price and the decrease in biodiversity. Second-generation biofuels are considered more sustainable as lignocellulosic residues are the most abundant, non-edible, renewable resources on the planet. Second-generation biofuels are not yet in large-scale commercial production because the recalcitrance of lignocellulose materials makes the conversion process costly. Research dedicated to overcoming the technical barriers for

second-generation biofuel production has increased tremendously because of the potential environmental and socio-economic advantages over its first-generation counterpart [3,4].

1.1.2 Potential lignocellulosic feedstocks

Most lignocellulosic materials can be categorized into agricultural residues, forest residues, and energy crops [5,6].

Agricultural crop residues are materials left in the field after crop harvesting. They are consisted of stalk, stems, leaves, and seed pods. In addition, husks, seeds, and roots obtained after the processing of the crops are also considered as agricultural residues. Potential sources of agricultural crop residues include those derived from corn, sorghum, barley, rice, wheat, and sugarcane. It has been estimated that between 0.7 and 11.9% of the gasoline consumed in Canada can be potentially produced from agricultural crop residues [5–7].

Forest residues are produced from forest harvest operations and products processing. Hardwood and softwood are the two major woody biomass species. Hardwoods include birch, aspen, and willow whereas softwoods include spruce and pine [6,8].

A class of dedicated non-food crops are also potential feedstocks. These energy crops have attracted much attention because they can be grown on marginal croplands that are not suitable for other agricultural production. Also, these crops can be genetically modified to better meet the need of bioconversion. Most energy crops are herbaceous species such as switch grass, miscanthus, and alfalfa. Dedicated energy crops are more cost effective as lower energy inputs are needed. Their high yield also makes them a promising source of feedstock for second-generation biofuels. It has been suggested that energy crops may become Canada's largest new

renewable source with the potential of producing up to 117 billion litres of bioethanol annually [6,8].

1.1.3 Lignocellulose components

Biomass is the general term for organic materials that are composed of carbon polymers. More specifically, lignocellulosic material is used to refer to biomass derived from non-starch components of plants. Lignocellulosic biomass is mainly composed of cellulose, hemicellulose, pectin, and lignin. The amount of each component varies among different species. For instance, hardwood species have more cellulose than softwood species [1,2].

Cellulose is composed of linear polymers of D-glucose sugars that are linked by β-1, 4 glycosidic bonds. On the other hand, hemicellulose is mainly constituted of the 5-carbon sugar xylose. Other sugars in hemicellulose include arabinose, mannose, and galactose. The major difference between cellulose and hemicellulose is that the latter contains heterogeneously branched polysaccharides, which means that side chains can be added to the polymer backbone through various linkages. Pectin is another complex polysaccharide found in plants. The backbone of pectin is composed of α-1, 4-linked galacturonic acids or alternating α-1, 2-rhamnopyranosyl residues and α-1, 4-linked galacturonic acids. Finally, there is lignin, which is built from different phenylpropane units [9].

The major challenge of second-generation biofuel production is to generate sugar monomers from lignocellulosic biomass. In addition to the β-1, 4 glycosidic bonds that link glucose monomers, multiple intrastrand hydrogen bonds cause crystallinity in cellulose structure which makes its degradation very difficult. Cellulose is also surrounded by hemicellulose and

pectin. The most recalcitrant component is lignin, which protects other components from degradation. Physiochemical treatments are often used to solubilize lignin and partially disturb the crystallinity of cellulose. Once cellulose, hemicellulose, and pectin are more accessible, their sugar monomers can be released through enzymatic hydrolysis. Microorganisms such as fungi and bacteria produce a wide array of polysaccharide-degrading enzymes [10,11].

1.1.4 Lignocellulolytic enzymes

Lignocellulosic biomass is composed of polysaccharides containing diverse sugar monomers and their modified forms joined together in different chemical linkages. Therefore, to efficiently hydrolyze biomass, a wide range of enzymes are required. Among them, glycoside hydrolases (GHs) are the most important as they are responsible for the hydrolysis of the various glycosidic bonds that link monosaccharides. Glycoside hydrolases can be classified based on the sequence similarity of their catalytic domains [12]. As of 2014, there are 113 GH families classified in the Carbohydrate-Active enZymes database, http://www.cazy.org/ (CAZy) [13]. Sequences within the same family share common characteristics such as structural folding and mode of action, which reflect the evolutionary relatedness among the family members [12,14]. Alternatively, GHs can also be classified based on the type of reaction they catalyze and on their substrate specificity. Enzyme Commission (EC) numbers based on the International Union of Biochemistry and Molecular Biology (IUBMB) are assigned to enzymes with different substrate specificity [15]. When analyzing lignocellulolytic enzymes, it is preferable to combine the two classification methods as one protein family may contain multiple enzyme activities or the same enzyme activity can be found in multiple GH families [14].

In addition to GHs, carbohydrate esterases (CE) and polysaccharide lyases (PL) are also involved in the degradation of lignocellulose. Carbohydrate esterases catalyze the deacylation of the substituted polysaccharides. In the CAZy database, they are grouped into 16 CE families. Polysaccharides lyases, which break the glycosidic bonds of uronic acid-containing polysaccharides through non-hydrolytic cleavage, are classified into 19 PL families [13].

1.2 Xylan hydrolysis

1.2.1 Xylan structure

Xylan is the most abundant type of hemicellulose with highly diverse structural features, depending on the plant sources (Table 1). The backbone of xylan is a linear polymer composed of D-xylose residues. The structure of xylan can be highly diverse as substituents can be added to the backbone. The most common side chains of xylan include α-D-glucuronic acid, 4-O-methyl-α-D-glucuronopyranoside, acetyl groups, and α-L-arabinofuranoside. The proportion of added substituents varies among plant species [16–18]. In general, xylan can be grouped into six structural subclasses: homoxylan, glucuronoxylan, (arabino)glucuronoxylan, arabinoxylan, (glucurono)arabinoxylan, and complex heteroxylan [17].

Homoxylan is a linear polysaccharide composed of xylose sugars. The sugars can be linked by β-1, 4 linkages (X4), β-1, 3 linkages (X3) as well as mixed β-1, 4 - β-1, 3 linkages (Xm). The essential feature of homoxylan is the absence of side chains on the xylose backbone. The β-1, 3 linkage and mixed β-1, 4 - β-1, 3 linkage homoxylans are commonly found in red

algae and green algae. The occurrence of homoxylan in higher plants is rare. Plants mostly contain heteroxylan consisting of a β-1, 4 linked D-xylose backbone and side chains [16].

Glucuronoxylan is a type of heteroxylan with α-D-glucuronic acid and/or its 4-O-methyl derivative attached at the O-2 of the xylose monomer (Figure 1A). Glucuronoxylans are mostly found in hardwoods and herbaceous plants of the temperate zone and can make up to 90% of the hemicellulose. In hardwoods, an acetyl group can also be added to the positions O-2 and/or 3 of the xylose backbone residue [17,19].

When an α-L-arabinofuranoside residue is added to the position O-3 of the previously described glucuronoxylan, the resulting heteroxylan is an (arabino)glucuronoxylan (Figure 1B). In temperate zone softwoods, this form of xylan is a minor hemicellulose component whereas in tropical softwood, it is about 50% of the hemicellulose. In addition, the proportion of 4-O-methyl-α-D-glucuronic acid is higher in softwood (arabino)glucuronoxylan than hardwood glucuronoxylan. The lignified tissues of grass and cereals are also rich sources of (arabino)glucuronoxylans [17,19]. Contrary to glucuronoxylan of hardwoods, (arabino)glucuronoxylans of softwoods are not acetylated and they are also shorter than hardwood xylans [20].

A β-1, 4 linked D-xyloses backbone can be mono-substituted at position O-2 or O-3 and/or di-substituted at both O-2 and O-3 position with α-L-arabinofuranoside residue to generate arabinoxylan (Figure 1C). In addition, the α-L-arabinofuranoside chain can be esterified with one or more phenolic acids such as ferulic acid. Arabinoxylan is a major hemicellulose component of the cell walls of cereal grains such as wheat, rye, barley, and oat [17,19].

A heteroxylan is called (glucurono)arabinoxylan when its backbone is di-substituted with α-L-arabinofuranoside residue as well as α-D-glucuronic acid and/or its 4-O-methyl derivative (Figure 1D). This form of xylan is found in the straw of cereal as well as in the grains of rice, maize and sorghum [16,17].

There is also another group of xylan, generally referred as heteroxylan, with very complex structure. The backbones of these xylans are heavily substituted with side chains. They can be isolated from cereal bran, seeds and gum exudate. Tropical dicots also contain highly diverse heteroxylan [17].

Table 1: Structural diversity of xylan and its occurrence in nature
This table lists the constituents of different forms of xylan. Abbreviation: **Araf**, α-L-arabinofuranoside; **GA**, α-D-glucuronic acid; **MeGA**, 4-O-methyl-D-glucuronic acid.

Form of xylan	Backbone	Side chain	Source
Heteroxylan	β-1, 4-linked D-xylose; β-1, 3-linked D-xylose; β-1, 3 - 1, 4-linked D-xylose	None	Green algae and seaweed
Arabinoxylan	β-1, 4-linked D-xylose	Araf	Cereal grains
Glucuronoxylan	β-1, 4-linked D-xylose	GA; MeGA	Hardwood;
(Arabino)glucuronoxylan	β-1, 4-linked D-xylose	MeGA; Araf	Softwood; Lignified tissues of cereal and grasses;
(Glucurono)arabinoxylan	β-1, 4-linked D-xylose	MeGA; Araf	Lignified tissues of cereal and grasses

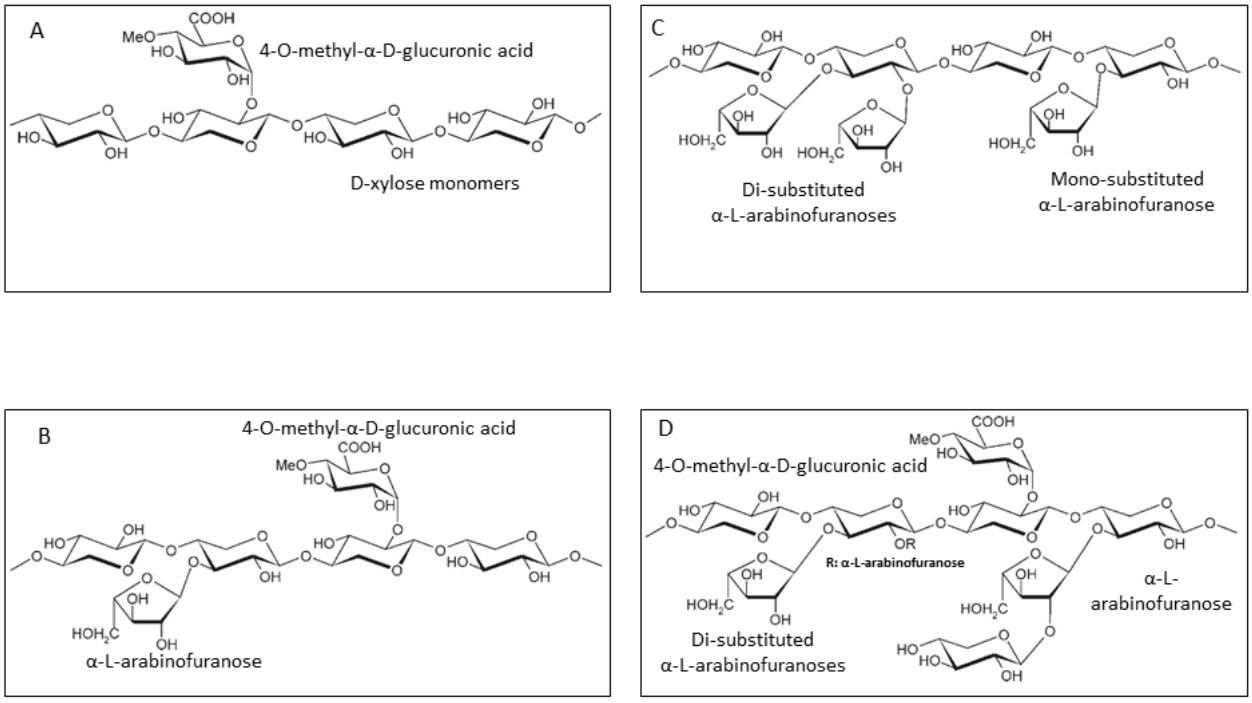


Figure 1: Diversity of heteroxylan
Structural features of (A) Glucuronoxylan, (B) (Arabino)glucuronoxylan, (C) Arabinoxylan, and (D) (Glucurono)arabinoxylan (Adapted from [17].)

1.2.2 Biorefinery of xylan

Biorefinery is the concept of processing lignocellulosic feedstocks into biofuels and other valuable bio-products. Xylan is, after cellulose, the most abundant source of lignocellulosic biomass. Depending on the source, xylan can occur up to 60% of the plant's dry mass (Table 2). In addition to biofuels, various valuable bio-products are xylan-based. Xylitol is considered as the most popular and marketable product derived from xylan fermentation. It can be used as a low-caloric sweetener and a preventive agent against dental cavities. Short xylooligosaccharides chains can be used as prebiotics in the food industry. It was also shown that these oligomers of xylose have a positive effect on human health. For instance, it was reported that

xylooligosaccharides can control the amount of ammonia in blood and can be used as an antioxidant against many diseases. The many uses of xylan derived products make it a promising feedstock for biorefinery systems [21].

Table 2: Xylan content value of various lignocellulose sources
The content of xylan is shown as the percentage of dry mass (Adapted from [21]).

Source	Category	Xylan content in dry mass (%)
Acacia dealbata	hardwood	16.4
Populus tremuloides (Aspen)	hardwood	17.7
Eucalyptus globulus	hardwood	16.6-18.0
Wheat straw	agricultural residue	18.1-29.4
Corn cob	agricultural residue	29.9-31.9
Corn stover	agricultural residue	17.3-22.8
Plantago ovata Forsk seed husk	agricultural residue	62.5
Miscanthus x *giganteus*	energy crops	19.0
Switchgrass	energy crops	17.7-25.3

1.2.3 Xylan-active enzymes

As described previously, the structure of xylan can be highly diverse due to the addition of various side chains. Hence different enzymes are needed for the complete hydrolysis of xylan. Enzymes involving in the degradation of xylan are referred as xylanolytic enzymes (Figure 2; Table 2) [20].

The internal β-1, 4 linkages of xylan backbone are hydrolyzed by xylanase, also known as endo-1, 4-β-xylanase (EC 3.2.1.8). Xylanases are highly diversified in terms of sequence similarity, structure, biochemical properties, and substrate specificity. Initially, xylanases were classified into families F and G [18]. It was suggested that while enzymes from family F have high molecular weight (>30kDa) and acidic pI, members of family G have low molecular weight (<30kDa) and basic pI. However, as new xylanases were discovered and experimentally characterized, only about 70% of the enzymes can be properly classified using this system [19]. Later on, xylanases were classified into glycoside hydrolase families based on sequence similarity and families F and G were renamed as families GH10 and GH11, respectively [22]. Structural analyses have shown that while the catalytic domain of GH10 xylanase forms a triosephosphate isomerase (TIM)-barrel fold consisting of eight α-helices and eight β-strands, its GH11 counterpart displays a β-jelly-roll architecture composed of two β-strands and a α-helix [23,24]. In addition, previous studies have shown that xylanases from these two families attack the same substrate differently (section 1.3) [25]. A few enzymes with endo-1, 4-β-xylanase activity have also been identified in families GH5, GH7, GH8, and GH43 [13]. Once endo-1, 4-β-xylanase breaks xylan polymer into shorter fragments, β-xylosidase hydrolyzes these oligosaccharides from their non-reducing ends to generate xylose monomers (EC 3.2.1.37). This enzyme is found in families GH3, GH39, GH43, GH52, and GH54 [14,18]. The α-L-arabinofuranoside side chains from heteroxylan are removed by α-L-arabinofuranosidases (EC 3.2.1.55) which are found in families GH43, GH51, GH54, and GH62 [20,26]. Xylan α-1, 2-glucuronosidase (EC 3.2.1.131) and α-glucuronosidase (EC 3.2.1.139) are two other de-

branching enzymes. They are responsible for the removal of α-D-glucuronic acid and/or its 4-O-methyl derivative from the xylan backbone. Xylan α-1, 2-glucuronosidase works specifically on hardwood glucuronoxylans and can be grouped into families GH67 and GH115. The major difference between these two families is that xylan α-1, 2-glucuronosidases from GH67 only target glucuronosyl linkage at the non-reducing ends of the xylooligosaccharides whereas enzymes from GH115 are capable of removing side chains from both the internal and terminal regions of the substrate [9,27–29].

In addition to GHs, a set of CEs also participate in the hydrolysis of xylan. As mentioned previously, acetyl groups can be added to positions O-2 and/or 3 of the xylose backbone residue in glucuronoxylan. These acetyl groups can be deacetylated by acetylxylan esterase (EC 3.1.1.72). The removal of acetyl groups is important as they may interfere with the interaction between glycoside hydrolases and the substrate by steric hindrance. Acetylxylan esterase can be found within CE families 1-7, and 12 [26,30,31]. Another esterase called acetylesterase (EC 3.1.1.6) has also been recently characterized and is assigned to CE16. While acetylxylan esterases from CE families 1-7 prefer polymeric xylan, the recently characterized CE16 acetylesterase (EC 3.1.1.6) removes acetyl groups linked to xylose or shorter xylooligosaccharides [9,32]. The α-L-arabinofuranoside side chains of arabinoxylan are frequently esterified with phenolic acids such as ferulic and *p*-coumaric acids. Ferulic acid esterase (EC 3.1.1.73), which belongs to CE1, is responsible for the hydrolysis of the ester bond between the arabinose side chain and ferulic acid or *p*-coumaric acid [28].

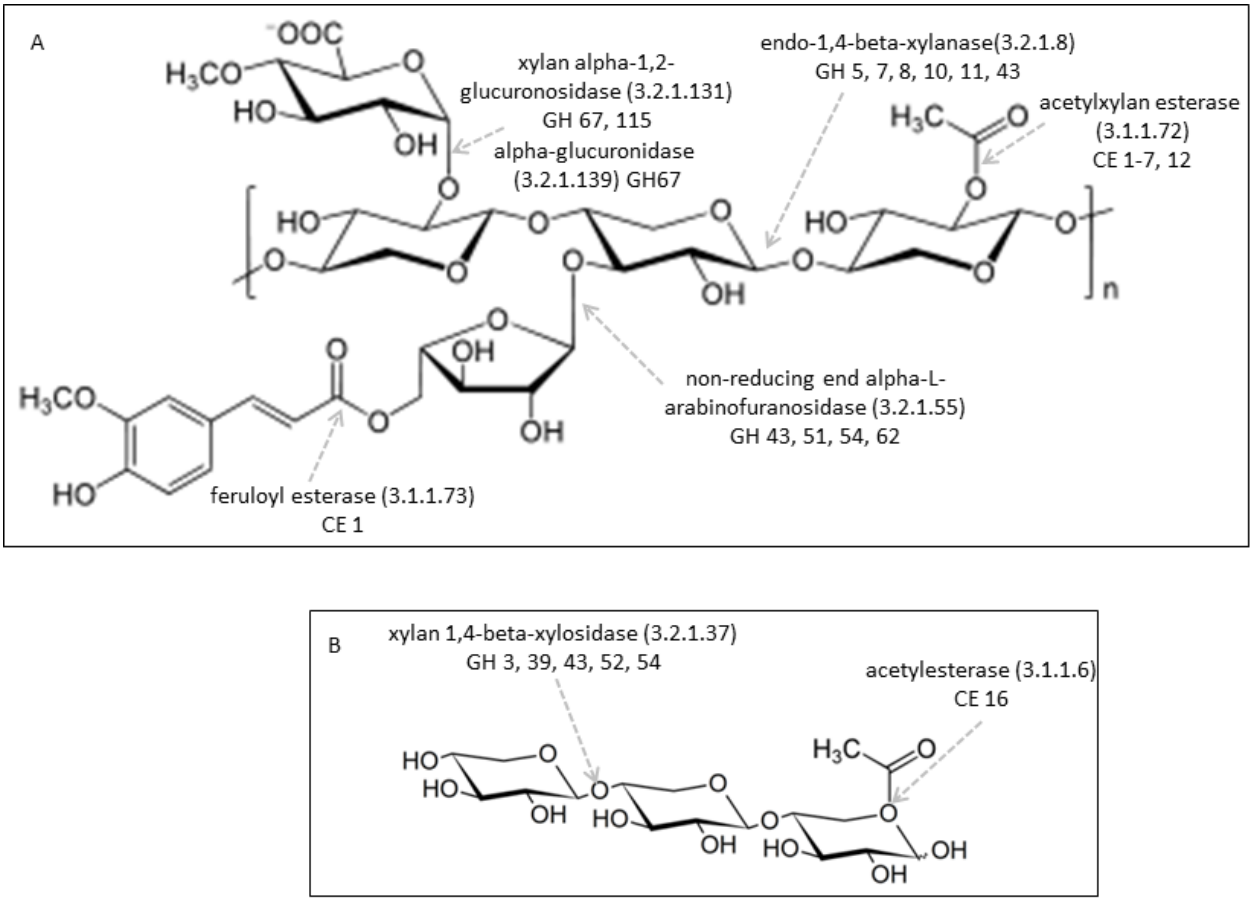


Figure 2: Degradation of xylan by xylanolytic enzymes

Panel A shows the hydrolysis of a polymeric xylan. The backbone is attacked by endo-1, 4-beta xylanase and the side chains are removed by various de-branching enzymes. Panel B shows the hydrolysis of a short xylooligosaccharides. Xylan 1, 4-beta-xylosidase releases a xylose monomer from the non-reducing end and the acetate group is removed by acetylesterase (Adapted from [26]).

Table 3: Hemicellulases involved in the hydrolysis of xylan

Enzyme recommended name	EC number	CAZy family	Catalysis
endo-1,4-beta-xylanase	3.2.1.8	GH 5,7,8,10,11,43	the endohydrolysis of 1,4-β-D-xylosidic linkages in xylans
xylan 1,4-beta-xylosidase	3.2.1.37	GH 3,39,43,52,54	the release of D-xylose residues from the non-reducing end of xylans
non-reducing end alpha-L-arabinofuranosidase	3.2.1.55	GH 43,51,54,62	the removal of α-L-arabinofuranosyl side chains from xylans
xylan alpha-1,2-glucuronosidase	3.2.1.131	GH 67,115	the removal of α-1,2-linked (4-O-methyl)glucuronosyl side chains in hardwood xylans
alpha- glucuronosidase	3.2.1.139	GH67	the removal of 4-O-methyl)glucuronosyl side chains in xylans
acetylxylan esterase	3.1.1.72	CE 1-7,12	the removal of acetyl esters from acetylated xylans
acetylesterase	3.1.1.6	CE16	the removal of acetyl esters from acetylated xylose & short xylooligosaccharides
feruloyl esterase	3.1.1.73	CE1	the removal of ferulic & coumaric acid from xylans

1.2.4 Xylan-utilizing organisms

A wide range of xylanolytic enzymes are needed for the degradation of xylan. Fungi and bacteria are the dominant xylan-utilizing microorganisms. Other xylan-degrading organisms include marine algae, protozoans, and land plants. Xylanolytic enzymes have also been found in archaea [33,34]. All of the aforementioned xylanolytic enzymes have been purified from fungi

and bacteria [20]. However, it has been shown that the amount of enzymes secreted by fungi is much higher than bacteria [35].

1.3 Glycoside Hydrolase family 10 xylanase

Glycoside hydrolase families 10 and 11 contain the majority of the endo-1, 4-beta-xylanases (EC 3.2.1.8). Although xylanases from both families catalyze the hydrolysis of internal beta-1,4-xylosidic linkages in xylan, they differ in their biochemical properties, amino acid sequences, tertiary structures, as well as reaction mechanisms [12,18]. When acting on heteroxylan, GH10 xylanases can hydrolyze the substrate to a higher degree. They are able to attack xylosidic bonds that are close to branched xylose residues and generate shorter products than their GH11 counterparts. Some GH10 xylanases also display "exo" activity as they can release terminal xylose residues attached to substituted residues. When acting on xylooligosaccharides, GH10 xylanases cleave shorter substrates more efficiently than GH11 xylanases. It has been proposed that GH10 xylanases contain a lower number of subsites where xylose residues can bind [25].

Three-dimensional structures of GH10 xylanases have been determined from bacteria and fungi. Figure 3 shows that GH10 xylanases fold into a TIM-barrel which is consisted of eight major β-sheets surrounded by eight α-helices. The catalytic cleft, where the substrate binds and gets cleaved, is located at the narrower end of the barrel close to the C-terminus of the enzyme. GH10 xylanases cleave xylosidic linkages through a double-displacement "retaining" mechanism using a proton donor and a nucleophile. Two invariant catalytic glutamate residues

located on β-sheets 4 and 7 have been identified as the proton donor and the nucleophile, respectively [36–39]. For the cleavage of the substrate, xylose residues bind to a series of binding subsites within the catalytic cleft. The subsites are labelled from $-n$ to $+n$ where the negative and positive integers correspond to the non-reducing end and the reducing end of the xylose chain, respectively. By convention, bond cleavage occurs between the xylose residues at the -1 and +1 subsites. In addition, the negatively labeled subsites are also referred as the glycone region whereas the positively labeled subsites are the aglycone region [40]. It has been shown that xylose residues at the glycone region make abundant and specific hydrogen bonds with a set of highly conserved amino acids. It is suggested that this region acts like a "substrate recognition area" that dictates the substrate specificity of the enzyme. On the other hand, the amino acids at the aglycone region of the cleft are less conserved across the family and their interactions with the ligand are weaker as only stacking interactions are observed. This region acts like a "product release area" where the products can be easily released after hydrolysis due to the low affinity of the subsites towards the xylose residues [41]. While the overall structure of the enzyme within the GH10 family is well conserved, the number of the subsites and their affinity towards xylose residues vary. These differences contribute to the binding preference of xylanases on different xylan substrates.

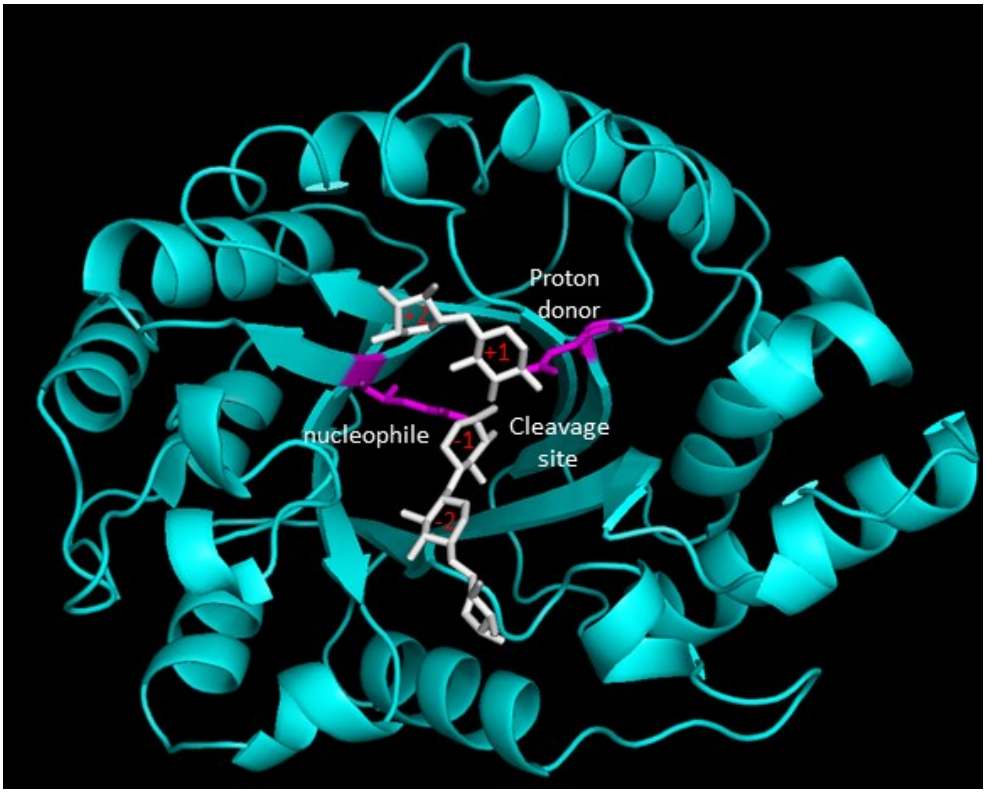


Figure 3: Tertiary structure of GH10 xylanase
Xylanase of *Penicillium simplicissimum* (cyan) is bound to xylopentose (white) PDB: 1b3z [41]. The five xylose rings of xylopentose occupy subsites -3 to +2 where it is cleaved at -1 and +1 subsites into xylotriose and xylobiose. The proton donor and nucleophile are colored in magenta.

1.4 Bioinformatic approaches

1.4.1 Genome databases

The number of sequenced genomes has increased rapidly due to the recent improvement in sequencing technology. The Genome Portal (http://genome.jgi.doe.gov) is a database established by The Department of Energy (DOE) Joint Genome Institute (JGI) to generate and store sequence data [42]. In addition, different tools are implemented within the portal for data

annotation and analysis. The Genome Portal covers sequenced genomes from four areas: plants, fungi, microbes, and metagenomes. MycoCosm (http://jgi.doe.gov/fungi) is a web-based database integrated inside the Genome Portal which contains all fungal genomes [43]. MycoCosm stores not only fungal genomes sequenced by JGI, but also fungal genomics data from other sources such as Fungal Genome Initiative of Broad institute [44]. At its release in March 2010, MycoCosm contained over 100 annotated fungal genomes and this number has increased rapidly since. The integrated microbial genomes database (IMG) is a data management, analysis, and annotation platform of the JGI Genome Portal that includes publicly available genomes from bacteria, archaea, and eukaryotes (https://img.jgi.doe.gov/cgi-bin/w/main.cgi). IMG was first released in 2005 and contained a total of 296 genomes. The current Version 4.0 (2014) of IMG contains 18,390 genomes including 13,178 bacterial, 442 archaeal, and 189 eukaryotic genomes along with others from plasmids, viruses and genome fragments. New genomes are added on a quarterly basis [45]. Phytozome (http://www.phytozome.net) is another web-based platform within The Genome Portal that focuses on plant genomes. Phytozome contains both JGI and non-JGI genomes [46].

1.4.2 Multiple sequence alignment

The quality of the multiple sequence alignment (MSA) plays a key role in the phylogenetic analysis as it directly affects the accuracy and the reliability of the subsequent result. Due to its importance, many programs have been introduced over the past decade. The available methods can be categorized into five algorithmic approaches: exact approach, structure-based method, progressive alignment, iterative approach, and consistence-based alignment [47]. Exact approach is computationally unfeasible for a dataset with more than a few

sequences. Structure-based method is useful when aligning distantly related sequences from different protein families. These two methods are deemed unsuitable for the purpose of this study and will not be explored further.

The most widely used progressive alignment program is ClustalW [48]. ClustalW aligns a set of proteins in three steps. First, a pairwise alignment score is calculated for every pair of proteins and converted into a distance matrix. Second, a guide tree is derived from the distance matrix. Based on the distance matrix, the algorithm selects the two most related sequences and creates a pairwise alignment. Finally, sequences are added progressively to the pairwise alignment according to their position on the guide tree. The final MSA profile generated using ClustalW depends on the order in which sequences are added. Once a sequence is aligned and a gap is introduced, no modification can be made even when it conflicts with sequences that are added later. Therefore progressive approach does not guarantee the best alignment profile [49,50].

Multiple sequence alignment algorithms that are based on iterative approach overcome the "uncorrectable alignment" limitation of the progressive alignment method. Basically, iterative programs use a progressive approach to generate an initial alignment. Then, they apply iterative refinement to modify and improve the quality of the alignment [49]. MAFFT (Multiple Alignment using Fast Fourier Transform) and MUSCLE (MUltiple Sequence Comparison by Log-Expectation) are the two most popular programs employing iterative approach [51,52]. These two methods differ in their iteration step. MUSCLE generates a draft progressive alignment profile based on a rooted tree. This rooted tree is derived from the distance matrix calculated from the similarity of the input sequences. Then, MUSCLE constructs another rooted

tree using a Kimura distance matrix which takes into consideration that multiple substitutions may occur at the same position. The main purpose of the second stage is to improve the quality of the tree and build a new progressive alignment. The new tree is compared to the previous tree to identify internal nodes with changed branching order. This step can be iterated. If all of the new trees have identified these changed nodes, one can conclude that the old tree can be improved. A new progressive alignment is only built for sequences with changed branching order, hence correcting the previous draft progressive alignment. In the last stage, MUSCLE performs a refinement step. Refinement starts with the deletion of a branch from the tree to obtain two subsets of sequences (create a bipartition). The MSA profile of each subset is extracted and empty columns are removed from the profile. The two profiles from the subsets are then re-aligned using profile-profile alignment. Finally, MUSCLE chooses to accept or reject the new alignment depending on the sum of pair (SP) score which assesses the quality of the alignment. The new alignment is accepted if the score increases. All the branches of the tree are deleted sequentially to create a bipartition. The refinement step stops when no change can be made after all the branches have been visited which means that the quality of the alignment cannot be improved further [51]. The other iterative approach MAFFT also generates an initial alignment using progressive alignment method and corrects it with iterative refinement. The major difference is that MAFFT incorporates information from homologous sequences from external databases to obtain a more accurate alignment of the submitted sequences. These extra sequences are then removed [50,52]. Both MUSCLE and MAFFT are shown to be very computational efficient.

The premise of consistency-based alignment is that if residue *x* of sequence A aligns with residue *y* of sequence B and *y* aligns with residue *z* of sequence C, then residues *x* and *z* should align with each other. The previously described progressive and iterative methods use a guide tree to generate the MSA profile. This guide tree is based on the pairwise sequence alignment score of the input sequences. Consistence-based approach also generates pairwise sequence alignment scores. However, when aligning two sequences, their alignments to other sequences are also taken into account [50]. T-Coffee (Tree-based Consistency Objective Function for alignment Evaluation) is based on this approach [53]. T-Coffee starts by generating a primary library consisting of global and local pairwise alignments of the input sequences. Then, every pair of aligned residues is assigned a weight using sequence identity. Take an example with sequences A, B, C, and D, six global pairwise sequence alignments are generated and a primary weight is assigned to each pair of sequences. In addition, T-Coffee generates local alignments between sequences and only those ten with the highest primary weight are included into the library. Then, a library extension is performed for each pair of sequences. During this step, the pairwise sequence alignments from the primary library are aligned to other sequences of the dataset. For example, there are three possible alignments to extend the global alignment of sequence A and B: align A and B, align A and B through C or align A and B through D. A weight is assigned to each of the three possible alignments. By doing so, the algorithm evaluates the residues aligned in A and B with the rest of the sequences in the library and gives the correct alignment. A position-specific substitution matrix, called "extended library" is calculated from these assigned weights. This position-specific substitution matrix is used to generate pairwise alignments of the input sequences which then can be aligned in a progressive manner. The

advantage of T-Coffee over ClustalW is that the pairwise alignment score generated by the former takes into account the consistency with other sequences of the database as a position-specific substitution matrix is used instead of a general substitution matrix [49,53].

Case studies have been done to assess the performance of different MSA tools. For example, five distantly related globins including beta globin, myoglobin, and neuroglobin from human as well as soybean leghemoglobin, and nonsymbiotic plant hemoglobin were aligned using ClustalW, MUSCLE, and T-Coffee [47]. The alignments of three highly conserved residues of the globin family were used to evaluate the quality of the MSA profile. These amino acids include a phenylalanine and two histidines. The result showed that ClustalW and MUSCLE were able to align the phenylalanine and the first histidine but failed to align the second histidine. On the other hand, T-Coffee aligned all three conserved residues properly. In addition, MAFFT was also used to align these sequences and compared to the results obtained from the above methods [47]. Table 4 summarizes the comparison of the different MSA tools.

Table 4: Comparison of different MSA tools

The advantages and disadvantages of the most popular MSA tools are summarized. The ability of aligning the three conserved residues of the globin family was used to evaluate the accuracy.

Program	Algorithm approach	Advantages	Disadvantages
ClustalW	progressive	fast	unable to make a correction once a misalignment is introduced; does not guarantee optimal alignment; only works well for closely related sequences
MUSCLE	iterative	fast; able to correct misaligned	less accurate; unable to align

Program	Algorithm approach	Advantages	Disadvantages
		position through iterative refinement steps	conserved residues of distantly related globins
MAFFT	iterative	fast; accurate; able to correct misaligned position through iterative refinement steps; external sequences are included to obtain a more accurate alignment; refinement step also includes consistency-based score	none
T-Coffee	consistency based	accurate; pairwise alignment score is supported by evidence from multiple sequences; both global & local alignment are assessed	slow

1.4.3 Phylogenetic inference

Phylogeny is the evolutionary history of species as they change over time. This history can be illustrated through a phylogenetic tree. Most phylogenetic tree-building methods can be grouped into two categories: distance-based and character-based. Distance-based phylogenetic methods use a distance matrix of pairwise dissimilarity which measures the evolutionary divergence between every pair of aligned sequences to generate the phylogenetic tree. The branch lengths of the phylogenetic tree should reflect as closely as possible the observed distances [54,55]. On the other hand, character-based methods examine characters (nucleotide for DNA sequence and amino acid for protein sequence) at every single site of the multiple sequence alignment to assess the reliability of each position on the basis of all other positions [56,57]. Most character-based methods rely on the use of optimality criterion. These methods compare alternative tree phylogenies based on a defined criterion and the goal is to search for the

optimal tree topology under that criterion. Different methods employ different optimality criterion [58]. The common feature of all tree building methods is that they all presume an evolutionary model to infer phylogeny. These models explain how one DNA nucleotide or a particular amino acid is substituted by another. Models differ by the mutation rule and pattern they incorporate [59,60]. The use of accurate model of evolution is critical to extract information from molecular sequence data. Models of sequences evolution were generated through the incorporation of biological, biochemical, and evolutionary knowledge. The use of inadequate and oversimplified models can lead to incorrectly inferred phylogenetic trees which reflect erroneous evolutionary relationships. In the past 30 years, the complexity of the models continues to increase as our knowledge of sequence evolution patterns accumulate. The use of accurate and realistic models allows robust evaluation of complex evolutionary hypotheses [61,62]. However, it is beyond the scope of this paper to include description and comparison of all existing models. Instead, I will describe and compare different methodologies that use models of sequence evolution to estimate phylogenetic trees in the following section.

Distance-based phylogenetic inference approach was pioneered by unweighted pair group method with arithmetic mean (UPGMA) which is based on a sequential clustering algorithm [63,64]. [64]The first step of UPGMA is to generate a distance matrix that shows the estimation of the similarity between every pair of sequences. Two sequences with the shortest evolutionary distance are selected first. These two most closely related sequences are grouped together to form a cluster which represents an internal node in the phylogenetic tree. From then on, these two sequences are treated as a single taxon and a new matrix is constructed to show the evolutionary distance between this newly assigned taxon to other sequences. From the new

matrix, the next pair of closest sequences is combined into a new cluster and another distance matrix is generated with these sequences treated as a single taxon. These steps are repeated until all the sequences are clustered. Trees generated by UPGMA are automatically rooted because this approach assumes that the rate of DNA or amino acid substitution is constant for all the branches in the tree [63,65]. This assumption is the major pitfall of UPGMA because it produces incorrect tree when there are unequal substitution rate along different branches [55,65].

Neighbor-Joining (NJ) method is the most widely used distance-based phylogenetic method introduced by Saitou and Nei in 1987 and later modified by Studier and Keppler [66,67]. NJ method first generates a starlike tree with all of the sequences. This tree topology has no hierarchical structure and is produced under the assumption that there is no clustering of sequences. From this starlike tree, NJ algorithm identifies and joins pairs of operational taxonomic units (OTUs). OTUs are joined based on a rate-corrected distance matrix which does not assume an equal substitution rate along all the branches. Once two OTUs are identified as neighbors, they are connected through a single interior node. Take for example a dataset of eight sequences denoted Seq1-Seq8 (N=8). At the beginning, these eight sequences are connected by a single node designated X to form a starlike tree. Based on the distance matrix, NJ algorithm identifies Seq1 and Seq2 to be a pair of neighbors with the shortest distance. Seq1 and Seq2 are joined together to form a new node U. At this point, the tree has two internal nodes: Seq1 and Seq2 form node U whereas Seq3-Seq8 are connected by node X. The two internal nodes are joined by the internal branch X-U. Once Seq1 and Seq2 are paired as neighbors, they are treated as a single taxon and the next step is to compute a new distance from node U that joins them together to other sequences to identify the next pair of neighbors. This new set can either be two

sequences or one sequence coupled with the previously assigned neighbors. For instance, Seq3 can be paired with Seq4, or Seq3 can be combined with OTU Seq1-Seq2. This procedure is repeated until all the internal branches are found. Since NJ algorithm does not assume a constant evolution rate, it produces an unrooted bifurcating tree in which all of the internal nodes are connected to only three other branches [55,66].

Maximum parsimony and maximum likelihood are the two most popular character-based methods to infer phylogeny. Maximum parsimony analysis is based on the premise that the best tree is the one that requires minimal evolutionary changes which is defined by the number of substitution among sequences. According to this theory, a simpler explanation for the observed data is preferred over the more complicated alternatives hence the phylogenetic tree obtained from maximum parsimony algorithm is the topology having the smallest total number of changes. This tree is referred as the most parsimonious tree. Maximum parsimony algorithm starts by categorizing aligned sequence sites into informative sites and non-informative sites. A column of the alignment is considered non-informative when the residues at this position are entirely conserved. A column is also defined as non-informative when only one sequence has a different residue. A position is informative when there are at least two character states (residues) with at least two sequences having each state. Non-informative sites are not analyzed by the algorithm as they do not contribute to the discrimination of the trees. Then, the algorithm generates a dataset of trees. A cost that represents the number of substitutions from hypothetical ancestral sequences to observed sequences is assigned to each tree. Maximum parsimony selects the tree with the lowest cost. All the possible trees are evaluated by the algorithm when the

dataset contains a dozen or fewer sequences. However, when analyzing a larger set of data, only trees that are most likely to be the most parsimonious one are evaluated [65,68].

Maximum likelihood (ML) approach seeks to find an adequate explanation for a given data set by varying all the parameters of a model of evolution until the highest possible likelihood is found. In the context of molecular evolution, the parameters of the model are the branching order and branch lengths of the phylogenetic tree whereas the given data set is the DNA or protein sequences. Provided with a model of evolution, the ML approach evaluates probability of generating the observed data under the chosen model [58,69]. In other words, it finds the evolutionary tree which yields the highest probability of evolving the observed data [70]. As mentioned previously, character-based methods examine characters at every single site of the aligned sequences. Therefore, a likelihood is calculated for each residue in an alignment. The likelihood for a particular site is the sum of the probabilities of every possible reconstruction of the ancestral state. The likelihood of the tree is the product of the likelihoods at each site. The tree with the highest likelihood is selected [58].

The major disadvantage of distance-based methods is that the actual character is not used to generate the tree. The tree is built based on the amount of dissimilarity between two sequences, hence, it is often less accurate. Despite its potential inaccuracy, distance-based methods are still widely used because they are computationally less intensive. The reason why distance-based methods are faster is that they generate only one tree using a specific algorithm. On the other hand, character-based methods generate and evaluate many trees and select the one that best answers the optimality criterion [58]. Among character-based methods, maximum likelihood usually outperforms maximum parsimony as shown by case studies [71–73]. The

major pitfall of maximum parsimony is the generation of an inconsistent tree with long-branch attraction. In long-branch attraction, two rapidly evolving sequences are clustered together on the tree because they both have many mutations. It is misleading as one may interpret from the tree that these two sequences are closely related [58]. Table 5 summarizes the features of the different tree building methods.

Table 5: Comparison of different tree building methods

Method	Category	Advantages	Disadvantages
Unweighted-Pair Group Method with Arithmetic Means	distance-based	fast; simple	does not use actual character data; provides only one tree; only works well when substitution rate is equal
Neighbor Joining	distance-based	fast; simple	does not use actual character data; provides only one tree; less accurate
Maximum Parsimony	character-based	actual character data are used; optimality criterion is used to evaluate alternative phylogenies	less accurate than ML; may produce misleading branch-attraction; relatively slow
Maximum Likelihood	character-based	actual character data are used; optimality criterion is used to evaluate alternative phylogenies; least affected by sample error; more accurate	relatively slow

1.5 Experimentally characterized xylanases

1.5.1 *myco*CLAP: A database for biochemically characterized fungal lignocellulose active genes

The increasing number of newly sequenced genes predicted to encode lignocellulose activities allows us to explore the distribution and abundance of xylanase genes through phylogenetic analysis. However, it is also essential to have a core set of biochemically characterized enzymes to help us further understand how proteins within the same family have evolved and how the major clades are structured. *myco*CLAP is a database that contains biochemically characterized lignocellulose active proteins of fungal origin (https://mycoclap.fungalgenomics.ca/mycoCLAP/). All the sequences stored in the *myco*CLAP database fulfill the following criteria: for a gene to be defined as characterized, its sequence has to be publically available; an experimental assay has to be performed on the gene product for its activity; and the biochemical properties of the enzyme have to be published in a peer-reviewed journal [74].

1.6 Rationale for this study

Due to the recent improvement in sequencing technology, the number of newly sequenced and predicted carbohydrate-active proteins is increasing rapidly. Currently, CAZy database contains about 340 000 CAZymes, which is a ~225% increase in five years [75]. For instance, glycoside hydrolase family 10 which contains industrially important xylanases has 1,765 sequences as of 2014. With genome sequencing becoming so efficient, it is impossible to

experimentally characterize all of the predicted genes. At this stage, a comprehensive analysis of the protein family using bioinformatic approaches is more realistic and valuable in framing future research. With this rationale in mind, I decided to investigate the large dataset of GH10 sequences. This thesis describes the phylogenetic analysis of GH10 proteins and further classification of the sequences into subfamilies. I also investigated annotated proteins with experimental evidence and incorporated these data into the analysis in the hope of establishing relationships, if any, between sequence similarity and biochemical properties of the gene product. I would like to explore whether the family phylogenetic tree can be used to predict the biochemical properties of an uncharacterized enzyme and eventually be used as a tool to select target proteins for further biochemical characterization. During the process, I manually curated a comprehensive set of non-fungal GH10 proteins that have been experimentally characterized. This dataset will be added to the existing *myco*CLAP database which currently only contains characterized glycoside hydrolases of fungal origin. In addition, the phylogenetic tree of the family also allows us to further understand the evolution and distribution of GH10 genes. Lastly, I performed subfamily discriminating residues analyses to identify subfamily-specific polymorphisms. These polymorphisms may cause proteins from different subfamilies to utilize xylan differently. It will be interesting to investigate how these residues affect the structure and function of the enzymes belonging to different subfamily.

In this study, different bioinformatic approaches were used to comprehensively analyze GH10 protein family. It is hoped that the approaches described in this study can be used towards a standardized framework to analyze other protein families.

CHAPTER TWO: MATERIALS AND METHODS

2.1 Sequence retrieval

Predicted protein sequences of sequenced genomes were retrieved from various genome databases. The main reason for using sequences from fully sequenced genomes over individually sequenced genes is that the number of paralogs from each genome is taken into account. The selected genomes represent a wide taxonomic spectrum of each Kingdom. Glycoside hydrolase family 10 protein sequences of fungal species were collected from MycoCosm (http://jgi.doe.gov/fungi) whereas sequences from bacteria, archaea, and non-fungal eukaryotes other than plants were retrieved from Integrated Microbial Genomes (https://img.jgi.doe.gov/cgi-bin/w/main.cgi) [42,45]. Finally, Phytozome was used to gather data from plants (http://www.phytozome.net) [46]. Sequences were retrieved from annotated genomes using pfam domain ID of interest (PF00331 for GH10 proteins). Pfam database (http://pfam.sanger.ac.uk/) contains a large collection of protein families. Within the database, a pfam domain ID based on hidden Markov models and multiple sequence alignment is assigned to each family [76]. The advantage of using domain ID over BLAST search is that relatively diverged family members are included in the analysis. Datasets collected from BLAST search may differ depending on the query used.

2.2 Multiple sequence alignment

Once all the sequences were retrieved, they were trimmed to their domain limits. Domain is the most conserved part of the protein that contains all the motifs as well as the catalytic residues. Multiple sequence alignment generated using protein domains is more significant as less ambiguously aligned sites are produced. In the case where a sequence had multiple domains, each domain was treated as an individual sequence. MAFFT (http://www.ebi.ac.uk/Tools/msa/mafft/) were used for MSA as it was shown to be more accurate and less time consuming (Section 1.4.2; Table 2) [52]. Then, the MSA profile was examined manually and sequences missing conserved residues or motifs were removed to improve the quality of the dataset.

2.3 Phylogenetic analysis

Maximum Likelihood trees were generated using RaxML (Section 1.4.3; Table 5) [77]. A bootstrap value of 1000 was used to estimate branch support. Subfamilies were assigned based on the topology of the phylogenetic tree. A subfamily was assigned if it included three or more sequences and supported by 55% or more of the bootstrap replicates. To validate the subfamilies, a sequence similarity analysis was performed using in-house scripts. Within group average pairwise percent identity, the arithmetic means of all of the individual pairwise percent identity between two sequences, was calculated for each subfamily (Figure 4A). Furthermore, between group average pairwise percent identity was calculated from the arithmetic means of all

individual pairwise percent identity between two inter-groups sequences (sequences from different subfamilies). This average pairwise identity was calculated for every pair of subfamilies (Figure 4B). The idea was that the average pairwise percent identity of sequences within the same subfamily should be higher than the average pairwise percent identity of sequences from different subfamilies since sequences belonging to the same subfamily should be more closely related.

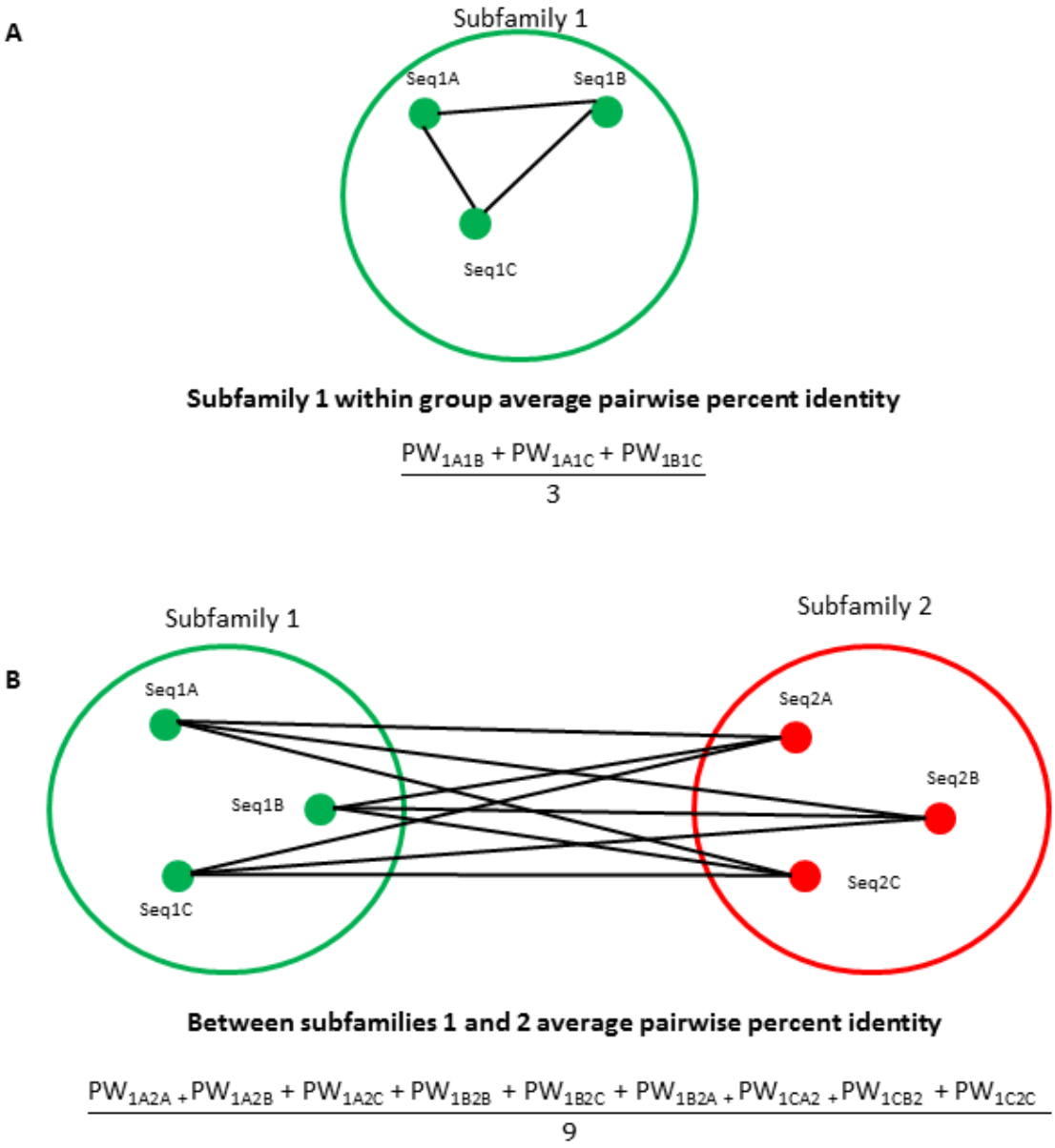


Figure 4: Sequence similarity analysis of subfamilies

Each subfamily is designated by a number and the letters represent the members of the subfamily. (A) Within group average percent identity of a subfamily containing three sequences. (B) Average percent identity between sequences from two subfamilies.

2.4 Experimentally characterized GH10 genes

Sequences encoding biochemically characterized GH10 xylanases of fungal origin were collected from the *myco*CLAP database (https://mycoclap.fungalgenomics.ca/mycoCLAP/) [74]. Crystal structures of GH10 xylanases were collected from Protein Data Bank (PDB) (http://www.rcsb.org/pdb/) [78]. pyMOL was used to display and align 3D structures [79].

Criteria used in the curation of fungal glycoside hydrolase genes [74] were applied to the curation of bacterial, non-fungal eukaryotes and archeal genes encoding biochemically characterized xylanases. Extracted sequence information and experimental data were organized into a spreadsheet as described by Murphy et al., [74].

2.5 Amino acid conservation analysis

Absolute conservation, hydrophobic conservation, and polar conservation methods described by Liu et al. [80] were used to assess the conservation level of each position in an alignment profile.

For the absolute conservation method, the conservation level of a particular position in the alignment is defined by the absolute conservation score $A_{conservation}$. For each position, an absolute conservation score is calculated for each amino acid *a* as follows:

$$A_{conservation}(a) = \frac{N(a)}{n} \text{ where } N(a) = \sum_{i=0}^{n} f(i) \quad f(i) = \begin{cases} 1, & C = a \\ 0, & C \neq a \end{cases}$$

In this formula, *C* represents the amino acid used by a particular sequence at that position. In other words, the absolute conservation score of amino acid *a* at position *n* is the number of

sequences that have residue *a* at this position divided by the total number of sequences in the alignment profile.

The hydrophobicity of a position in an alignment is defined as conserved when the amino acids occurring at this position belong to the same hydrophobic class (C_i). Similar to the absolute conservation method, a hydrophobicity conservation score $H_{conservation}$ is calculated for each position:

$$H_{conservation}(a) = \frac{N(a)}{n} \text{ where } N(a) = \sum_{i=0}^{n} f(i)$$

$$f(i) = \begin{cases} 1, & hydrophobic\ class\ (Ci) = hydrophobic\ class\ (a) \\ 0, & hydrophobic\ class\ (Ci) \neq hydrophobic\ class\ (a) \end{cases}$$

The amino acids are categorized according to their hydrophobicity as follows:

- (CVLIMFW): hydrophobic
- (RKEDQN): hydrophilic
- (PHYGAST): neutral

The hydrophobicity score is defined as the number of sequences with amino acids that belong to one hydrophobic class divided by the total number of sequences. Since there are three hydrophobic classes, each position will have three hydrophobicity conservation scores.

The polarity of a position in an alignment is defined as conserved when the amino acids occurring at this position belong to the same polar class (C). The polarity conservation score $P_{conservation}$ for each position is calculated similarly to the hydrophobicity conservation score:

$$H_{conservation}(a) = \frac{N(a)}{n} \text{ where } N(a) = \sum_{i=0}^{n} f(i)$$

$$f(i) = \begin{cases} 1, & polar\ class\ (Ci) = polar\ class\ (a) \\ 0, & polar\ class\ (Ci) \neq polar\ class\ (a) \end{cases}$$

The polarity of the amino acids is defined as follows:

- (GAVLIPFWM): non-polar
- (SCNQYT): polar-uncharged
- (EDHRK): polar-charged

The conservation level of each position within the alignment was calculated using all three methods. A position is considered conserved when one of the three conservation scores is greater than 0.90.

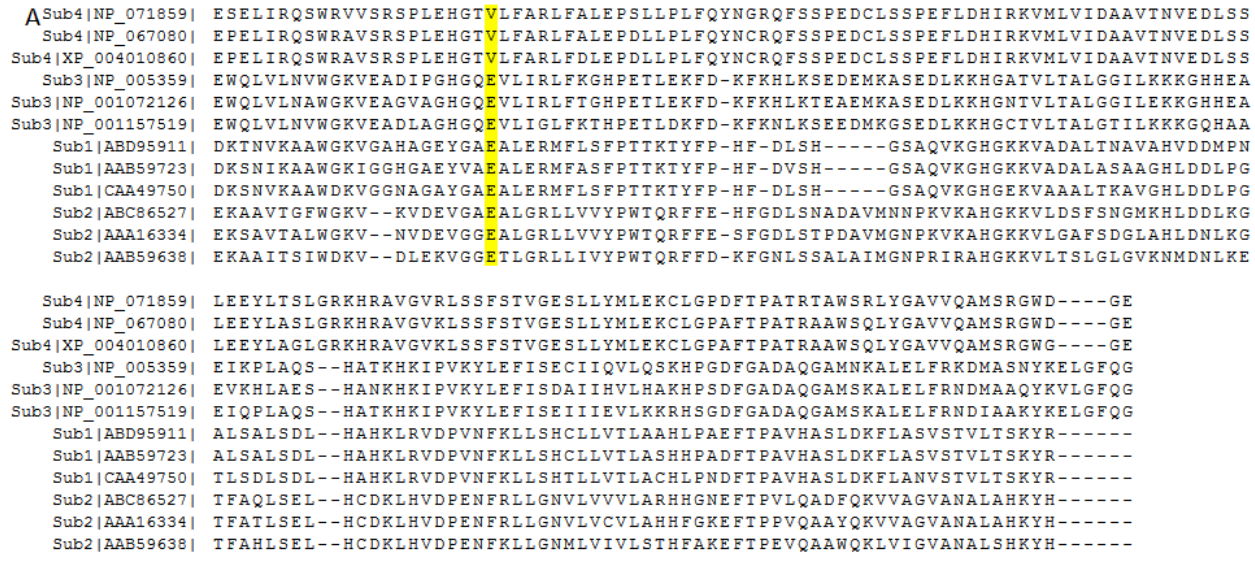
2.6 Subfamily discriminating residues analysis

Absolute, hydrophobicity, and polarity conservation scores were re-calculated separately for each subfamily to evaluate its conservation level [80]. Only residues with subfamily-conservation scores as well as global conservation score that exceeded 60% were used for discrimination analysis. In other words, a particular residue has to be more than 60% conserved within each subfamily as well as across the whole family. A position was defined as subfamily discriminating when the following conditions were met [81]:

1) The subfamily conservation score exceeded 60% for all subfamilies.
2) At this position, the amino acid or the property (hydrophobicity or polarity) of the amino acid used by the discriminating subfamily was different from the other subfamilies.

Take an example of an alignment profile consisting of globin sequences (Figure 5A). According to the phylogenetic tree, these globin sequences are clustered into four subfamilies: alpha globins, beta globins, myoglobins, and neuroglobins (Figure 5B). At position 22 of the

multiple sequence alignment (highlighted in yellow), the absolute conservation score of the amino acid calculated from all the sequences of the family is 0.75 for residue glutamate (E), which means that 75% of the sequences from this family have a glutamate at position 22. On the other hand, neuroglobins from subfamily 4 have a valine (V) at this position with a conservation level of 1.0. In this case, amino acid 22 is conserved across the global family as 75% of the sequences use glutamate at this position. However, for subfamily 4, valine is used instead of glutamate hence amino acid 22 is defined as a subfamily discriminating position for neuroglobins subfamily 4.



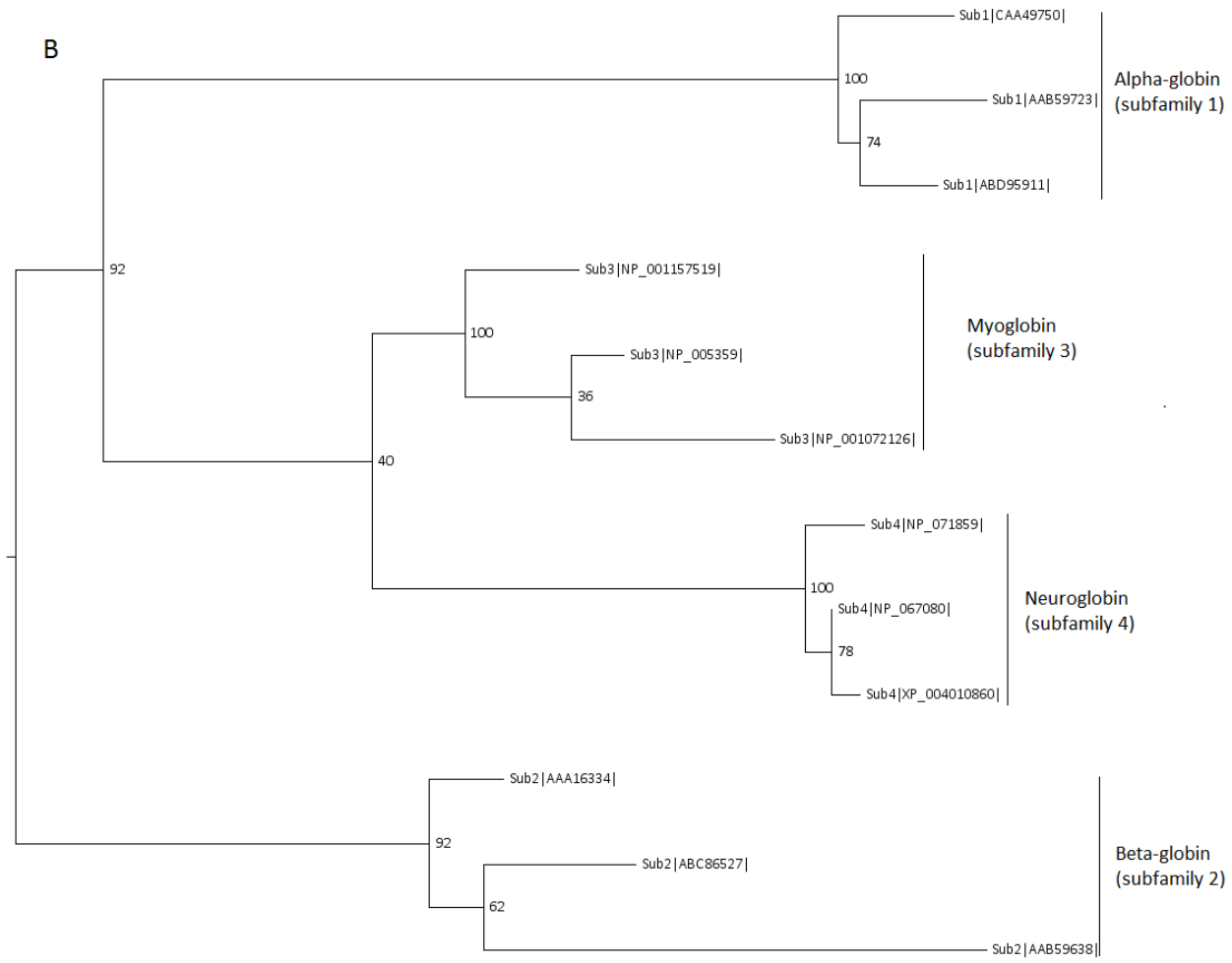


Figure 5: Subfamily discriminating residues analysis

(A) MSA profile of four globin subfamilies: alpha globin, beta globin, myoglobin, neuroglobin. Amino acid at position 22 (highlighted in yellow) is an example of subfamily discriminating residues. While subfamilies 1, 2, and 3 use glutamate (E) at this position, valine (V) is used by sequences of subfamily 4. (B) Maximum likelihood tree of globin sequences. The tree is rooted at mid-point and a bootstrap value of 100 was used. The four subfamilies are well supported.

CHAPTER THREE: RESULTS

3.1 Distribution of genes encoding Glycoside Hydrolase Family 10 proteins

3.1.1 Fungi

Publicly available fungal genomes from MycoCosm of the Department of Energy Joint Genome Institute (http://genome.jgi.doe.gov/programs/fungi/index) [42] were analyzed for the presence of GH10 protein-encoding genes (Figure 6). At the time of the last analysis for the thesis (January 2014), MycoCosm held 354 fungal genomes. Among them, 251 contain one or more GH10 xylanase-encoding genes. Ascomycota and Basidiomycota are the major phyla with the most sequenced genomes.

The phylum Ascomycota can be divided into three subphyla: Pezizomycotina, Saccharomycotina, and Taphrinomycotina [82]. Taphrinomycotina is considered to be the earlier diverged lineage within Ascomycota. Fungi within this subphylum include facultative biotrophic plant pathogens, yeast-like species, and fission yeasts, which are highly diverse in terms of morphology and ecology [83,84]. Within this subphylum, the genomes of seven fungi have been sequenced and *Taphrina deforman* of the class *Taphrinomycetes* is the only sequenced species harboring GH10 genes. This fungus is a pathogen that mainly causes peach leaf curl disease. The remaining six fungi of this subphylum are from the classes *Schizosaccharomycetes* and *Pneumocystidomycetes*. *Schizosaccharomycetes* contain fission yeasts whereas species from *Pneumocystidomycetes* are pathogens found in the lungs of mammals [84]. All these fungi lack GH10 genes which is consistent to their ecological niches. The subphylum Saccharomycotina

contains ascomycete yeasts. Ascomycete yeasts share morphological similarities and their mode of life as saprobes. However, phylogenetic analyses have shown that the genomes of yeast can be quite diverse even when they are from the same order [85]. Of the 31 ascomycete yeasts with sequenced genomes, *Scheffersomyces stipitis* is the only one that contains GH10 genes. It has been shown that *Scheffersomyces stipitis* belongs to a clade of yeasts that are capable of fermenting xylose, a rather rare trait among yeasts [85]. *Scheffersomyces stipitis* is the only member of this clade with a sequenced genome. It will be interesting to see if other species from this clade also contain GH10 genes. The largest subphylum Pezizomycotina has more than 32 000 filamentous and ascoma-producing fungi classified. Species belonging to this subphylum can be further assigned into 11 classes based on their morphologies and molecular phylogenies [86,87]. MycoCosm contained 176 sequenced fungal genomes belonging to 8 classes: *Dothideomycetes* (52), *Eurotiomycetes* (53), *Leotiomycetes* (10), *Sodariomycetes* (49), *Orbiliomycetes* (2), *Lecanoromycetes (2)*, *Xylonomycetes* (1), and *Pezizomycetes* (7). Among these classes, only species from *Xylonomycetes* and *Lecanoromycetes* lack GH10 protein-encoding genes. However, it is impossible to judge if the lack of GH10 protein-encoding genes reflects the whole group or the limited number of sequenced genomes available. On the other hand, MycoCosm does not contain fungal genome from the classes *Arthoniomycetes*, *Lichinomycetes*, and *Laboulbeniomycetes*. BLAST search against the non-redundant protein sequences database of NCBI did not retrieve any GH10 ortholog from these classes.

The other extensively studied phylum Basidiomycota is composed of subphyla Agaricomycotina, Pucciniomycotina, and Ustilaginomycotina [88]. Agaricomycotina is the largest subphylum representing one third of the described species in the Fungal Kingdom.

Members of this subphylum include mushrooms, jelly fungi and yeasts. Genome sequencing of Agaricomycotina species are of interest as they are mostly wood and litter decomposers. Some of the members are also pathogens of plants and humans. *Dacrymycetes*, *Tremellomycetes*, and *Agaricomycetes* are the three classes of the subphylum [89,90]. Among the 94 sequenced Agaricomycotina species in MycoCosm, 86 are from *Agaricomycetes* with 80 of them containing GH10 genes. Species of *Dacrymycetes* and *Tremellomycetes* also contain GH10 genes. Eight sequenced fungal genomes of the subphylum Pucciniomycotina were available for the analysis; five of them harbor GH10 genes. All five species belong to the order *Pucciniales* which contains rust fungi that are obligate plant parasites derived from insect and non-vascular plant parasite lineages [91]. Members of the subphylum Ustilaginomycotina are basidiomycetous plant parasites mostly of angiosperms. Species from this subphylum can be further grouped into three classes based on their morphological characteristics: *Entorrhizomycetes*, *Ustilaginomycetes*, and *Exobasidiomycetes* [92]. MycoCosm currently contains seven fungal genomes that belong to the two latter classes. All fungi from *Ustilaginomycetes* contain GH10 genes. The Malasseziales species from *Exobasidiomycetes* represent a unique order within the Ustilaginomycotina subphylum as they are isolated from the skin of warm-blooded animals [92]. Species from this order lack GH10 protein-coding genes.

Subphyla Mucoromycotina, Entomophthoromycotina, and Kickxellomycotina were previously classified into the Zygomyta phylum. However, this classification is obsolete as phylogenetic analysis showed that this phylum is artificial [93]. While Entomophthoromycotina contains insect pathogens, the subphylum Kickxellomycotina consists of saprobes, mycoparasites, and symbionts of aquatic arthropods. MycoCosm contained a single genome for

each of these subphyla and none of them harbor GH10 protein-encoding genes. Mucoromycotina is another subphylum which contains mostly saprobes. Currently, eight fungal genomes from this subphylum have been sequenced and only *Umbelopsis ramanniana* contains GH10 genes. It has been shown that this fungus represents an early diverging lineage within Mucoralean fungi [43,94]. *Piromyces sp.* is an anaerobic fungus isolated from the gut of elephant. This species has an expansion in the number of GH10 protein-encoding genes (28 copies). Microsporidia is a basal phylum containing eukaryotic parasites that are intracellular [95]. None of the eight sequenced fungi of this phylum have GH10 protein-encoding genes. One fungal genome for each of the basal lineages Cryptomycota, Blastocladiomycota, and Chytridiomycota has been sequenced [96] and they all lack GH10 protein-encoding genes. Glomeromycota contains arbuscular mycorrhizal fungi, which are mutualistic symbionts between land plants and fungi. Only one genome is available for this phylum and it lacks GH10 genes.

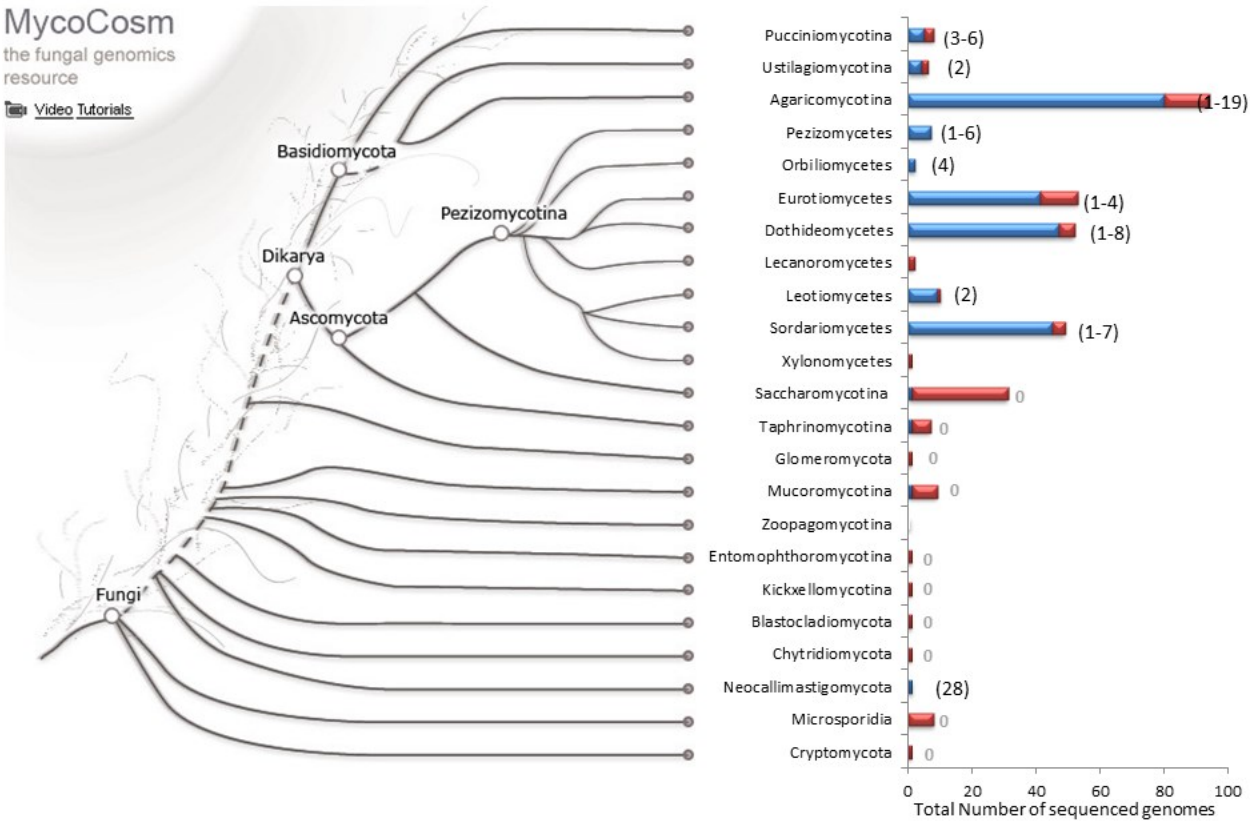


Figure 6: Abundance of GH10 protein-encoding genes within the Fungal Kingdom
Fungal taxonomy tree is obtained from MycoCosm, the Genome Portal of the Department of Energy's Joint Genome Institute [42]. The bar graph indicates the total number of sequenced genomes within the phylum or subphylum. The blue portion represents the number of the genomes that encode GH10 genes whereas the red portion represents genomes lacking GH10 genes. The number in bracket represents, if applicable, the range of gene copies produced by the species of the phylum or subphylum.

All fungi are characterized by their heterotrophic nutrition mode, which means they obtain energy (carbon-based compounds) from other organisms. The majority of fungi live off other organisms as saprotrophs or symbionts. Symbiotic fungi share intimate association with another species. This association can be pathogenic (parasitism) or beneficial (mutualism).

Furthermore, pathogenic fungi can be classified as plant or animal pathogens. On the other hand, saprotrophic fungi decompose dead plant matters to meet their energy requirement [84]. Here, the correlation between the ecological niche of each fungus and the abundance of GH10 protein-encoding genes were investigated. The majority of the analyzed fungi can be classified as saprobe, plant pathogen, plant mutualist, or animal pathogen. Some species can adopt more than one strategy. For instance, *Fomitiporia mediterranea* is a saprobe that also infects the wine grape, *Vitis vinifera* [97]. Some fungi also adopt other alternatives and live as gut symbionts, animal opportunists, or nematode-trapping fungi. The examination of the distribution of genes encoding GH10 proteins in fungi showed that closely related organisms do not necessarily share the same lifestyle and a similar number of GH10 encoding genes. It is observed that there is a stronger correlation between lifestyle and copy number than taxonomic relatedness. For example, *Ophiostma piceae* and *Grosmannia clavigera* are both Sodariomycetes from the order *Ophistomatales*. While *O. piceae* lives as a saprobe, *G. clavigera* is a bark beetle-associated pine pathogen [98,99]. Whereas *O. piceae* contains one GH10 gene, *G. clavigera* harbours none. The general trend observed is that while animal pathogens do not contain GH10 protein-encoding genes, only 8 of the 79 analyzed saprotrophic and plant pathogenic fungi lack these genes. Among these saprobes and plant pathogens, 64 species produce multiple GH10 genes. As for analyzed plant mutualists, 71% (10 of 14 surveyed) lack GH10 protein-encoding genes or only contain a single-copy gene. These plant mutualists are from *Agaricomycetes*, *Pezizomycetes*, and *Lecanoromycetes*. On the other hand, plant mutualists from the class *Leotiomycetes* harbour between 2-4 copies of GH10 genes. With 28 copies, the anaerobic fungus *Piromyces sp.*, isolated

from the gut of an elephant, an extreme and competitive ecological environment, has the highest number of GH10 genes.

3.1.2 Green plants

Green plant genome sequences obtained from Phytozome (http://www.phytozome.net) [46] were analyzed for the presence of GH10 protein-encoding genes. Green plants belong to the Kingdom Viridiplantae, which contains two major phyla. Members of the first lineage, Chlorophyta, are green algae such as large seaweeds. The other lineage, Streptophyta, contains mainly land plants and closely related green algae [100]. At the time of the last analysis for the thesis (January 2014), Phytozome contained 35 sequenced genomes of land plants (Streptophyta) and seven sequenced genomes of green algae (Chlorophyta). All land plants from Streptophyta harbor multiple GH10 genes. These land plants possess a higher GH10 gene copy number than most fungi. *Volvox carteri* is the only species from Chlorophyta that contains GH10 genes. This multicellular green alga has two GH10 gene copies (Figure 7).

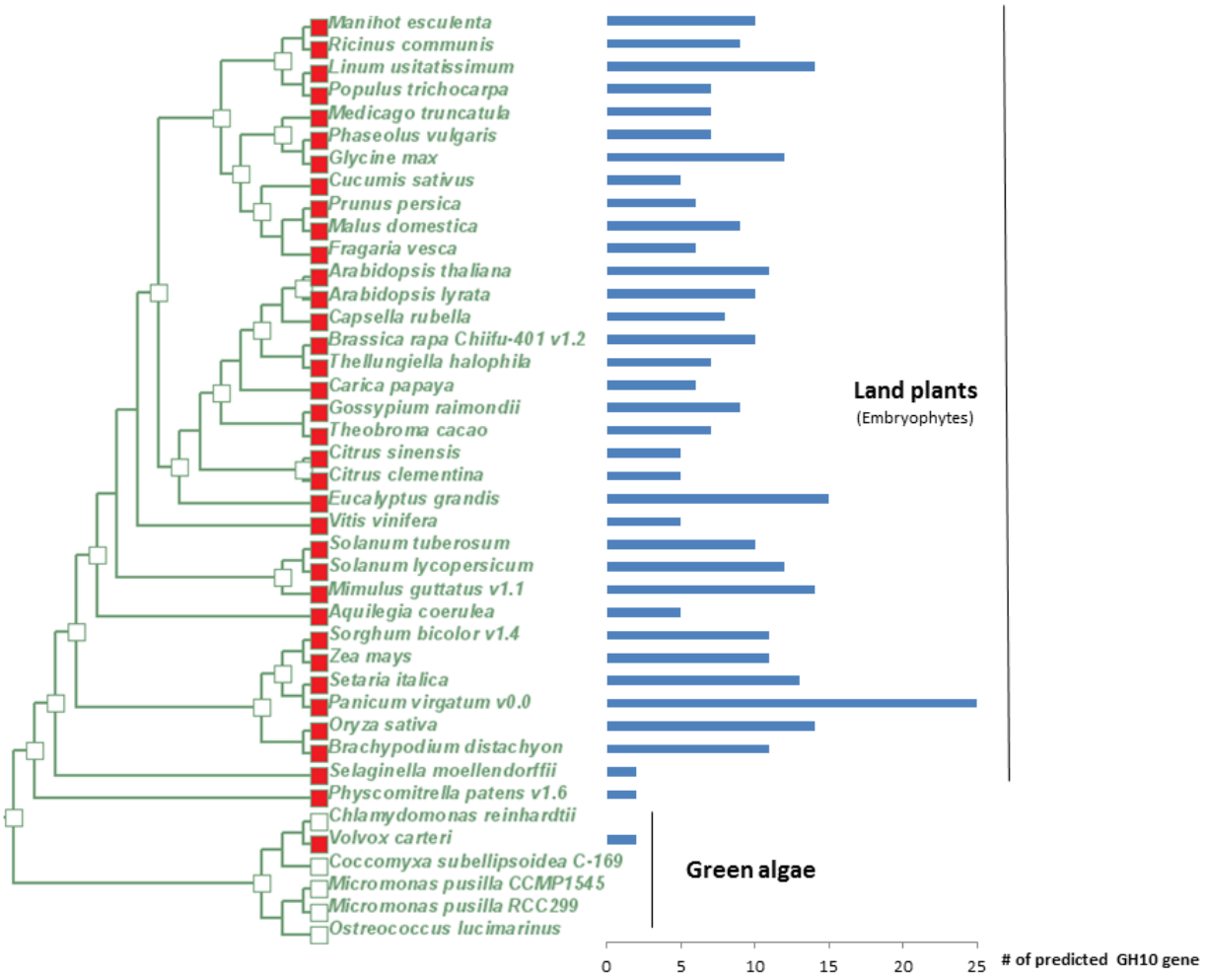


Figure 7: Abundance of GH10 protein-encoding genes within the Viridiplantae Kingdom
The Viridiplantae taxonomy tree is obtained from Phytozome [46]. Red nodes indicate GH10 xylanase-producing species. The bar graph indicates predicted GH10 gene copy number for each species.

3.1.3 Other eukaryotes

The Metazoa (Animal) tree of life was first introduced by Ernst Haeckel in 1866. Today, the Metazoan Kingdom contains 35-40 phyla including 1.3 million described species [101]. At the time of the last analysis for the thesis (January 2014), Integrated Microbial Genomes of the

Department of Energy Joint Genome Institute (https://img.jgi.doe.gov/cgi-bin/w/main.cgi) held 17 Metazoan genomes sequences from the phyla Annelida, Mollusca, Cnidaria, Chordata, Arthropoda, and Placoza [45]. Table 6 shows the presence of GH10 genes within these genomes.

Table 6: Abundance of GH10 protein-encoding genes within the Metazoan Kingdom

This table lists the presence of GH10 genes within the publicly available metazoan genomes sequences from IMG [45]. The available GenBank common names are obtained from NCBI (when available).

Species	Phylum	Genbank common name	Number of GH10 gene
Capitella teleta	Annelida	n.a	23
Helobdella robusta	Annelida	n.a	0
Daphnia pulex	Arthropoda	water flea	0
Branchiostoma floridae	Chordata	Florida lancelet	0
Ciona intestinalis	Chordata	vase tunicate	0
Fugu rubripes	Chordata	n.a	0
Xenopus tropicalis	Chordata	western clawed frog	0
Nematostella vectensis	Cnidaria	sea anemone	1
Lottia gigantea	Mollusca	owl limpet	13
Trichoplax adhaerens	Placozoa	n.a	0

Other than metazoan and plant species, the genomes of eukaryotic species without an assigned Kingdom were also sequenced and stored in IMG [102]. In total, 17 genomes were analyzed for the presence of GH10 protein-encoding genes (Table 7). Genomes of diatoms, oomycetes, and labyrinthulids are from the group Stramenophiles which is characterized by the presence of flagella with hairs. GH10 genes are present in oomycetes and diatoms but absent

from labyrinthulids (slime nets). Oomycetes are water molds and downy mildews whereas diatoms are single-celled algae. *Emiliania huxleyi* belongs to a group of organisms called haptophytes. It has been shown that this group is closely related to Stramenophiles. *Monosiga brevicollis* is a choanoflagellate which is a lineage of Opisthokonts, along with metazoa and fungi. *Dictyostelium discoideum* and *Acanthamoeba castellanii* are members of the group Amoebozoa which have been shown to be the sister group of Opisthokonts [103,104].

Table 7: Abundance of GH10 protein-encoding genes within other non-fungal eukaryotes
This table lists the presence of GH10 genes within non-fungal eukaryotes that have no taxonomy classification [45]. The available Genbank common names are obtained from NCBI (when available).

Species	Group	Genbank common name	# GH10 gene
Dictyostelium discoideum	Amoebozoa	cellular slime molds	0
Acanthamoeba castellanii	Amoebozoa	n.a	2
Monosiga brevicollis	Opisthokonta	choanoflagellates	1
Naegleria gruberi	n.a	n.a	0
Bigelowiella natans	Rhizaria	cercozoans	0
Emiliania huxleyi	n.a	haptophyte	2
Guillardia theta	n.a	cryptomonads	0
Aplanochytrium kerguelense	Stramenophiles	labyrinthulids	0
Aurantiochytrium limacinum	Stramenophiles	labyrinthulids	0
Aureococcus anophagefferens	Stramenophiles	n.a	0
Fragilariopsis cylindrus	Stramenophiles	diatoms	1
Phaeodactylum tricornutum	Stramenophiles	diatoms	1
Phytophthora capsici	Stramenophiles	oomycetes	6
Phytophthora ramorum	Stramenophiles	oomycetes	5
Pseudo-nitzschia multiseries	Stramenophiles	diatoms	1

Species	Group	Genbank common name	# GH10 gene
Schizochytrium aggregatum	Stramenophiles	slime nets	0
Thalassiosira pseudonana	Stramenophiles	diatoms	0

3.1.4 Bacteria

Bacterial genomes from Integrated Microbial Genomes (IMG) (https://img.jgi.doe.gov/cgi-bin/w/main.cgi) were analyzed for the presence of GH10 protein-encoding genes [45]. At the time of the last analysis for the thesis (January 2014), IMG contained 12,920 sequenced bacterial genomes belonging to 34 phyla. In addition, there were also 264 unclassified bacterial genomes. Table 8 shows the number of sequenced genomes for each phylum and the number of genomes encoding GH10 genes. Bacteria from 16 phyla lack GH10 genes. The distribution of the GH10 gene is not even as 1,001 out of 1,126 xylanase-producing bacteria are from Actinobacteria, Bacteroidetes, Firmicutes, and Proteobacteria.

Table 8: Abundance of GH10 protein-encoding genes within the Bacterial Kingdom

Column from left to right: bacterial phylum; number of publicly available sequenced bacterial genomes in IMG [45]; number of bacteria that contain GH10 genes; range of gene copy number (if applicable).

Phylum	# of sequenced Genome	# of Genomes containing GH10 gene	Range of gene copy number
Acidobacteria	23	6	1-3
Actinobacteria	1321	172	1-14
Aquificae	18	0	n.a
Armatimonadetes	3	3	1-2

Phylum	# of sequenced Genome	# of Genomes containing GH10 gene	Range of gene copy number
Bacteroidetes	516	173	1-7
Caldiserica	2	0	n.a
Chlamydiae	116	0	n.a
Chlorobi	13	0	n.a
Chloroflexi	32	4	1-2
Chrysiogenetes	1	0	n.a
Cloacimonetes	1	0	n.a
Cyanobacteria	199	49	1-3
Deferribacteres	6	0	n.a
Deinococcus-Thermus	42	8	1-2
Dictyoglomi	2	2	n.a
Elusimicrobia	3	0	n.a
Fibrobacteres	2	2	n.a
Firmicutes	3391	186	1-12
Fusobacteria	50	0	n.a
Gemmatimonadetes	7	1	n.a
Ignavibacteria	1	0	n.a
Ignavibacteriae	1	1	n.a
Lentisphaerae	3	2	n.a
Nitrospinae	1	0	n.a
Nitrospirae	9	0	n.a
Planctomycetes	28	16	1-5
Poribacteria	6	0	n.a
Proteobacteria	6490	470	1-12
Spirochaetes	412	10	1-5
Synergistetes	16	0	n.a
Tenericutes	147	0	n.a
Thermodesulfobacteria	5	0	n.a
Thermotogae	22	12	1-4
Verrucomicrobia	30	10	1-6

Actinobacteria is one of the largest phyla containing mostly gram-positive bacteria with high GC content. Most of the GH10 xylanase-producing Actinobacteria were isolated from soil, which have been shown to play a crucial role in the decomposition of biomaterials [105]. The phylum Bacteroidetes contains Gram-negative bacteria that can be further grouped into four classes: *Bacteroidia*, *Cytophagia*, *Flavobacteria*, and *Sphingobacteria* [106]. Bacteria from all four classes possess GH10 genes. Firmicutes is a phenotypically diverse phylum containing mostly Gram-positive bacteria. Firmicutes includes the following classes: *Bacilli*, *Clostridia*, and *Erysipelotrichia*. Only bacteria from *Erysipelotrichia* lack GH10 genes. Proteobacteria accounts for more than 40% of all published prokaryotic species [107]. IMG contained 6,490 available genomes from Proteobacteria and 470 harbor GH10 genes. Members of Proteobacteria can be further classified as *Alphaproteobacteria*, *Betaproteobacteria*, *Deltaproteobacteria*, *Gammaproteobacteria*, and *Episilonproteobacteria* [106,108]. Bacteria from all classes, except *Episilonproteobacteria,* contain GH10 genes.

3.1.5 Archaea

The sequences of 438 archaeal genomes from Integrated Microbial Genomes (https://img.jgi.doe.gov/cgi-bin/w/main.cgi) were used to analyze the presence of GH10 protein-encoding genes [42,45]. Euryarchaeota and Crenarchaeota are the two first described and established phyla. Most of the described archaea fall into these two major lineages. The phylum Euryarchaeota contains mostly methanogens, halobacteria, and thermophiles. On the other hand,

thermoacidophiles, extreme thermophiles, and sulfur-dependent archaea are members of the Crenarchaeota phylum [109]. Korarchaeota, Nanoarchaeota, and Thaumarchaeota are recently introduced phyla with a lower number of sequenced genomes [110–112]. Among all of the sequenced genomes with taxonomic classification, only species from the class *Halobacteria* of the phylum Euryarchaeota contain GH10 protein-encoding genes (Table 9). *Thaumarchaeota archaeon*, an unclassified archaeon also harbors a GH10 gene.

Table 9: Abundance of GH10 protein-encoding genes within the Archaeal Kingdom
Presence of GH10 genes within publicly available archaeal genomes from IMG [45].

Phylum	# of sequenced genomes	# of genomes containing GH10 genes
Crenarchaeota	109	0
Euryarchaeota	260	10
Korarchaeota	1	0
Nanoarchaeota	2	0
Thaumarchaeota	25	0
unclassified	45	1

3.2 Phylogenetic analysis of Glycoside Hydrolase Family 10

Because of the uneven number of sequenced genomes in different classes of fungi, green plants and bacteria, preliminary trees were generated to select representative genomes. *Agaricomycete*s from the phylum Basidiomycota as well as *Eurotiomycetes*, *Sodariomycetes*, and *Dothideomycetes* of the phylum Ascomycota have a significantly higher number of sequenced

genomes than the other classes in the Fungal Kingdom. Preliminary trees were generated for each of these classes to select representative fungal genomes. For bacterial GH10 proteins, preliminary trees were generated for the phyla Actinobacteria, Firmicutes, Proteobacteria, and Bacteroidetes. A preliminary tree was created for GH10 proteins from plants as well. After removing partial sequences and those missing conserved residues, a total of 508 predicted GH10 proteins from 165 sequenced genomes of different Kingdoms were selected for phylogenetic analysis (Table 10; Supplementary file 1).

Table 10: GH10 proteins used in phylogenetic analysis

Kingdoms	# of Genomes	# of GH10 sequences
Fungi	53	184
Plants	14	79
Metazoa	3	20
Other Eukaryotes	8	19
Bacteria	79	180
Archaea	8	26

In addition to sequences from sequenced genomes, experimentally characterized proteins were also included in the dataset. Sequences encoding biochemically characterized xylanases of fungal origin were retrieved from the *myco*CLAP database [74]. At the time of the last analysis for the thesis (January 2014), there were 31 experimentally characterized fungal GH10 proteins in *myco*CLAP. In addition, the protein data bank (PDB) contains five fungal GH10 proteins with experimental crystal structures. Following the criteria used by *myco*CLAP, a set of

experimentally characterized GH10 genes from organisms of other Kingdoms were manually curated. In total, 103 experimentally characterized bacterial GH10 enzymes were used in the analysis. Among them, 17 sequences also have available tertiary structures deposited in PDB. Two other uncharacterized proteins also have available crystal structures. The number of experimentally characterized GH10 genes in other Kingdoms is very scarce comparing to fungi and bacteria. None of the GH10 proteins has been characterized in Archaea. Two proteins from Metazoa have been characterized as well as a protozoan GH10 protein. In total, there are 143 GH10 proteins with functional data (Supplementary file 2). Sequences from these experimentally characterized proteins were combined with the previously selected GH10 proteins from sequenced genomes. After removing redundant sequences, 626 GH10 proteins were used in our phylogenetic analysis.

As mentioned in Materials and Methods, subfamilies were assigned based on the topology of the phylogenetic tree. Each subfamily must include three or more sequences and is supported by 55% or more of the bootstrap replicates. Phylogenetic analysis showed that 584 sequences can be clustered into 50 well supported subfamilies (Figure 8; Table 11; Supplementary file 3). Among them, 11, 28, 3, 1, and 2 subfamilies contain exclusively sequences from fungi, bacteria, land plants, metazoa, and archaea, respectively. Subfamilies containing sequences from species of different Kingdoms were also observed. For instance, subfamily 2 contains both sequences from fungi and oomycetes. Other multi-Kingdoms subfamilies are subfamily 4 (fungi and bacteria), subfamily 36 (bacteria and archaea), and subfamily 12 (fungi, bacteria, and protozoan). In addition, there are 42 sequences which remain unclustered in the phylogenetic tree. For example, eight fungal sequences fail to cluster with any

of the subfamilies. These sequences are from *Exobasidiomycetes*, *Tremellomycetes*, *Orbiliomycetes*, and *Pezizomycetes*. Since these classes have a limited number of sequenced genomes, it is tempting to predict that these unclustered sequences will eventually form other subfamilies when more genomic data become available. Other eukaryotic GH10 proteins from green algae, diatoms, and amoebae also failed to form subfamily due to the limited number of sequences. To validate the assignment of the subfamilies, within group average pairwise percent identity and between groups average pairwise percent identity were calculated for each subfamily and every pair of subfamilies, respectively (Supplementary file 4). As expected, the within group average pairwise percent identity of each subfamily is higher than the between groups average pairwise percent identity with all other subfamilies.


F (Basidiomycota)
F (Neocallimastigomycota)
F (Ascomycota/Basidiomycota)
Fungi & Bacteria (Tomatinase)
F (Basidiomycota)
F (Basidiomycota)
Fungi & Oomycetes
F (Ascomycota/Basidiomycota)
F (Ascomycota/Basidiomycota)
F (Basidiomycota)
Haptophytes
F (Neocallimastigomycota) & anaerobic Bacteria & Protozoan
Archaea & Bacteria
Land Plants
Metazoa
Green Algae

Figure 8: Maximum Likelihood phylogeny of Glycoside Hydrolase Family 10

The phylogenetic tree is rooted at mid-point. GH10 sequences are clustered into 50 subfamilies. Shown are subfamilies from: eukaryotes, blue; bacteria, grey; and archaea, pink. Clades with members from multiple Kingdoms are uncolored. Unclustered sequences are in black. Subfamilies containing biochemically characterized sequences are indicated with a red circle. The taxonomic distribution of eukaryotic as well as multi-Kingdoms subfamilies is also shown on the tree.

Table 11: GH10 subfamily classification

This table lists the number of sequences in each subfamily, the bootstrap value, and the average percent identity within each subfamily. The taxonomy distribution of each subfamily is also shown. Abbreviation: **F**, fungi; **B**, bacteria; **A**, Archaea. The number of experimentally characterized sequences found in each subfamily is also indicated.

Subfamily	Taxonomy	# of sequences	Bootstrap value	Within subfamily pairwise identity (%)	# of characterized proteins
S1	F (Basidiomycota; Ascomycota)	58	85	62.6	23
S2	F (Basidiomycota; Ascomycota) & Oomycetes	73	57	47.9	10
S3	F (Basidiomycota; Ascomycota)	12	100	57.8	N
S4	F (Ascomycota) & B (Actinobacteria)	11	100	66.0	1
S5	F Basidiomycota)	6	100	66.9	N
S6	F (Basidiomycota)	7	86	47.5	N
S7	F (Basidiomycota)	9	100	62.0	N
S8	F (Basidiomycota; Ascomycota)	5	90	54.3	N
S9	F (Basidiomycota; Ascomycota)	5	99	66.9	N
S10	F (Basidiomycota)	4	100	78.1	N
S11	F (Ascomycota)	5	100	61.0	N
S12	F (Neocallimastigomycota) & B (Firmicutes) & Protozoans	22	99	36.3	7
S13	F (Neocallimastigomycota)	5	100	57.6	N
S14	F (Neocallimastigomycota)	12	97	59.6	N

Subfamily	Taxonomy	# of sequences	Bootstrap value	Within subfamily pairwise identity (%)	# of characterized proteins
S15	A (Halobacteria)	9	100	62.6	N
S16	A (Halobacteria)	15	100	46.8	N
S17	Plants	14	100	70.6	N
S18	Plants	27	100	54.8	N
S19	Plants	36	97	64.2	N
S20	Metazoa	20	58	38.6	1
S21	B (Bacteroidetes)	7	100	50.4	2
S22	B (Deinococcus)	4	99	79.1	N
S23	B (Proteobacteria; Acidobacteria)	5	67	53.0	1
S24	B (Fibrobacteres)	3	100	72.9	N
S25	B (Firmicutes)	3	73	61.4	1
S26	B (Thermotogae; Chloroflexi)	7	100	76.8	3
S27	B (Actinobacteria; Chloroflexi)	17	84	64.9	13
S28	B (Bacteroidetes; Proteobacteria)	4	100	46.3	4
S29	B (Actinobacteria)	10	78	50.6	4
S30	B (Bacteroidetes; Proteobacteria; Ignavibacteriae)	6	76	52.5	N
S31	B (Verrucomicrobia; Lentisphaerae; Bacteroidetes)	7	100	39.6	N
S32	B (Actinobacteria; Firmicutes; Spirochaetes: Thermotogae)	8	100	61.6	1
S33	B (Spirochaetes)	3	99	51.8	N
S34	B (Firmicutes)	5	96	59.1	1
S35	B (Firmicutes)	6	89	70.3	4

Subfamily	Taxonomy	# of sequences	Bootstrap value	Within subfamily pairwise identity (%)	# of characterized proteins
S36	B (Proteobacteria; Planctomycetes) & A	4	75	37.3	N
S37	B (Proteobacteria)	4	100	49.0	N
S38	B (Bacteroidetes)	7	82	52.4	1
S39	B (Proteobacteria; Bacteroidetes; Acidobacteria; Verrucomicrobia; Lentisphaerae; Ignavibacteriae)	18	100	46.9	6
S40	B (Firmicutes; Proteobacteria)	23	98	60.9	18
S41	B (Firmicutes; Thermotogae; Dictyoglomi)	17	97	63.8	6
S42	B (Firmicutes)	6	100	89.8	4
S43	B (Actinobacteria)	6	100	45.6	1
S44	B (Cyanobacteria)	7	100	63.7	N
S45	B (Firmicutes)	16	95	58.7	10
S46	B (Proteobacteria)	4	100	56.4	N
S47	B (Firmicutes; Spirochaetes; Gemmatimonadetes)	8	89	61.7	6
S48	B (Actinobacteria)	4	100	71.7	N
S49	B (Fibrobacteres)	5	100	58.1	3
S50	haptophytes	3	100	80.1	N

The phylogenetic tree shows that the number of bacterial and fungal GH10 subfamilies is much higher than those of land plants, metazoa, and archaea, suggesting multiple gene duplication events in the former lineages (Figure 8; Table 11). Furthermore, the fact that some of the recovered subfamilies contain GH10 genes specific to organisms of certain phyla suggests extensive lineage specific losses within these subfamilies following duplication. The analysis also showed that fungal GH10 genes are more closely related to bacterial GH10 genes than those from other eukaryotes. In addition, subfamily 34 which contains GH10 sequences from the bacterial phylum Firmicutes is shown to be a well-supported sister group of the land Plants GH10 subfamilies. The close relationship between prokaryotic and eukaryotic GH10 sequences observed in the Maximum Likelihood phylogeny suggests that the divergence of GH10 genes preceded the appearance of Eukaryotic lineage. This suggestion is further supported by the fact that the phylogeny of GH10 family does not reflect the established taxonomic relationships. In addition to subfamilies containing members from the same Kingdom, the phylogenetic analysis also recovered several well-supported subfamilies comprising GH10 sequences from organisms of different Kingdoms. For instance, subfamily 4 contains both fungal and bacterial GH10 sequences. It was shown that while sequences within this subfamily share about 58% amino acid identity, they only show about 25% identity with other subfamilies. The only experimentally characterized fungal sequence of subfamily 4 is shown to be a tomatinase which hydrolyzes α-tomatine, a secondary anti-fungal metabolite produced by plants [113]. The development of this substrate specificity may be correlated to the ecological niches of the organisms as most of the members of this subfamily belong to plant pathogens. *Piromyces sp.*, an anaerobic fungus from the phylum Neocallimastigomycota, has an expansion in GH10 gene number. The 28 GH10

sequences from this organism are clustered into three subfamilies (Sub12-14). Subfamilies 13 and 14 hold exclusively *Piromyces sp.* GH10 sequences. On the other hand, subfamily 12 contains GH10 proteins from *Piromyces sp.*, bacteria, and a protozoon. Members of this subfamily share the same lifestyle as anaerobic organisms. Subfamily 2 is another multi-Kingdoms subfamily which contains fungal sequences and those from pathogenic oomycetes. It is worth mentioning that a close phylogenetic relationship was also observed between an oomycete and fungi in a recently published analysis of cytochrome b proteins [114]. In the analysis, the cytochrome b protein sequence of the oomycete *Pseudoperonsopora cubensis* is nested within a fungal cluster instead of grouping with other oomycetes. The cytochrome b amino acid sequence of *P. cubensis* is 91% identical to that of *Verticillium dahliae*, an ascomycete pathogen. The authors concluded that this oomycete acquired its cytochrome b gene from fungi through horizontal gene transfer. Contrary to the phylogeny observed in the cytochrome b protein tree, all ten GH10 xylanases from the two oomycetes (*Phytophthora capsici* and *Phytophthora ramorum*) cluster with fungal xylanases in my analysis (Figure 8). In addition, the fungal and the oomycetes GH10 sequences of this subfamily only share about 40% sequence identity. Based on these observations and that a highly complex set of CAZy homologs has been identified in the species of the genus *Phytophthora* [115], it is unlikely that oomycetes inherited GH10 xylanase genes from fungi through horizontal gene transfer as in the case of the cytochrome b gene. It should be mentioned that although the assigned subfamilies are well-supported by high bootstrap value, the deep level relationships between subfamilies only have moderate or poor support in the phylogenetic tree, hence preventing further inference of the evolution within the gene family.

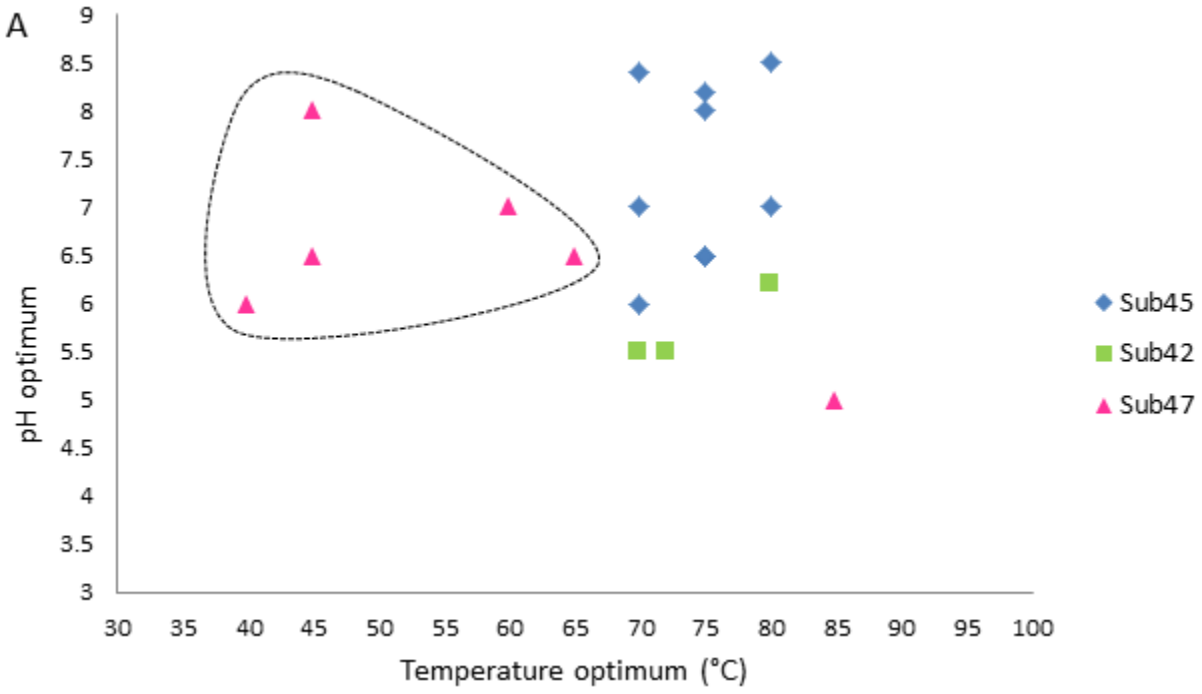
3.3 Functional diversity of GH10 proteins

As shown by phylogenetic analyses, proteins within glycoside hydrolase family 10 display great diversity in terms of amino acid sequences and can be clustered into well supported subfamilies (Figure 8). It was previously established that, within the same family, more closely related sequences also have similar function [74,116,117]. I have mapped experimental data from biochemically characterized GH10 proteins onto the phylogenetic tree. The purpose is to investigate the correlation between sequences clustering and function of proteins. The phylogenetic tree shows that among the 50 subfamilies, 24 contain sequences encoding biochemically characterized proteins. In total, 12 subfamilies have a sufficient amount of functional data to establish correlations between sequences clustering and biochemical properties (Table 12; Supplementary file 2). Figure 9 shows the pH and temperature optima of enzymes from different subfamilies and how they are clustered together. Only proteins for which both biochemical parameters have been determined were included in the analysis.

Table 12: Correlation between characterized GH10 proteins and subfamily clustering
This table shows the biochemical properties patterns of well-characterized subfamilies.

Subfamily	# of enzymes	pH optimum pattern	Temp. optimum pattern	Additional characteristic
Sub1	22	5.0-6.0	no correlation	open catalytic cleft
Sub2	8	4.5-6.0	70-85°C	narrow catalytic cleft
Sub12	6	< 6.0	no correlation	n.a
Sub26	3	6.5	> 90°C	n.a
Sub27	13	6.0-8.0	50-60°C	wide pH stability range
Sub35	4	5.0-6.5	60-70°C	n.a

Subfamily	# of enzymes	pH optimum pattern	Temp. optimum pattern	Additional characteristic
Sub39	6	6.0-7.5	< 50°C	narrow pH stability range
Sub40	18	6.0-8.0	no correlation	signal peptide-less high activity on small xylooligosaccharides
Sub41	6	5.5-6.5	no correlation	n.a
Sub42	5	5.5-6.2	70-80°C	narrow pH stability range
Sub45	11	6.0-8.0	70-80°C	wide pH stability range
Sub47	6	6.0-8.0	no correlation	high activity on polymeric substrates



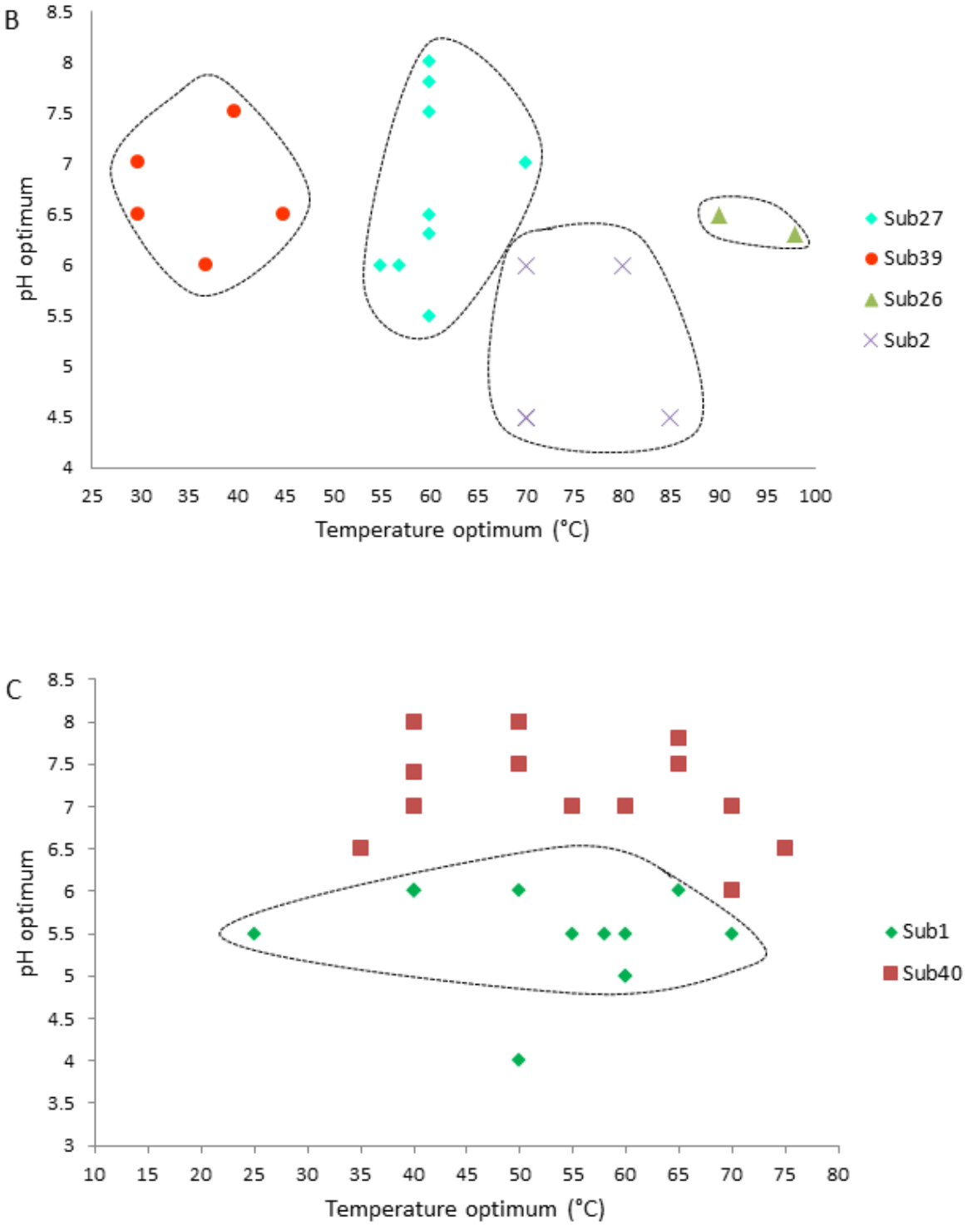


Figure 9: Temperature and pH optima of biochemically characterized GH10 enzymes

3.3.1 Experimentally characterized fungal GH10 genes

At the time of the analysis, there were 31 experimentally characterized fungal GH10 xylanases in *myco*CLAP [74]. In addition, five crystal structures from fungal species have been published and deposited in PDB. The analysis showed subfamilies 1 and 2 contain 22 and 8 experimentally characterized sequences, respectively. As for sequences with crystal structures, three belong to subfamily 1 whereas two are grouped in subfamily 2.

Figure 9C shows that the most frequent pH optimum for subfamily 1 characterized proteins is between pH 5.0 and 6.0 with one exception. On the other hand, these enzymes do not display a consistent temperature optimum pattern as the range is between 25°C and 70°C. As shown by Figure 9B, subfamily 2 xylanases have a wider optimal pH range which is between 4.0 and 6.0. Also, members of subfamily 2 have a higher optimal temperature range than their subfamily 1 counterparts, which is between 70°C to 85°C. The functional data showed that xylanases from different strains of the same species sharing 99% amino acid identity can display very different biochemical properties. In subfamily 1, xylanases from three different strains of *Penicillium chrysogenum* have been characterized. XYN10P_PENCH and XYN10A_PENCH are derived from strains Q176 and A3969.2, respectively and they both display an optimal temperature at 40°C [118,119]. However, XYN10B_PENCH isolated from the cold adaptive strain FS010, is most active at 25°C (Supplementary file 2) [119].

It was previously suggested that while the overall structures of all of the members of the GH10 family are well conserved, differences are often observed in the loop regions and the

length of the α-helices, which may account for the difference in substrate binding and specificity [120]. According to the phylogenetic analysis, three of the five fungal GH10 xylanases with crystal structures belong to subfamily 1 while two are members of subfamily 2. The 3D structures of these sequences were aligned using pyMOL [121]. All the sequences are folded into the typical $(\beta/\alpha)_8$ TIM-barrel and are relatively well aligned except at places where two extra loops are inserted. The proton donor and the nucleophile of xylanases are two conserved glutamic acid (E) residues located in the active site. Structures of subfamily 1 *Thermoacus aurantiacus* (PDB: 1gok) and subfamily 2 *Penicillium simplicissimum* (PDB: 4f8x) were used to highlight the structural differences. Figure 10A shows that α-helix 7 of *P. canescens* xylanase is longer and has an extended loop. Figure 10B demonstrates a second extra loop between α-helices 8 and 9 on this subfamily 2 xylanase. Both loops are found on the barrel top of the catalytic site of *P. canescens* xylanase and are in close proximity to each other, which suggests possible interactions between these two loops (Figure 10C). It seems that the insertion of the extra loop caused α-helix 8 of the subfamily 2 xylanase to partially shield the catalytic site. It has been shown that xylanases with these two loops display distinct degradation pattern from those without the loops. Two xylanases were characterized from *Myceliophthora thermophile* recently and it was proposed that they have different substrate specificities [122]. While XYN10A_MYCTH is more active on wheat arabinoxylan, a substrate highly substituted with arabinose (32%), the preferred substrate of XYN10C_MYCTH is oat spelt xylan, a more linear substrate with only 7% arabinose substitutions [122]. According to my phylogenetic tree, XYN10A_MYCTH belongs to subfamily 1 whereas XYN10C_MYCTH is a member of subfamily 2. The author concluded that XYN10A_MYCTH has a more open cleft because of the

absence of the two loops hence its ability to hydrolyze branched xylooligosaccharides more efficiently. In contrast, the presence of two extra loops causes XYN10C_MYCTH to have a more closed cleft that leads to its preference for linear xylans [122]. Multiple sequence alignment of all of GH10 fungal xylanases demonstrated that these two extra loops are found in all subfamily 2 sequences but absent from subfamily 1 counterparts. The clustering pattern is also supported by the study of *T. aurantiacus* xylanase. It was shown that this xylanase, which belongs to subfamily 1 according to my analysis, has fourfold more activity on a xylotriose substituted with arabinose than undecorated xylotriose [123]. Another interesting observation is that although α-helix 8 of sequences from subfamilies 1 and 2 are well aligned at their primary sequence level, their crystal structures cannot be superimposed (Figure 10C). Based on the crystal structure of *T. aurantiacus*, it was discovered that the highly conserved tryptophan (W) located on α-helix 8 is more disordered in subfamily 1 xylanases. It was shown that this tryptophan residue and two other adjacent amino acids, arginine and glutamate, have two conformations (A and B) in the native form of the enzyme. All three residues are located at the catalytic site. On the other hand, the same tryptophan within the subfamily 2 xylanases is more ordered, adopting only one conformation. It was proposed that the extra residues on the inserted loop of the subfamily 2 sequences form additional hydrophobic/aromatic interactions with the tryptophan, thus making it less flexible. It was also shown that once the *T. aurantiacus* xylanase, belonging to subfamily 1, forms a complex with a xylooligosaccharide, the B conformation of the three residues disappeared and the disorder of the tryptophan is reduced. The authors suspected that the mobility of the tryptophan contributes to the substrate specificity of the enzyme. They speculated that subfamily 1 xylanases prefer longer xylooligosaccharides for the

stabilization of this disordered tryptophan. On the other hand, the more rigid catalytic site of the subfamily 2 xylanases makes them better at cleaving shorter xylooligosaccharides [124].

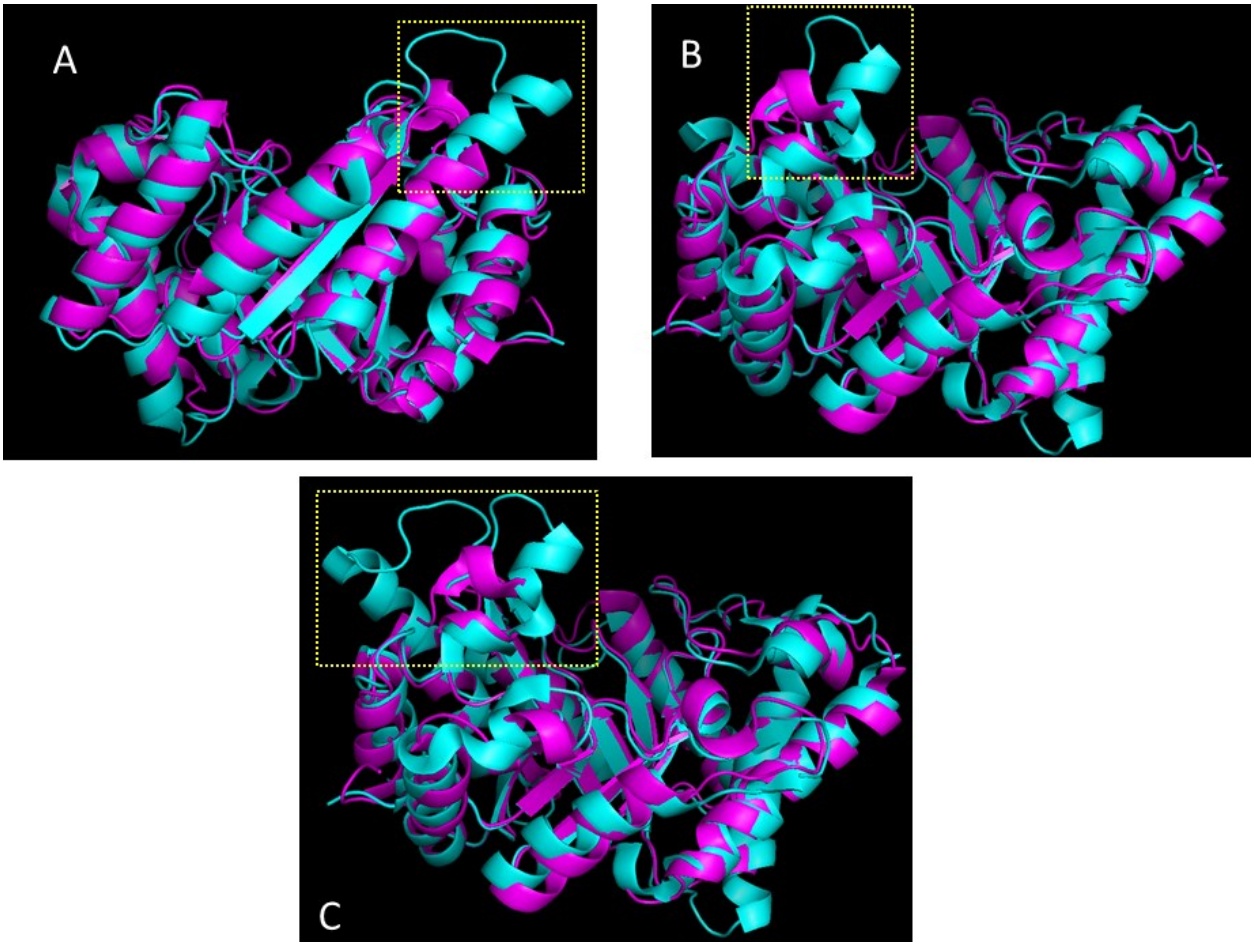


Figure 10: Superposition of fungal GH10 xylanase crystal structures

The crystal structures of subfamily 1 *Thermoacus aurantiacus* (magenta) and subfamily 2 *Penicillium canescens* (cyan).Panel A shows the extended loop and α-helix 7 of *P. canescens*. Panel B shows the loop inserted between alpha-helices 8 and 9 of *P. canescens*. Panel C shows the two extra loops found in *P. canescens*. are in close proximity and above the catalytic site and the different conformation of α-helix 8.

In summary, fungal subfamilies 1 and 2 GH10 xylanases show significant structural differences which are believed to be correlated with their substrate specificites [120,122,124]. To validate this correlation, the substrate specificity of experimentally characterized fungal GH10 xylanases was mapped onto the tree. Table 13 shows the experimentally characterized fungal GH10 xylanases with available data on substrate preference. For subfamily 1, three xylanases display higher specific activity towards the more branched wheat arabinoxylan which is in accordance with the proposed structure-function relationship. However, one sequence (XYN10B_PENCH) shows preference towards the more linear birchwood and oat spelt xylan. As for subfamily 2, most of the characterized xylanases have not been tested on more branched substrates hence they cannot be used to validate the hypothesis that enzymes from this subfamily prefer more linear xylan. Only one subfamily 2 xylanase (XYN10D_PENFN) was tested on both wheat arabinoxylan and birchwood xylan. The assay showed that the enzyme is more active on wheat arabinoxylan, which disagrees with the aforementioned prediction pattern. From the experimentally characterized fungal GH10 xylanases collected from the *myco*CLAP database, it seems that one cannot validate the proposed prediction pattern as some of the data disagree with it and only a limited amount of information is available. In addition, one should keep in mind that all of these xylanases are assayed under different experimental conditions which may cause discrepancies in the results. To confidently confirm the substrate specificites of subfamilies 1 and 2 xylanases, more experimental characterization has to be done and it is necessary to assay enzymes of interest under the same assay conditions.

Table 13: Biochemically characterized GH10 proteins in fungi

This table lists the experimentally characterized fungal GH10 xylanases from the *myco*CLAP database with available data on pH optimum, temperature optimum or substrate preference. The subfamily is assigned according to the phylogenetic tree. The substrates are: birchwood xylan (BiWX), beechwood xylan (BeWX), oat spelt xylan (OSX), and wheat arabinoxylan (WAX).

***myco*CLAP Entry Name**	**Subfamily**	**Host**	**Substrate preference**	**Reference**
XYN10B_PENCH	sub1	*E. coli*	BiWX>OSX>WAX	[119]
XYN10A_PENPU	sub1	native	WAX≈OSX>BiWX	[125]
XYN10A_MYCTH	sub1	*M. thermophila*	WAX>BeWX>BiWX≈OSX	[122]
XYN10C_GIBZE	sub1	*E .coli*	WAX>OSX>BiWX	[126]
XYN10P_PENCH	sub1	native	OSX≈BiWX	[118]
XYN10D_PENFN	sub2	native	WAX>BiWX	[127]
XYN10A_PHACH	sub2	*A. niger*	OSX>BeWX≈BiWX	[128]
XYN10C_PHACH	sub2	*A. niger*	OSX>BeWX≈BiWX	[128]
XYN10C_MYCTH	sub2	*M. thermophila*	OSX>WAX	[122]
XYN10B_AURPU	sub2	native	OSX≈BiWX	[129]
XYN10C_BISSP	sub2	*P. pastoris*	OSX>BiWX	[130]

One experimentally characterized GH10 sequence from *Fusarium oxysporum f. sp. Lycopersici* is placed in subfamily 4. The biochemical characterization showed that this sequence does not have xylanase activity but hydrolyzes α-tomatine, an antifungal agent produced by plants [113]. Multiple sequence alignment confirmed that sequences of this subfamily have conserved motifs that are unique, which may contribute to the development of their substrate specificity towards α-tomatine. The identification of a subfamily with a new function further supports the idea that phylogenetic trees can be used to predict the substrate specificites of the enzymes within the same family.

3.3.2 Experimentally characterized bacterial GH10 genes

A set of bacterial GH10 enzymes were manually curated using the criteria described by Murphy et al. [74]. In total, there are 103 experimentally characterized bacterial GH10 proteins as of January 2014 (supplementary file 2). Among them, 17 sequences also have available crystal structures deposited in PDB. Two other uncharacterized xylanases also have available crystal structures. The phylogenetic tree shows that 20 subfamilies contain experimentally characterized bacterial sequences (Figure 8). Furthermore, 15 out of the 17 sequences with crystal structures are distributed across 8 subfamilies. The remaining two are unclustered (Supplementary file 3).

The functional data of these biochemically characterized bacterial GH10 proteins were mapped onto the phylogenetic tree. Of the 20 characterized subfamilies, 10 have a sufficient amount of functional data to demonstrate correlations between sequence clustering and biochemical properties (Table 12). For instance, Figure 9B demonstrates that three bacterial subfamilies display distinct temperature optima ranges. While proteins of subfamily 29 are optimally active at temperatures lower than 50°C, members of subfamily 26 have temperature optima above 85°C. In addition, enzymes of subfamily 27 have a narrow temperature optimum range that is between 55 and 65°C. Figure 9A shows that all, except one, of the experimentally characterized proteins of subfamily 47 are optimally active at a pH range between 6.0 and 8.0. In the same figure, it is observed that the clustering of subfamily 45 is not correlated with the pH optima of its members but instead with the temperature optima. All of the characterized proteins are optimally active at temperatures between 70 and 80°C. Other than a correlation to pH and temperature optimum, I observed that some subfamilies are clustered according to substrate

specificity as well. For example, biochemical assays demonstrated that while sequences of subfamily 40 show high activity on small xylooligosaccharides, GH10 proteins of subfamily 47 are highly active on polymeric substrates. In the following sections, well-characterized sequences from different subfamilies were compared in more depth to explore how differences in structures and amino acid sequence correlate with the formation of the subfamilies.

3.3.2.1 Bacterial subfamilies 32 and 40: Signal peptide-less xylanases

Bacterial subfamily 32 contains one characterized xylanase (XynA4-2) which is from *Alicyclobacillus sp. A4*. Its characterization showed that the protein is intracellular, which is consistent with the lack of a predicted signal peptide [131]. Furthermore, SignalP analysis indicated that all sequences found within this subfamily lack a predicted signal peptide [132]. In addition, bacterial subfamily 40 is another subfamily that contains exclusively signal peptide-less xylanases. Multiple sequence alignment of representative subfamilies 32 and 40 xylanases showed that sequences from these two subfamilies are very dissimilar and each subfamily has well conserved unique motifs (Figure 11). In addition to the aforementioned XynA4-2 belonging to subfamily 32, *Alicyclobacillus sp. A4* also contains a second xylanase XynA4 which clusters within subfamily 40. These two paralogs share less than 20% amino acid sequence identity and their characterization showed distinct properties. While XynA4-2 hydrolyzes xylan mostly to xylose (92.7%) with a minor amount of xylobiose (7.3%), only about half of the hydrolysis products generated by XynA4 using xylan as the substrate is xylose (51.5%) with 34.3%, 7.53%, and 6.65% of xylobiose, xylotriose and xylotetraose, respectively [133]. The MSA profile of the

two subfamilies showed that subfamily 32 xylanases have truncated N-terminai lacking α-helices 0 and 1 as well as β-sheets 1 and 2 (Figure 11). Also, subfamily 32 xylanases contain additional inserted regions. One of these regions is between β-sheet 4 and α-helix 4 and another region is located between β-sheet 7 and α-helix 7. Both regions are in proximity of the proton donor and the nucleophile, respectively, which suggests that these residues may play a role in catalysis. Subfamily 40 xylanases also have inserted regions compared to their subfamily 32 counterparts. Currently, the crystal structure of subfamily 40 IXT6 (PDB: 2q8x) from *Geobacillus sterothermophilus* is available [134]. On the other hand, XynA4-2 is the only experimentally characterized xylanase of subfamily 32 and no crystal structure is available for this subfamily. Characterization of other enzymes from this subfamily will determine whether they display similar exo-acting properties as XynA4-2. The determination of crystal structures of proteins belonging to subfamily 32 will allow comparison with IXT6 to evaluate how these insertion/deletion regions affect the hydrolysis mechanism of the enzymes.

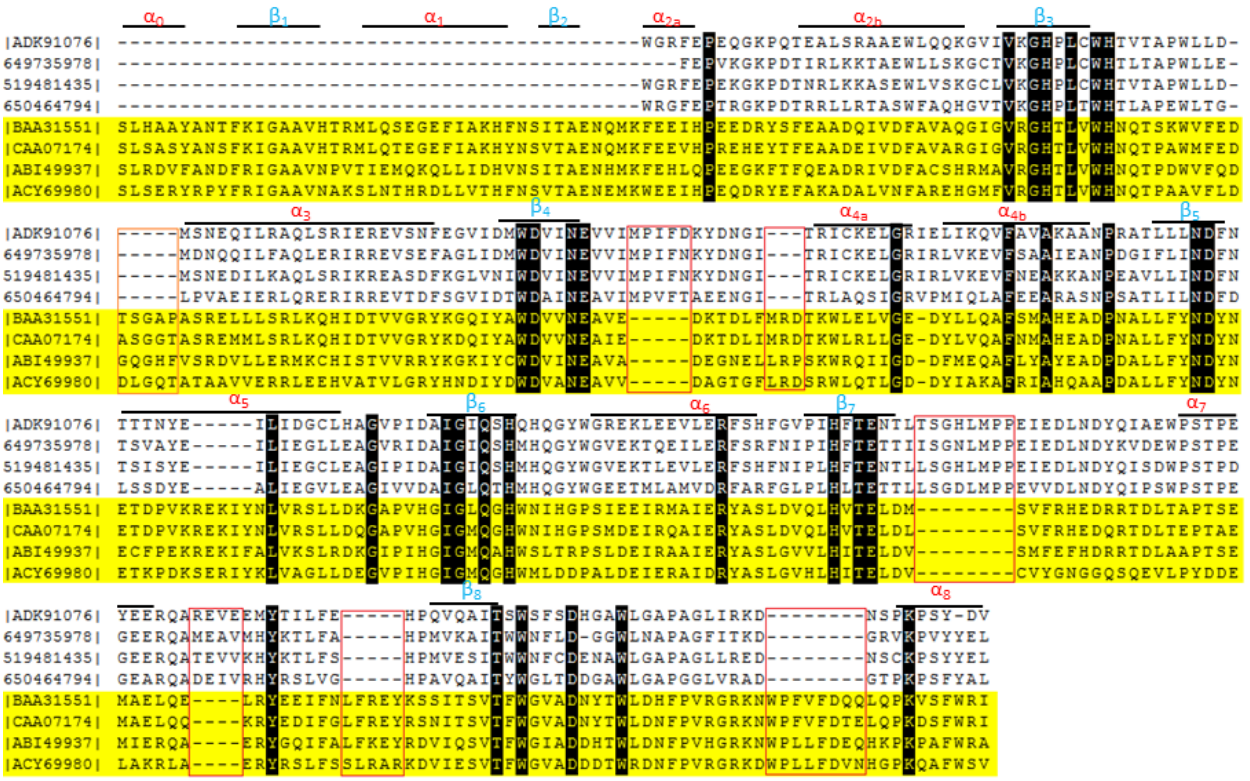
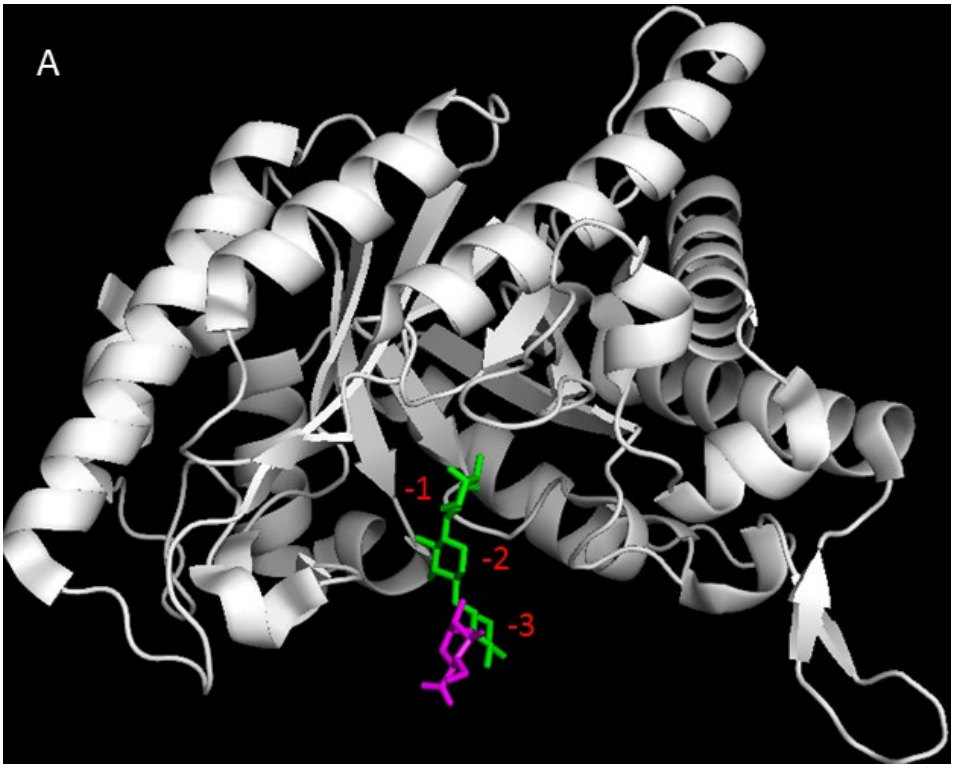


Figure 11: Comparison of bacterial subfamilies 32 and 40 signal peptide-less xylanases
This figure shows the alignment of subfamilies 32 and 40 representative xylanases. Secondary structure formation is predicted based on the crystal structure of *Geobacillus sterothermophilus* (PDB: 2q8x). Subfamily 40 xylanases are colored in yellow. Residues highlighted in black are 100% conserved. The major insertion/deletion regions are boxed.

3.3.2.2 Bacterial subfamilies 40 and 47: Structural differences at the substrate recognition area

The phylogenetic analysis showed that bacterial subfamily 40 contains exclusively signal peptide-less xylanases. Experimental characterization data indicated that these xylanases have high activity on small substrates which is consistent with their intracellular location as shorter xylooligosaccharides are generated by extracellular xylanases and subsequently imported into the cells [135,136]. On the other hand, experimental assays showed that xylanases from subfamily

47 have higher specificity on polymeric xylan substrates [137–139]. It was suggested that subfamily 47 xylanases are capable of utilizing polymeric substrates due to the presence of S-layer homology domains (SLH) which allow them to be anchored on the cell surface [138]. However, these SLH domains are not universal within subfamily 47, which suggests that the localization of the protein is not the only factor that contributes to the difference in substrate specificity of these two subfamilies. The determination of crystal structure of *Panebacillus sp.* XynA, a member of subfamily 47, revealed that the enzyme has a relatively open substrate recognition area (negative binding subsites), which allows it to accommodate branched xylan. The crystal structure of the enzyme complex with aldotetrauronic acid ($MeGX_3$), a xylotriose substituted with a 4-*O*-methyl-α-D-glucuronic acid, showed that the ligand is bound to the subsites -3 to -1 of the enzyme with the glucuronic side chain attached to the xylose residue that occupied subsite -3 (PDB: 3rdk) (Figure 12A). It was shown that the substrate only makes direct contact with subsites -1 and -2 of the enzyme [140]. Furthermore, the structure of subfamily 40 *Panebacillus barcinonenesis* XynB (PDB: 3emc) is superimposed onto its subfamily 47 counterpart [141]. The superposition showed that the glycone region of XynB is significantly narrower due to the presence of an inserted loop (amino acids 302 to 306). Aromatic residues Phe303 and Arg306 seem to cause steric hindrance at subsite -2 and -3, respectively, which is consistent with the preference of XynB for small xylooligosaccharides (Figure 12B). This loop is found in all the bacterial subfamily 40 xylanases.


A
-1
-2
-3

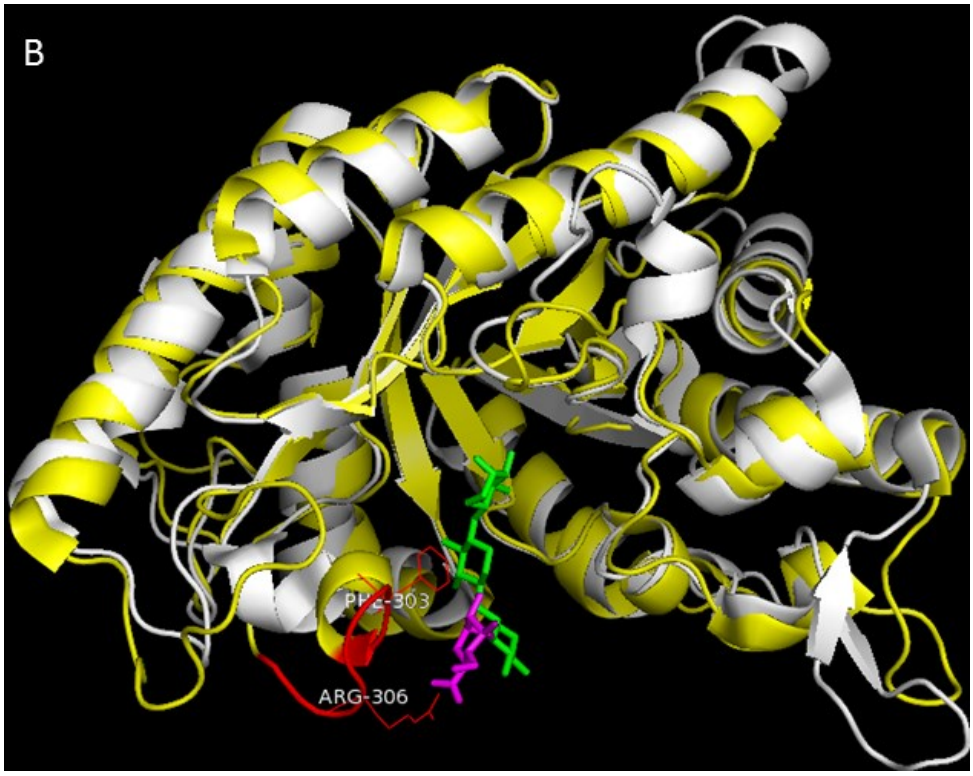

B
ARG-306

Figure 12: Comparison of bacterial subfamilies 40 and 47 xylanases

(A) Crystal structure of *Panebacillus sp.* XynA complex with $MeGX_3$ (PDB: 3rdk). The xylotriose (green) is bound to subsites -3 to -1. The 4-*O*-methyl-α-D-glucuronic acid side chain (magenta) is attached to the xylose ring at subsite -3. (B) Superposition of subfamily 47 *Panebacillus sp.* XynA complex with $MeGX_3$ (white) and subfamily 40 *Panebacillus barcinonenesis* XynB PDB: 3emc (yellow). The inserted loop (residues 302-308) of XynB is colored in red. The aromatic residues Phe303 and Arg306 are shown in red.

3.3.2.3 Bacterial subfamilies 26 and 39: Low temperature-active vs Hyperthermophilic xylanases

Bacteria subfamily 39 contains four experimentally characterized xylanases. These xylanases have optimal temperatures from 30°C to 45°C as well as low thermostability [142–146]. Contrary to subfamily 39 xylanases that are active at low temperatures, the experimentally characterized xylanases of subfamily 26 can thrive at an optimum temperature of 90°C [147–149]. Xylanases of these two subfamilies share about 30% amino acid identity. It has been proposed that various parameters such as the number of salt bridges and hydrogen bonds as well as amino acid composition affect the thermostability of the enzyme [150,151]. Structures from subfamily 39 *Cellvibrio mixtus* CmXyn10B (PDB: 2cnc) and subfamily 26 *Thermotogae maritime* TmxB (PDB: 1vbu) were used to evaluate the differences between the low temperature active and the hyperthermophilic xylanases [152,153]. It has been suggested that the increasing number of easily decomposed amino acids such as serine and threonine as well as thermolabile asparagine and glutamine can decrease the thermostability of the proteins. However, Table 14 shows that CmXyn10B and TmxB have very similar composition with regard to these residues, suggesting that they do not contribute greatly to the thermostability of TmxB. The number of salt

bridges and the number of hydrogen bonds of the proteins are predicted using VMD (http://www.ks.edu/Research/vmd/) [154] and USFC Chimera (www.cgl.ucsf.edu/chimera) [155]. The results showed that the low temperature active CmXyn10B possess fewer hydrogen bonds and salt bridges, which are responsible for stabilizing the outer helices and loops regions of the protein [152]. In addition to the predicted single salt bridges, five triad bridges were also identified in TmxB [152]. Among them, two were also found in CmXyn10B. Loops as well as N and C termini are believed to be the regions where denaturation most likely to begin. One of the unique triad bridges of TmxB is found in the C-terminus which might contribute to the stability of the protein. The C-terminus of CmXyn10B seems much more vulnerable to denaturation due to the absence of this triad salt bridge as well as the presence of a loop (Figure 13B). The length of secondary structural elements is believed to be positively correlated to its thermostability which is consistent with our data showing 72% of the residues on TmxB are involved in secondaru structure whereas as only 64% of the residues in CmXyn10B form α-helices and β-sheets (Table 14). Tertiary structural alignment of these two xylanase showed that the hyperthermophilic TmxB has a more compacted structure and contains fewer loop regions (Figure 13B). As shown by the alignment of representative subfamily 26 and 39 sequences, the low temperature active xylanases have an insertion of 24 amino acids forming two short helices and two long loops, which may contribute to the destabilization of the protein (Figure 13A).

Table 14: Comparison of low temperature active and hyperthermophilc bacterial xylanases

This table compares potential parameters affecting the optimum temperature of the enzyme between low temperature active CmXyn10B and hyperthermophile TmxB.

	CmXyn10B (PDB: 2cnc)	TmxB (PDB: 1vbu)
Optimum Temperature	40°C [142]	90°C [149]
Number of salt bridges	16	24
Number of hydrogen bonds	364	706
Percentage of aromatic residues (FWYH)	12.9%	15.9%
Percentage of easily decomposed residues (ST)	9.2%	7.1%
Percentage of thermolabile residues (NQ)	7.4%	7.8%
Percentage of residues involved in secondary structure formation	64.5%	72.1%

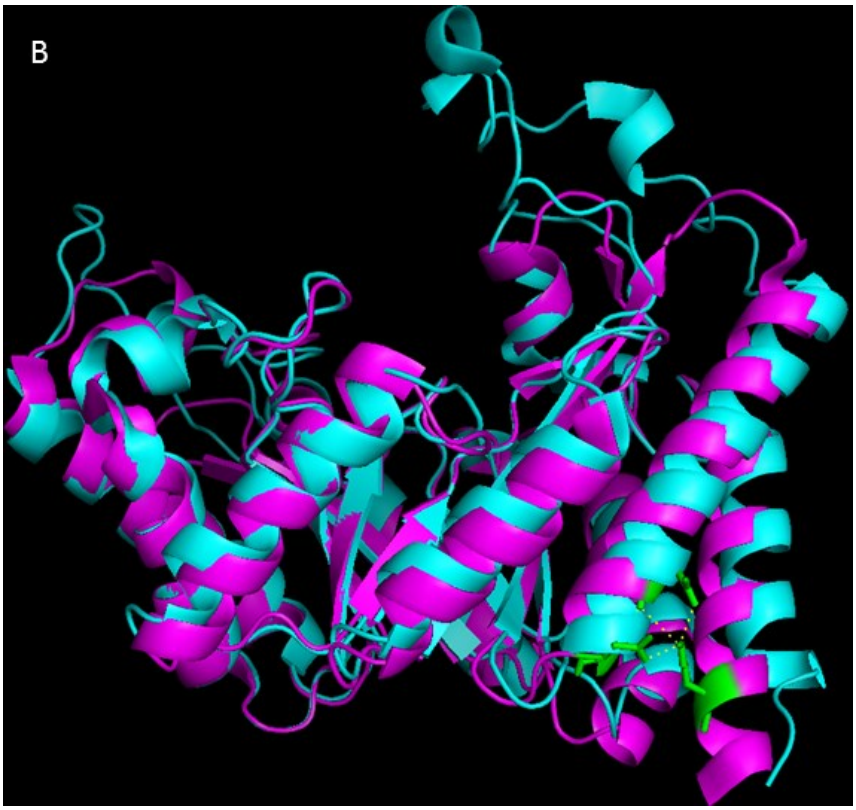


Figure 13: Sequences and structures alignment of bacterial subfamilies 26 and 39 xylanases
(A) Multiple sequences alignment of representative bacterial subfamilies 26 and 39 xylanases. Secondary structure formation is predicted based on the crystal structure of *Cellvibrio mixtus* (PDB: 2cnc) and *Thermotogae maritima* (PDB: 1vbu). Residues involved in the formation of α-helices and β-sheets are colored in red and yellow, respectively. The amino acid insertion of subfamily 39 xylanases is boxed. (B) Superposition of the crystal structure of *C. mixtus* (cyan) and *T. maritima* (magenta). The residues that form the triad salt bridge which stabilizes the C-terminus of TmxB are shown in green.

3.3.2.4 Bacterial subfamilies 42 and 45: Alkaline-active vs Alkaline-inactive xylanases

Characterized bacterial xylanases from subfamily 45 are encoded by alkalophiles of the phylum Firmicutes and are stable at relatively elevated temperature and pH. The optimum

temperature of these enzymes is between 70 and 80°C. The optimum pH ranges from 6.0 to 8.0 and the enzymes remain stable up to a pH between 10.0 and 12.0 [156–163]. Sequences from subfamily 42 are also from Firmicutes and the characterized xylanases display a similar temperature optimal range. However, these enzymes are not alkaline active as they are optimally active at pHs less than 6.0 and only remain stable up to pH 7.0 [164–166]. It has been reported that amino acid composition affects the adaptation of alkaline enzymes to high pH. For instance, it has been shown that the negatively charged residues glutamate and aspartate occur more frequently within alkaline xylanase and mostly exhibit on the surface of the protein. Arginine is another residue that is believed to be involved in the stabilization of the enzyme as its high pKa allows the formation of hydrogen bonds at high pH. On the other hand, the alkali-labile residue asparagine is less frequently found in alkaline active enzymes [167]. The amino acid compositions of subfamilies 42 and 45 were analyzed to validate these previous findings. As shown in Table 15, xylanases of subfamily 45 do contain more arginine and fewer asparagine residues compared to their non-alkaline active counterparts. Also, most of the subfamily 45 xylanases display a higher percentage of acidic residues than the non-alkaline active enzymes. In addition, when comparing the crystal structures of subfamily 42 *Thermoanaerobacterium saccharolyticum* TsXylA (PDB: 3w24) to subfamily 45 *Bacillus halodurans* Xyn10A (PDB: 1vbu), the negatively charged residues of the latter are mostly found on the surface of the enzyme [167,168].

Table 15: Comparison of subfamilies 42 and 45 xylanases
This table compares the amino acid composition of subfamily 45 alkaline active xylanases and subfamily 42 alkaline-inactive xylanases.

	Subfamily 42	Subfamily 45
Acidic residue (ASP & GLU)	12.3-13.3% (12.8%)	12.5-17.5% (15.1%)
Arginine	1.6-2.6% (1.9%)	3.2-6.9% (4.5%)
Asparagine	7.6-9.3% (8.5%)	4.1-5.7% (4.9%)

3.3.2.5 Bacterial Subfamilies 42 and 45: Structural differences at the product release area

In addition to different pH stability, experimentally characterized sequences from subfamily 42 and 45 also display different modes of action on xylooligosaccharides and branched xylan. Experimental assays showed that xylooligosaccharide products obtained from the hydrolysis of heteroxylan using subfamily 45 xylanases were mainly xyloses, xylobiose, and xylotriose [111,116]. On the other hand, xylose was not detected from subfamily 42 xylanase hydrolysis when the same substrate was used [168,169]. Through xylan-binding subsite mapping, it was shown that XT6 from *Geobacillus stearothermophilus*, a subfamily 45 xylanase can hydrolyze xylotriose into xylose and xylobiose whereas subfamily 42 TsXylA from *T. saccharolyticum* requires a minimum length of the xylopentose chain for cleavage [168,170]. Structures of XT6 complex with xylopentose (PDB: 1r87) and TsXylA complex with xylotriose (PDB: 3w26) were used to analyze the binding preference of these two subfamilies. It was shown that the xylotriose occupies subsite -3 to -1 in TsXylA. On the other hand, the complex of XT6 and xylopentose display a xylotriose at subsite -3 to -1 and a xylobiose at +1 and +2, illustrating

that the xylopentose is cleaved at the subsites -1 and +1. Figure 14A shows that the xylose residues at negative subsites of the two enzymes aligned perfectly, suggesting that the difference in their binding preference must be caused by their interaction with xylooligosaccharides at the positive subsites. Figure 14B shows that an arginine (Arg238) residue of subfamily 45 XT6 forms hydrogen bonds with the xylose residues at the +1 and +2 subsites. No hydrogen bonds is found at the +1 and +2 subsites of subfamily 42 TsXylA as the equivalent Arg238 is missing from this xylanase. In addition, XT6 has an extra loop close to the +2 subsite. A tryptophan (Trp273) located in this loops forms a stacking interaction with the xylose residue at +2. It was shown that Trp241 forms a stacking interaction with subsite +3 [171]. At this equivalent position, subfamily 42 TsXylA has a serine (S228) which is much smaller than the aromatic tryptophan, hence it cannot interact with the xylose residues at +3 subsite through stacking interaction (Figure 14B). Multiple sequence alignment showed that all subfamily 45 sequences have this inserted loop at the proximity of subsite +2 with the tryptophan residue being conserved whereas all subfamily 42 xylanases lack this aromatic residue. In addition, the substitution of Trp241 of the subfamily 45 XT6 by a serine residue is universal within subfamily 42 xylanases. Finally, the arginine that forms hydrogen bonds with xylose residues at +1 and +2 subsites of subfamily 45 XT6 is missing from all subfamily 42 xylanases. From this analysis, it can be concluded that XT6 is able to cleave xylooligosaccharides as short as xylotriose because the enzyme has a higher binder affinity at its aglycone region. This high binding affinity is provided by two tryptophan residues through stacking interaction as well as hydrogen bonding through an arginine. On the other hand, subfamily 42 xylanases have a weaker binding affinity at their positive subsites. Therefore, at least two positive subsites have to be occupied for the proper

binding of the substrate, which is consistent with the fact that no xylose is observed in the hydrolysis product.

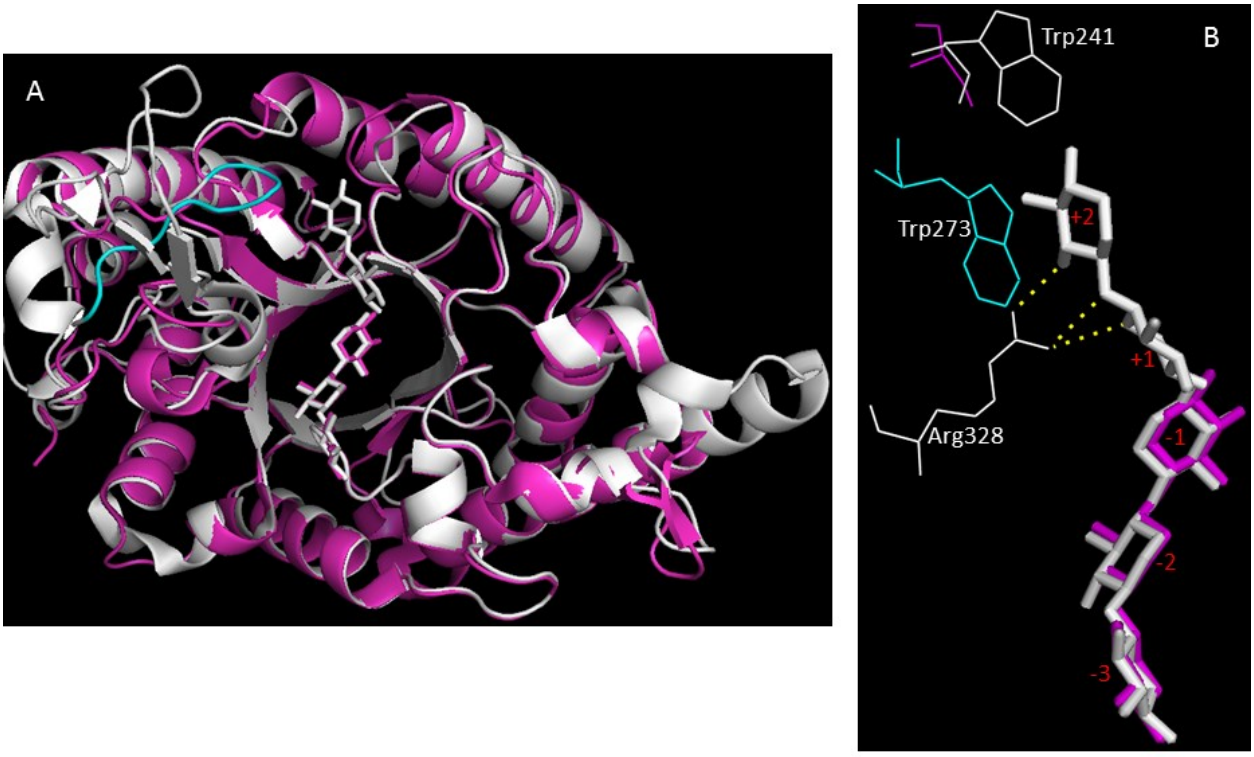


Figure 14: Xylooligosaccharide binding preference of bacterial subfamilies 42 and 45 xylanases

(A) Superposition of subfamily 42 TsXylA complex with xylotriose PDB: 3w26 (magenta) and subfamily 45 XT6 complex with xylopentose PDB: 1r87 (white). The extra loop of XT6 at the proximity of subsite +2 is colored in cyan. (B) View of xylotriose (magenta) occupying subsites -3 to -1 of TsXylA and xylopentose positioned at subsite -3 to +2 of XT6 (white). Arg328 forms hydrogen bonds (yellow dashes) with xylose residues at +1 and +2 subsites. Trp273 (cyan) located at the extra loop of XT6 is shown to be at close proximity to the xylose ring at +2 subsite. Trp241 (white) of XT6 is replaced by Ser278 (magenta) in TsXylA.

3.3.3 Experimentally characterized GH10 genes in other Kingdoms

The number of experimentally characterized GH10 genes in other Kingdoms is very low as compared to those in fungi and bacteria. None of the GH10 proteins from Archaea have been characterized. GH10 sequences of *Ampullaria crossean* and *Hypothenemus hampei* are the only two metazoan xylanases that have been experimentally characterized. *A. crossean* belongs to the phylum Mollusca and GH10 sequences encoded by this organism are nested within the Metazoan GH10 subfamily and its closest orthologs are those from *Lottia gigantean*, also from the phylum Mollusca. On the other hand, *H. hampei*, an insect pathogen of coffee, is from the phylum Arthropoda. The GH10 sequence encoded by this organism remains unclustered. Finally, an experimentally characterized GH10 xylanase from the rumen anaerobic protozoan *Polyplastron multivesiculatum* was also curated. This sequence is found within subfamily 12 that contains GH10 proteins from the anaerobic fungus *Piromyces sp.* as well as anaerobic bacteria.

3.4 GH10 sequences conservation analysis

It is well known that globally conserved residues within a protein family are crucial for the structure as well as the function of the enzyme. In addition, mutations accumulated during evolution may result in the development of conformational changes and/or new functions within the family. Therefore it is crucial to recognize subfamily with mutation at these positions. Amino acid conservation analyses were performed to identify globally conserved amino acids as well as discriminating residues that contribute to the divergence of subfamilies.

3.4.1 Globally conserved amino acids

The absolute conservation scores were calculated for each position of the alignment to identify conserved amino acids in GH10 sequences [80]. A position in the alignment is considered globally conserved when the conservation scores exceed 0.90. Using this criterion, 23 amino acids that are highly conserved across all Kingdoms were identified. These residues must be crucial for the function of the GH10 as 12 of these conserved residues are involved in the formation of β-sheets that surround the catalytic cleft, where the substrate binds and gets cleaved (Table 16).

Table 16: Globally conserved amino acids of GH10 family identified using the absolute method

Alignment position	Globally conserved Amino Acid	Absolute conservation score	Amino Acid location
6	G	0.90	β_1
42	N	0.92	loop
45	K	0.97	α_{2a}
79	H	0.98	β_3
83	W	0.97	loop
161	D	0.96	β_4
162	V	0.98	β_4
164	N	0.98	β_4
165	E (proton donor)	0.98	β_4
237	A	0.93	α_4
252	L	0.94	β_5
255	N	0.99	β_5
304	G	0.97	β_6
308	H	0.97	loop

Alignment position	Globally conserved Amino Acid	Absolute conservation score	Amino Acid location
345	T	0.90	β_7
346	E (nucleophile)	1.0	β_7
348	D	0.95	β_7
413	W	0.97	loop

In addition to the absolute method, hydrophobicity and polarity conservation methods were also used to identify amino acids that are conserved in terms of hydrophobicity and/or polarity. I identified 36 such amino acids. These residues are found on both α-helices as well as β-sheets which reflect their importance in the function of the protein (Table 17).

Table 17: Globally conserved amino acids of GH10 family identified using the hydrophobicity and polarity methods

Alignment position	Globally conserved Amino Acid	Conservation score	Amino Acid location
29	non-polar	0.91	α_1
40	non-polar	0.91	β_2
46	non-polar	0.97	α_{2a}
63	hydrophilic	0.91	α_{2b}
76	non-polar hydrophobic	0.96 0.93	β_3
77	polar-charged	0.91	β_3
78	non-polar	0.92	β_3
81	non-polar hydrophobic	0.91 0.96	β_3

Alignment position	Globally conserved Amino Acid	Conservation score	Amino Acid location
82	non-polar hydrophobic	0.93 0.92	β_3
91	non-polar hydrophobic	0.92 0.90	α_{3a}
92	non-polar hydrophobic	0.97 0.97	α_{3a}
128	non-polar	0.91	α_{3b}
132	non-polar hydrophobic	0.92 0.90	α_{3b}
136	non-polar hydrophobic	0.95 0.98	α_{3b}
140	non-polar	0.96	α_{3b}
157	hydrophobic	0.97	loop
160	non-polar hydrophobic	0.91 0.91	β_4
167	non-polar hydrophobic	0.98 0.97	loop
233	non-polar	0.95	α_4
234	non-polar hydrophobic	0.91 0.91	α_4
256	polar-charged hydrophilic	1.0 1.0	β_5
280	non-polar hydrophobic	0.99 0.93	α_5
283	non-polar hydrophobic	0.99 0.99	α_5
297	non-polar hydrophobic	0.95 0.92	loop
302	non-polar neutral	0.92 0.91	β_6

Alignment position	Globally conserved Amino Acid	Conservation score	Amino Acid location
303	non-polar hydrophobic	0.98 0.98	β_6
306	hydrophilic	0.98	loop
322	non-polar hydrophobic	0.98 0.97	α_6
342	non-polar hydrophobic	0.92 0.96	β_7
344	non-polar hydrophobic	0.96 0.95	β_7
345	Polar-uncharged neutral	0.98 1.0	β_7
347	non-polar hydrophobic	0.96 0.95	β_7
349	non-polar hydrophobic	0.97 0.97	β_7
397	non-polar	0.90	α_7
407	hydrophobic	0.97	β_8
410	non-polar hydrophobic	0.98 0.98	β_8

3.4.2 Subfamily discriminating residues

In addition to amino acids that are conserved in the majority of GH10 members, it is also important to identify subfamily discriminating residues that contribute to the divergence of the subfamilies. These residues may be responsible for the development of properties that are specific to a subfamily.

Absolute conservation scores were re-calculated for each subfamily separately to obtain a subfamily-conservation score ($A_{subfamily}$). Only residues with all of the subfamily-conservation scores as well as absolute conservation score across the whole family ($A_{conservation}$) that exceed 60% were used for discrimination analysis. Using these criteria, five amino acids were identified as discriminating residues for one or more subfamilies (Table 18). Three residues are found in the list of highly conserved amino acids identified previously. This result suggests that while these amino acids are conserved in most of the GH10 member, some subfamilies have mutations at these positions that are crucial for the function of the protein. For instance, it was previously established that a glutamic acid (E165) located on β-sheet 4 acts as the proton donor of the enzymes [38]. This amino acid is conserved across all subfamilies except for members from subfamily 36 which have histidine (H165) at this position. Since none of the sequences from this subfamily have experimental data, it is unknown how a mutation at this position may affect the function of the enzymes. Identification of this discriminating residue in the members of subfamily 36 makes them interesting targets for further experimental characterization.

Table 18: Subfamily discriminating residues according to the absolute conservation method

Columns from left to right: alignment position of the amino acid; residue used by the majority sequences of the whole family with the conservation level in brackets; discriminated subfamily and amino acid used in the discriminated subfamily with the subfamily conservation level shown in brackets.

Alignment position	Whole family consensus ($A_{conservation}$)	Discriminated subfamily	Residue used ($A_{subfamily}$)
83	W (0.97)	Sub4 (Fungi & Bacteria)	S (1.0)
161	D (0.96)	Sub16 (Archaea) Sub30 (Bacteria)	E (0.67) E (0.67)

Alignment position	Whole family consensus ($A_{conservation}$)	Discriminated subfamily	Residue used ($A_{subfamily}$)
165	E (0.98)	Sub36 (Bacteria)	H (1.0)
304	G (0.97)	Sub4 (Fungi & Bacteria)	A (0.91)
306	Q (0.90)	Sub17 (Land plants) Sub18 (Land plants) Sub36 (Bacteria & Archaea) Sub48 (Bacteria)	E (1.0) E (0.71) M (1.0) E (1.0)

I also examined the subfamily-discriminating amino acids in terms of hydrophobicity and polarity. The hydrophobicity conservation score and the polarity conservation score were re-calculated for each subfamily separately to obtain a hydrophobicity subfamily-conservation score ($H_{subfamily}$) and a polarity subfamily-conservation score ($P_{subfamily}$), respectively. Again, the threshold of 60% conservation level was used in these analyses. After removing redundant discriminating residues already identified by the absolute conservation method, 23 amino acids are assigned as subfamily discriminating residues according to hydrophobicity and/or polarity (Table 19; Table 20).

Table 19: Subfamily discriminating residues according to the hydrophobicity conservation method

Column from left to right: alignment position of the amino acid; hydrophobicity class of the residue used by the majority of sequences of the whole family with the conservation level in brackets; discriminated subfamily; hydrophobicity class of the amino acid used in the discriminated subfamily and its conservation level in brackets.

Alignment position	Whole family Hydrophobicity conservation ($H_{conservation}$)	Discriminated subfamily: Discriminated subfamily Hydrophobicity ($H_{subfamily}$)
78	neutral (0.86)	S2 (Fungi & Oomycetes): hydrophobic (0.62) S30 (Bacteria): hydrophobic (0.83) S33 (Bacteria): hydrophobic (0.67) S35 (Bacteria): hydrophobic (1.0) S45 (Bacteria): hydrophobic (1.0) S49 (Bacteria): hydrophobic (1.0)
132	hydrophobic (0.90)	S6 (Fungi): neutral (0.71) S16 (Archaea): neutral (1.0) S18 (Plants): neutral (0.61) S32 (Bacteria): hydrophilic (0.88)
139	hydrophobic (0.88)	S3 (Fungi): hydrophilic (0.83) S32 (Bacteria): hydrophilic (1.0) S34 (Bacteria): neutral (0.60) S35 (Bacteria): neutral (1.0) S36 (Bacteria & Archaea): hydrophilic (0.67) S48 (Bacteria): hydrophilic (1.0) S50 (Haptophytes): neutral (0.67)
162	hydrophobic (0.99)	S36 (Bacteria & Archaea): neutral (0.67)
163	hydrophobic (0.82)	S11 (Fungi): neutral (1.0) S17 (Plants): neutral (0.71)

Alignment position	Whole family Hydrophobicity conservation ($H_{conservation}$)	Discriminated subfamily: Discriminated subfamily Hydrophobicity ($H_{subfamily}$)
		S19 (Plants): hydrophilic (1.0) S20 (Metazoa): hydrophilic (0.75) S29 (Bacteria): neutral (0.60) S34 (Bacteria): hydrophilic (1.0)
164	hydrophilic (0.98)	S4 (Fungi & Bacteria): neutral (0.91)
233	neutral (0.75)	S10 (Fungi): hydrophobic (1.0) S13 (Fungi): hydrophobic (1.0) S16 (Archaea): hydrophobic (1.0) S17 (Plants): hydrophobic (1.0) S18 (Plants): hydrophobic (0.82) S19 (Plants): hydrophobic (1.0) S20 (Metazoa): hydrophobic (0.95) S32 (Bacteria): hydrophobic (0.75) S34 (Bacteria): hydrophobic (1.0) S36 (Bacteria & Archaea): hydrophobic (1.0) S50 (Haptophytes): hydrophobic (0.67)
234	hydrophobic (0.91)	S16 (Archaea): neutral (0.80) S18 (Plants): neutral (0.71)
237	neutral (0.97)	S10 (Fungi): hydrophobic (1.0) S20 (Metazoa): hydrophobic (0.60)
303	hydrophobic (0.98)	S49 (Bacteria): neutral (1.0)
342	hydrophobic (0.93)	S3 (Fungi): neutral (0.75) S29 (Bacteria): neutral (0.80) S43 (Bacteria): hydrophilic (1.0)

Alignment position	Whole family Hydrophobicity conservation ($H_{conservation}$)	Discriminated subfamily: Discriminated subfamily Hydrophobicity ($H_{subfamily}$)
		S45 (Bacteria): hydrophilic (0.88) S46 (Bacteria): hydrophilic (1.0)
347	hydrophobic (0.95)	S32 (Bacteria): hydrophilic (0.63) S34 (Bacteria): neutral (1.0) S49 (Bacteria): neutral (1.0)
397	hydrophobic (0.89)	S8 (Fungi): neutral (0.60) S9 (Fungi): neutral (0.80) S31 (Bacteria): neutral (1.0) S32 (Bacteria): neutral (0.88) S34 (Bacteria): neutral (1.0)
407	hydrophobic (0.97)	S21 (Bacteria): hydrophilic (1.0) S28 (Bacteria): hydrophilic (1.0)

Table 20: Subfamily discriminating residues according to the polarity conservation method

Column from left to right: alignment position of the amino acid; polarity class of the residue used by the majority sequences of the whole family with the conservation level in brackets; discriminated subfamily; polarity class of the amino acid used in the discriminated subfamily and its conservation level in brackets.

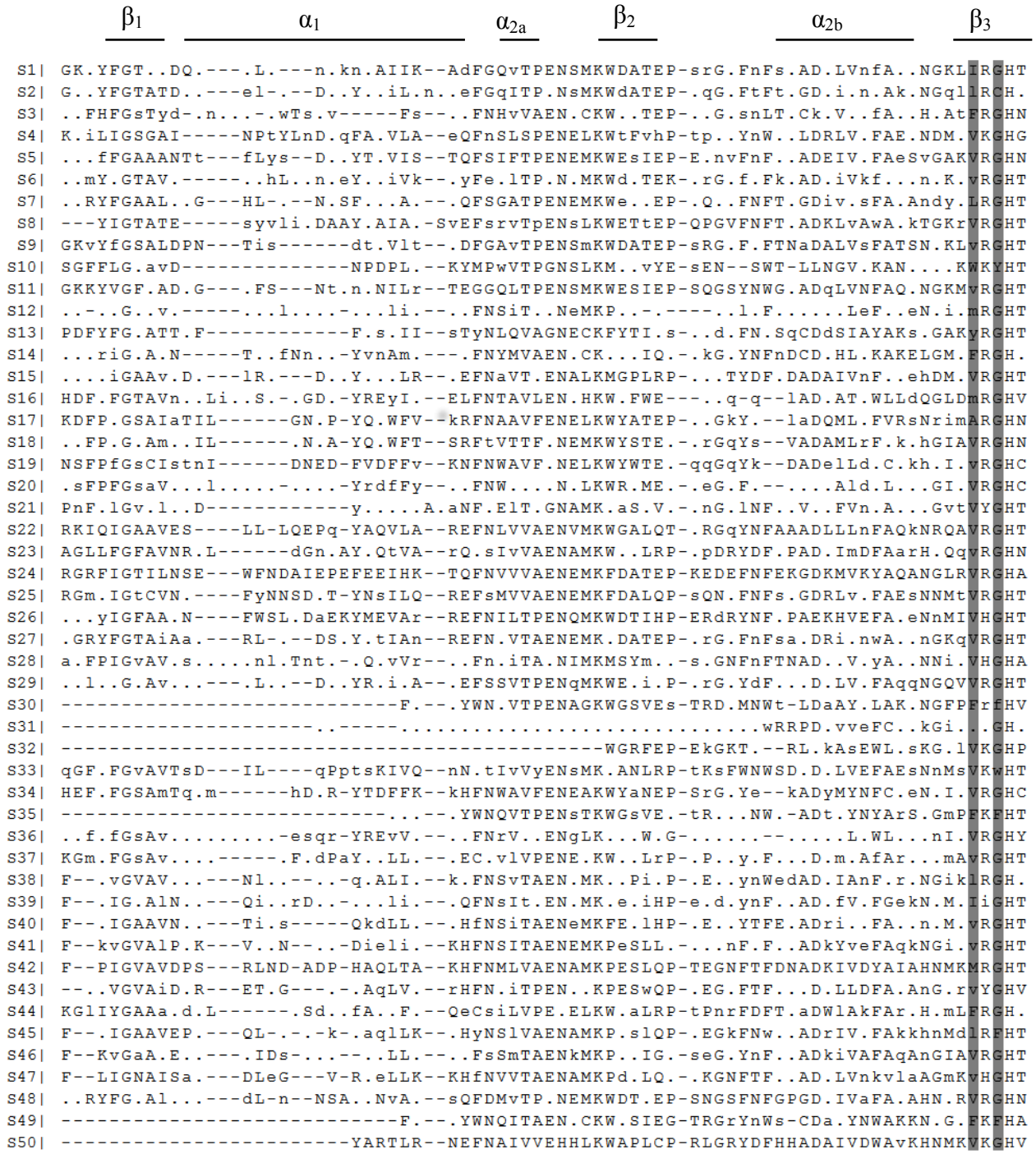
Alignment position	Whole family Polarity conservation ($P_{conservation}$)	Discriminated subfamily: Discriminated subfamily Polarity ($P_{subfamily}$)
76	non polar (0.96)	S13 (Fungi): polar-uncharged (0.60)
78	non polar (0.91)	S2 (Fungi & Oomycetes): polar-uncharged (0.62) S10 (Fungi): polar-uncharged (1.0)
81	non polar (0.91)	S3 (Fungi): polar-uncharged (0.92) S16 (Archaea): polar-uncharged (0.87)
157	non polar (0.90)	S2 (Fungi & Oomycetes): polar-uncharged (0.64) S35 (Bacteria): polar-uncharged (0.67)
163	non polar (0.84)	S6 (Fungi): polar-uncharged (1.0) S17 (Plants): polar-uncharged (1.0) S19 (Plants): polar-uncharged (1.0) S20 (Metazoa): polar-uncharged (0.85) S34 (Bacteria): polar-uncharged (1.0)
233	non polar (0.94)	S30 (Bacteria): polar-uncharged (0.67) S35 (Bacteria): polar-uncharged (1.0) S50 (Haptophytes): polar-uncharged (1.0)

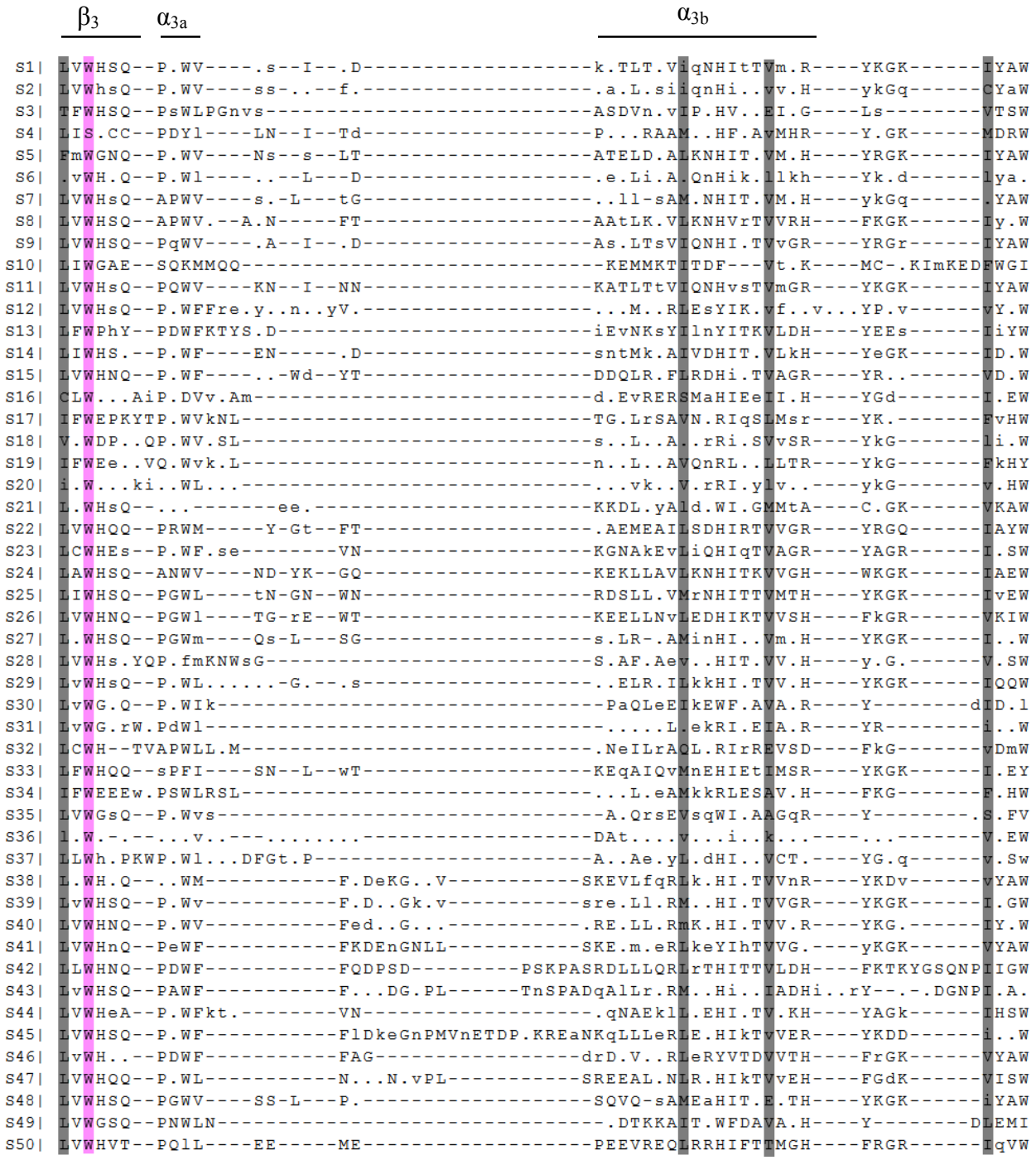
Alignment position	Whole family Polarity conservation ($P_{conservation}$)	Discriminated subfamily: Discriminated subfamily Polarity ($P_{subfamily}$)
234	non polar (0.91)	S16 (Archaea): polar-uncharged (0.67) S18 (Plants): polar-uncharged (0.71)
303	non polar (0.98)	S49 (Bacteria): polar-uncharged (1.0) S50 (Haptophytes): polar-uncharged (1.0)
342	non polar (0.92)	S29 (Bacteria): polar-uncharged (0.80) S43 (Bacteria): polar-uncharged (1.0) S45 (Bacteria): polar-uncharged (0.94) S46 (Bacteria): polar-uncharged (1.0)
345	polar-uncharged (0.98)	S4 (Fungi & Bacteria): non polar (0.82)
347	non polar (0.96)	S32 (Bacteria): polar-uncharged (1.0) S34 (Bacteria): polar-uncharged (1.0) S49 (Bacteria): polar-uncharged (1.0)
348	polar charged (0.95)	S31 (Bacteria): polar-uncharged (1.0) S32 (Bacteria): polar-uncharged (1.0)
349	non polar (0.97)	S34 (Bacteria): polar-uncharged (1.0) S48 (Bacteria): polar-charged (1.0)
396	non polar (0.87)	S3 (Fungi): polar-uncharged (0.92) S23 (Bacteria): polar-uncharged (1.0) S32 (Bacteria): polar-charged (0.63) S35 (Bacteria): polar-charged (1.0) S37 (Bacteria): polar-uncharged (1.0) S44 (Bacteria): polar-uncharged (1.0) S49 (Bacteria): polar-uncharged (0.67)

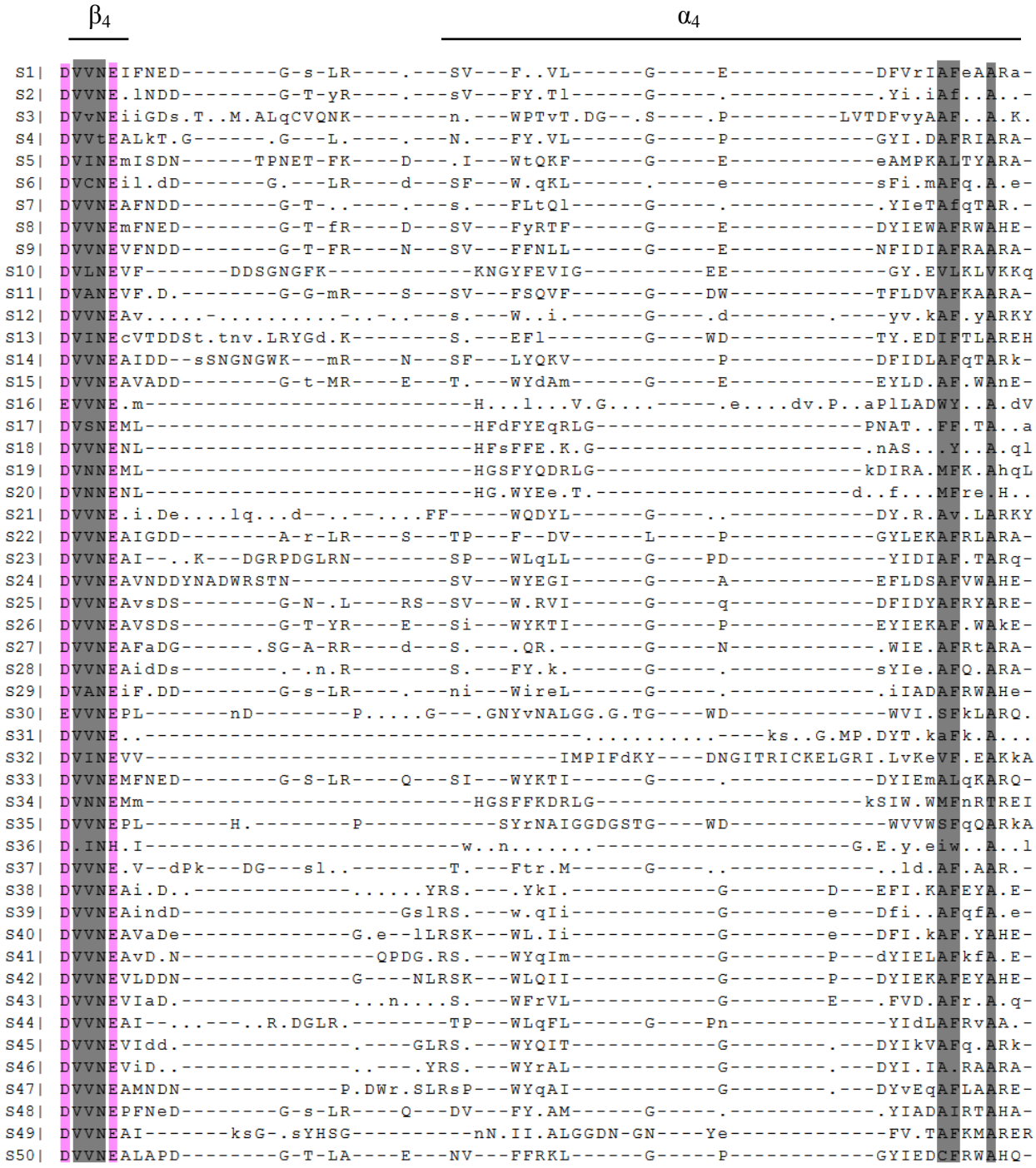
Alignment position	Whole family Polarity conservation ($P_{conservation}$)	Discriminated subfamily: Discriminated subfamily Polarity ($P_{subfamily}$)
397	non polar (0.91)	S8 (Fungi): polar-uncharged (0.60) S9 (Fungi): polar-uncharged (0.60) S31 (Bacteria): polar-uncharged (1.0) S32 (Bacteria): polar-uncharged (0.88) S34 (Bacteria): polar-uncharged (1.0)
411	polar uncharged (0.65)	S3 (Fungi): non polar (0.92) S14 (Fungi): non polar (1.0) S15 (Archaea): non polar (0.89) S16 (Archaea): non polar (1.0) S17 (Plants): non polar (1.0) S18 (Plants): non polar (0.96) S19 (Plants): non polar (1.0) S20 (Metazoa): non polar (1.0) S23 (Bacteria): non polar (1.0) S24 (Bacteria): non polar (1.0) S25 (Bacteria): non polar (1.0) S33 (Bacteria): non polar (1.0) S34 (Bacteria): non polar (1.0) S36 (Bacteria & Archaea): non polar (1.0) S37 (Bacteria): non polar (1.0) S42 (Bacteria): non polar (1.0) S44 (Bacteria): non polar (1.0) S50 (Haptophytes): non polar (1.0)
412	non polar (0.87)	S3 (Fungi): polar-uncharged (0.83) S7 (Fungi): polar-uncharged (1.0)

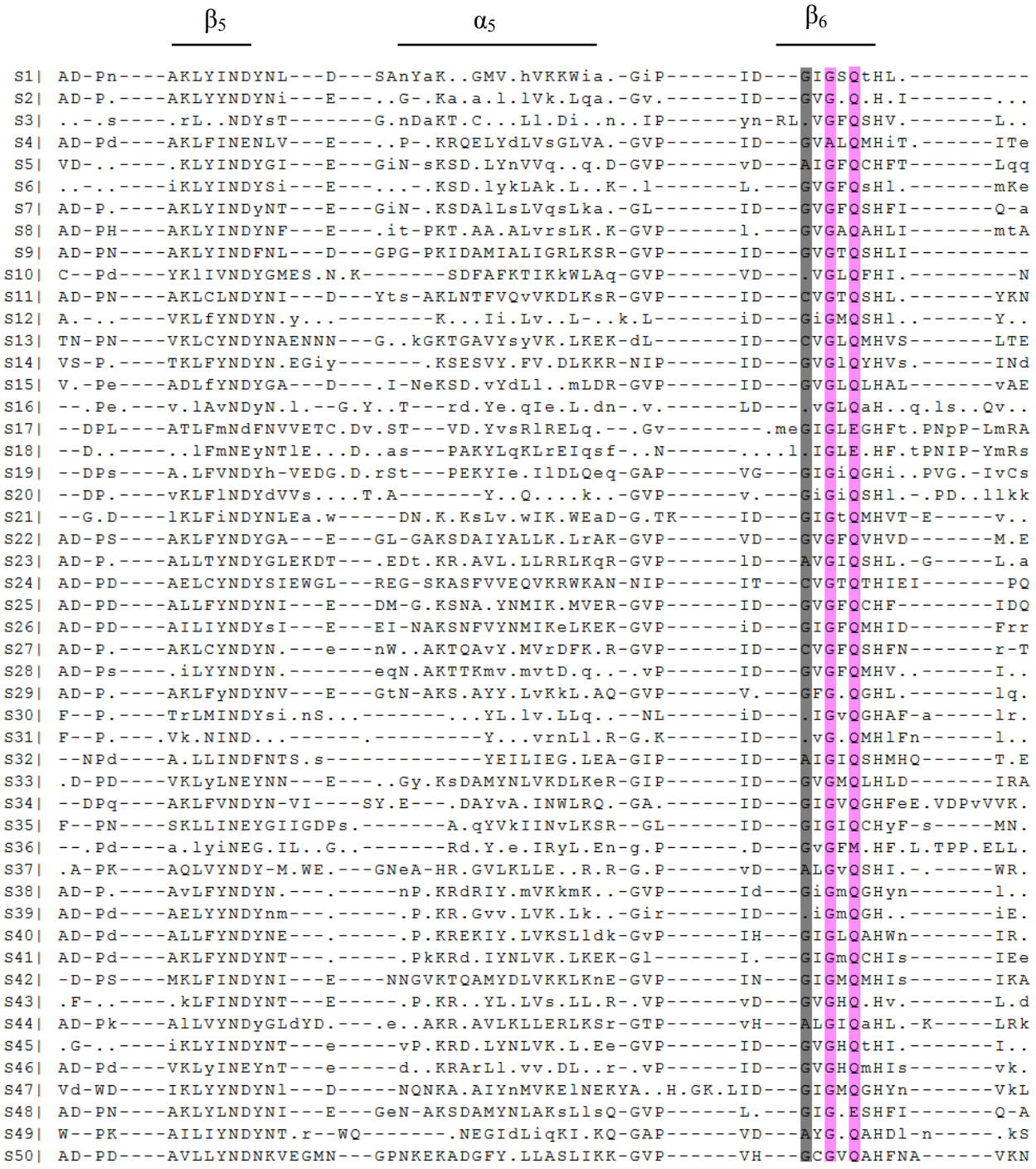
Alignment position	Whole family Polarity conservation ($P_{conservation}$)	Discriminated subfamily: Discriminated subfamily Polarity ($P_{subfamily}$)
		S15 (Archaea): polar-uncharged (0.67)
		S21 (Bacteria): polar-uncharged (0.86)
		S23 (Bacteria): polar-uncharged (1.0)
		S36 (Bacteria & Archaea): polar-uncharged (0.67)
		S44 (Bacteria): polar-uncharged (1.0)
		S48 (Bacteria): polar-uncharged (1.0)

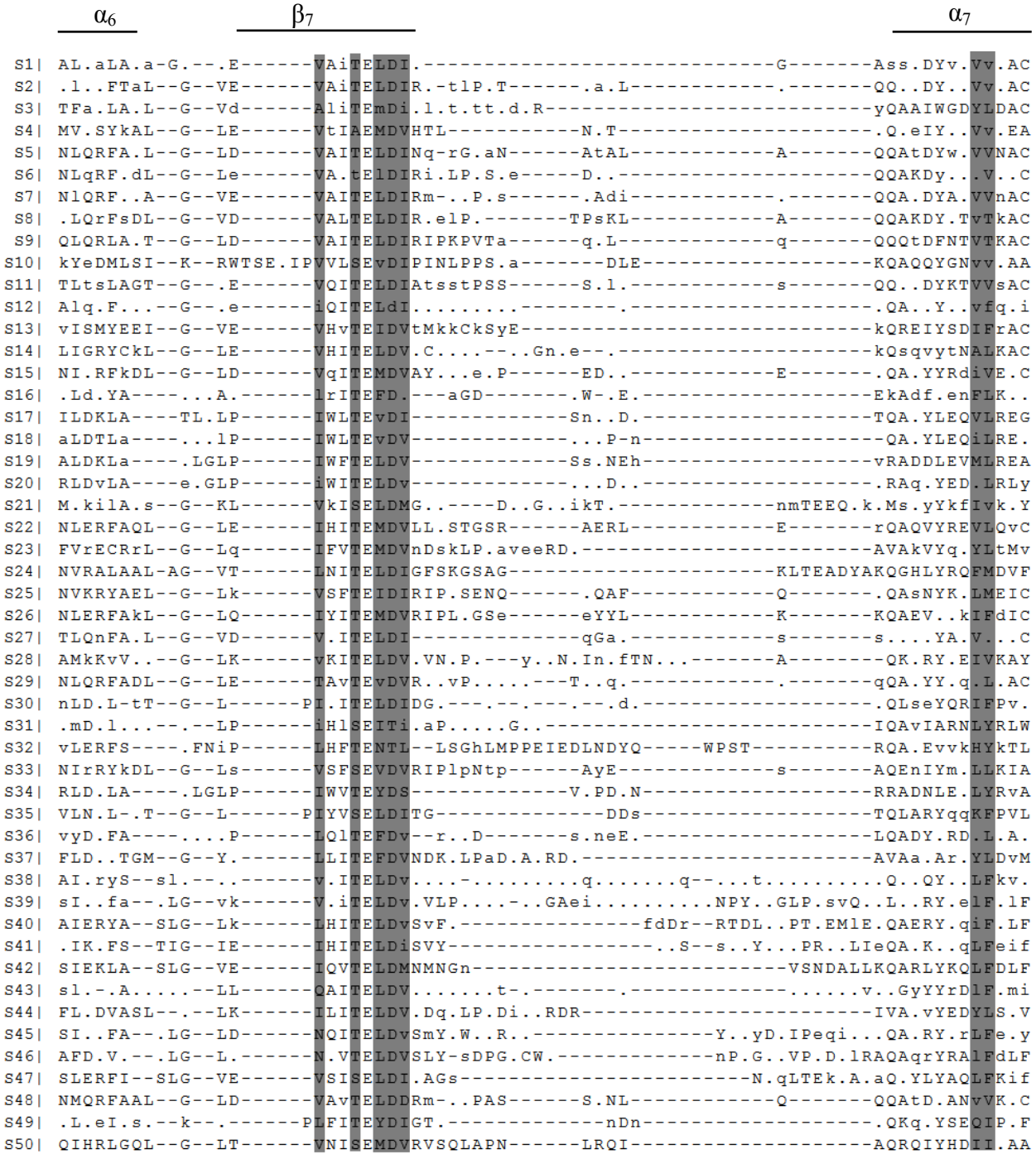
A consensus sequence is obtained from the MSA profile of each subfamily using consensus finder [173]. The subfamily discriminating residues are highlighted in the MSA of the consensus sequences. All except one amino acid are involved in secondary structure formation as predicted from the 3D structures. In summary, a particular position is defined as subfamily discriminating when the residue or the properties (hydrophobicity and/or polarity) of the residue used by the subfamily was different from other subfamilies (Figure 15).


β1
α1
α2a
β2
α2b
β3
S1|
S2|
S3|
S4|
S5|
S6|
S7|
S8|
S9|
S10|
S11|
S12|
S13|
S14|
S15|
S16|
S17|
S18|
S19|
S20|
S21|
S22|
S23|
S24|
S25|
S26|
S27|
S28|
S29|
S30|
S31|
S32|
S33|
S34|
S35|
S36|
S37|
S38|
S39|
S40|
S41|
S42|
S43|
S44|
S45|
S46|
S47|
S48|
S49|
S50|


β3
α3a
α3b


β4
α4


β5
α5
β6


α6
β7
α7

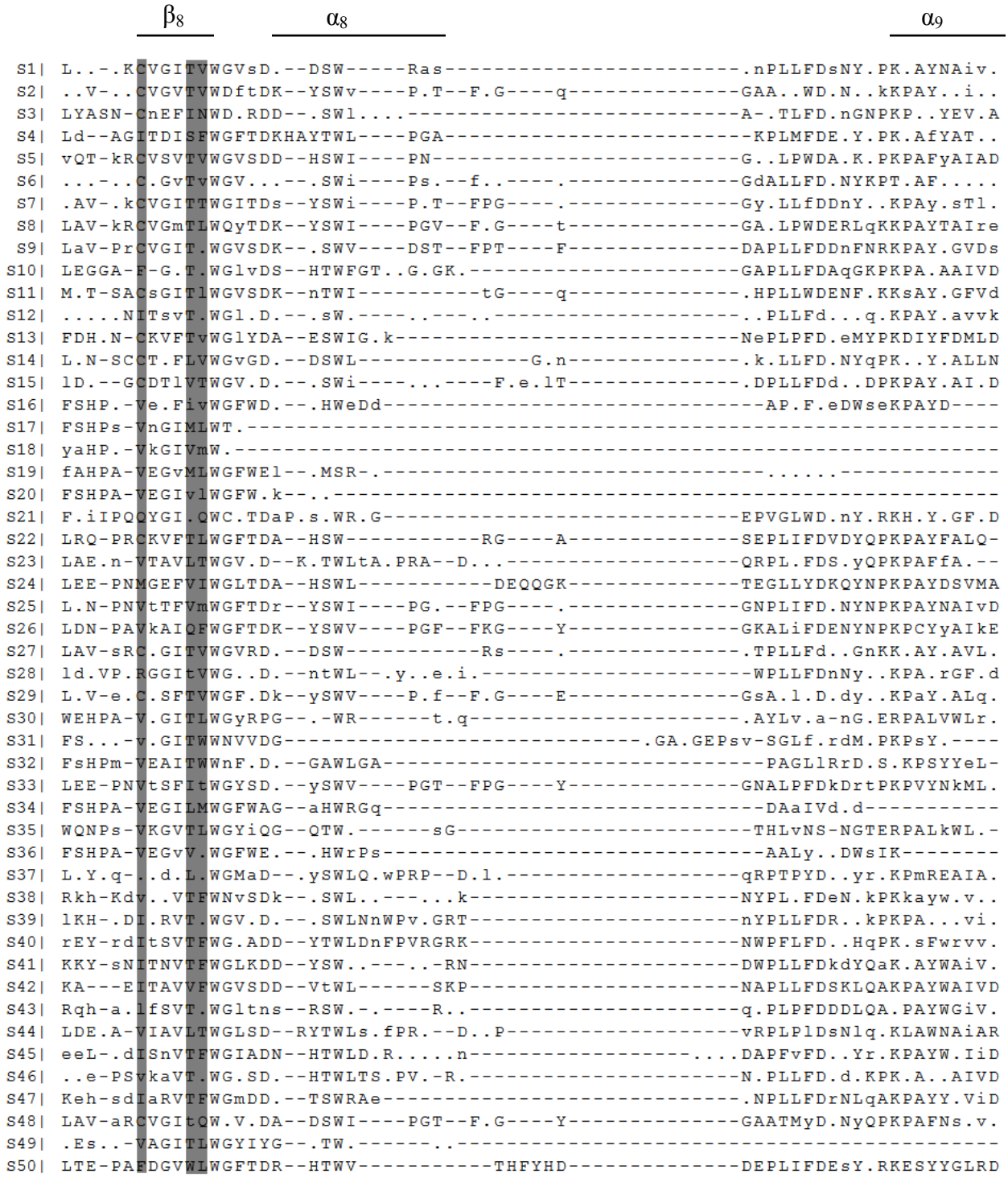

β8
α8
α9

Figure 15: Subfamily discriminating residues of GH10

The MSA of subfamily consensus sequences is shown. Positions highlighted in pink are discriminating residues identified using the absolute conservation method. Discriminating residues highlighted in grey are identified using the hydrophobicity and/or polarity methods. The secondary structure elements α-helices and β-sheets are denoted as α and β, respectively, and are assigned based on experimental 3D structures.

CHAPTER FOUR: DISCUSSION

4.1 Towards a standardized framework for subfamily classification

Sequences belonging to the same protein family are evolutionarily related, hence share similar amino acid sequences and higher order structure as well as mode of action. The size of protein families is increasing continuously due to the rapid accumulation of genomic data. Often, sequences within the same family show significant diversity. Hence, further classification into subfamilies can provide information on evolutionary relationship and functional diversity within the family. Here, I propose a framework of analysis for subfamily classification using glycoside hydrolase family 10 as the template. Protein sequences were retrieved from sequenced genomes across different Kingdoms. Phylogenetic trees were built using the Maximum Likelihood method. Subfamily assignment is based on tree topologies and validated using sequence similarity analysis. The phylogenetic tree shows that GH10 sequences can be clustered in 50 subfamilies (Figure 8). Among those, 46 subfamilies are restricted to a single Kingdom. For instance, 11 subfamilies contain only fungal sequences whereas the 28, 2, and 3 subfamilies contain exclusively sequences from bacteria, archaea, and land plants, respectively. Among the 626 analyzed GH10 sequences, 42 failed to be grouped into any subfamilies. This may be caused by the limited number of sequenced genomes that are closely related to those unclustered sequences. It is likely that these sequences will eventually form new subfamilies when more genomic data become available.

The classification of sequences into subfamilies was first used on GH13 [174]. In that analysis, the sequences of GH13 proteins were retrieved from the CAZy database and the

subfamilies were identified based on the phylogenetic analysis. The major difference between my analysis and GH13 phylogenetic analysis lies in the data sampling step. In the phylogenetic analysis of GH13, the sequences were obtained solely from the CAZy database [75], which resulted in a biased dataset. For instance, the dataset used to generate GH13 phylogenetic tree contains only 62 fungal sequences whereas the number of bacterial sequences is 872 [174]. On the other hand, for my analysis, I retrieved sequences from genomes that represent a wide taxonomic spectrum of different Kingdoms to ensure that the dataset is unbiased. For example, my dataset includes fungal sequences from basal lineages to Ascomycota and Basidiomycota as well as anaerobic species. Presently, CAZy database only contains 171 fungal GH10 sequences, which represent 9.8 % of the total number of available GH10 sequences. I collected all of the fungal sequences and inserted them into my phylogenetic tree to determine where they situate. The sequences collected from CAZy database fell within subfamilies 1, 2, 4, and 14. In other words, if I had only used sequences from CAZy to generate the GH10 fungal phylogenetic tree, I would not have discovered the other subfamilies. This comparison demonstrated the importance of extensive coverage, as complete as possible, of family members in phylogenetic analysis.

4.2 Phylogenetic tree as a screening and prediction tool

Due to the recent improvement in sequencing technology, the accumulation of electronically predicted proteins is increasing rapidly. Presently, it is impossible to experimentally characterize them all individually. Bioinformatic analysis of whole protein families is more realistic, and phylogenetic analysis is one important approach.

The phylogenetic analysis of GH10 has identified 50 subfamilies where 24 of them contain experimentally characterized sequences and/or crystal structures (Figure 8; Table 11). This result shows that a large portion of the family still remains unexplored. The phylogenetic tree can be used as a screening tool to select representative targets from uncharacterized subfamiles for further biochemical characterization.

It was previously established that, within the same family, more closely related sequences also have similar functions [74,117,175]. To investigate the correlation between sequence similarity and biochemical properties, the biochemical properties of experimentally characterized enzymes were mapped onto the phylogenetic tree. The aim was to evaluate if the phylogenetic tree of the protein family can be used to predict the function and the biochemical properties of an uncharacterized sequence. A set of experimentally characterized fungal GH10 proteins was obtained from *myco*CLAP, a database containing fungal lignocellulose-active proteins with manually curated biochemical properties and functional annotations [74]. Following the criteria used in *myco*CLAP, a set of biochemically characterized GH10 sequences from other Kingdoms were manually curated. It is worth mentioning that CAZy database also contains a set of experimentally characterized sequences. The phylogenetic analysis of GH13 also contained experimentally characterized sequences. However, there is a significant discrepancy between the number of characterized sequence harbored in CAZy and *myco*CLAP. For instance, while CAZy contains 60 characterized fungal GH10 proteins as of August 2014, *myco*CLAP holds 31. This discrepancy is caused by the fact that these two databases use different curation criteria. In *myco*CLAP, all the characterized genes have been sequenced and their sequences have been deposited in a public database. In addition, the specific activities of the gene products have been

assayed and the biochemical properties have been published in a peer-reviewed journal. For each entry in the database, the pertinent information is collected manually by curators and the reference papers supporting the evidence are provided. Contrary to *myco*CLAP which follows a set of vigorous rules, the curation process of CAZy characterized proteins seems more ambiguous with less solid supporting evidence. For example, in the CAZy database, Xyn10A (accession number: ACH15005) of *Chrysosporium lucknowense* is curated as a characterized GH10 xylanase. However, no supporting publication that demonstrates the characterization of this enzyme is linked to the entry. Both the phylogenetic analysis of GH13 and my analysis included experimentally characterized data in the common objective of function prediction. However, by propagating annotation based on "characterized" sequences that have less reliable evidence, functional prediction may become less dependable. Based on the comparison between *myco*CLAP and CAZy, it seems the criteria used by the former produce a more reliable set of manually curated sequences.

So far, all but three GH10 sequences encoding biochemically characterized xylanases are of bacterial and fungal origins. Biochemically characterized fungal GH10 sequences are found in subfamilies 1, 2, and 4 (Figure 8; Table 11). Experimental data showed that while enzymes of subfamily 1 show no correlation with their optimal temperature, all except one of them are optimally active at a pH range between 5.0 and 6.0. On the other hand, subfamily 2 enzymes display optimum temperature between 70 and 80°C (Figure 9). No correlation between sequence similarity and pH optimum was observed in this subfamily. For subfamilies containing biochemically characterized proteins of bacterial origin, 10 subfamilies showed a correlation between sequence clustering and pH and/or temperature optimum (Figure 9; Table 12). In

addition, subfamilies with different substrate specificity were also identified. For instance, bacterial xylanases from subfamily 40 are more active on small xylooligosaccharides whereas bacterial enzymes from subfamily 47 prefer polymeric substrates. Through crystal structure comparison, it was demonstrated that subfamily 40 xylanases have narrower catalytic clefts due to the insertion of a loop near the negative subsites which hinders the binding of longer and more branched substrates. The binding preferences of xylanases from different subfamilies were explored further to gain an understanding of the hydrolysis pattern of the enzymes. For example, experimental assays have shown that xylanases belonging to subfamily 45 are able to cleave xylooligosaccharides as short as xylotriose whereas subfamily 42 counterparts require longer substrate. I proposed that the ability of subfamily 45 xylanases to generate xylose from xylooligosaccharides is due to the high binding affinity at their aglycone regions. Among fungal GH10 proteins, only subfamilies 1 and 2 contain sequences with solved structure. Through sequence and structure alignments, it was shown that subfamily 2 xylanases have two extra loops. These xylanases have more closed catalytic clefts compared to their subfamily 1 counterparts. It was proposed by different publications that these two extra loops affect the substrate specificity of the enzymes [120,122,124]. It was suggested that subfamily 2 xylanases which have more closed catalytic clefts prefer linear xylan whereas subfamily 1 enzymes are more active on branched xylan due to their open catalytic clefts. However, some of the data collected from experimentally characterized fungal GH10 sequences disagree with this hypothesis. To confidently confirm the substrate specificity of subfamilies 1 and 2 xylanases, further experimental assays need to be performed. In addition, a clade which is composed of both fungal and bacterial proteins (subfamily 4) also contains an experimentally characterized

fungal GH10 protein. Sequences within this clade are well conserved and show considerable variation as compared to sequences from other subfamilies. The experimentally characterized protein within this clade is a tomatinase. Moreover, 26 subfamilies still lack biochemically characterized members. In this case, phylogenetic tree can be used to select target proteins from uncharacterized subfamilies for further study.

4.3 Glycoside Hydrolase Family 10: An ancient protein family with great diversity

The global distribution analysis showed that GH10 protein-encoding genes are found in fungi, green plants, metazoa, bacteria, and archaea. Other eukaryotes such as oomycetes, diatoms, haptophytes, and choanoflagellates also harbor GH10 genes. Phylogenetic analysis revealed that glycoside hydrolase family 10 proteins can be clustered into 50 subfamilies, suggesting a highly complex evolutionary pathway for the family. This tree topology can be explained by multiple gene duplications followed by lineage specific gene loss. In addition, the phylogeny of the protein tree does not reflect the evolution of species. For instance, the tree shows that fungal GH10 sequences are more related to bacterial genes than to those of metazoa which is incongruent with the previously established evolutionary relationship between fungi and metazoa [103]. The presence of GH10 genes in the archaea, bacteria, and eukaryotic domains of life and the complex topology of the family tree suggest the existence of an ancient form of GH10 gene prior to the appearance of the eukaryotic lineages.

According to amino acid sequence similarity analysis, sequences from different subfamilies display considerable variation at their amino acid sequence level. It is well known

that gene duplications could generate redundant genes, which might eventually result in new functions. Amino acid conservation of GH10 proteins was analyzed. In total, I have identified 47 globally conserved residues, based on identity and/or class similarity, across the whole family (Table 16; Table 17). These residues are found on both α-helices and β-sheets which reflects their importance in the function of the protein. The accumulation of mutations among subfamilies reflects how they have diverged during evolution and may be responsible for the development of new properties and/or functions of a subfamily. By performing subfamily discriminating residue analyses, detailed lists of subfamily discriminating residues were obtained (Table 18; Table 19; Table 20). These analyses would provide a guide to investigate how these residues affect the structure and function of the enzymes belonging to different subfamilies.

CHAPTER FIVE: CONCLUSION

With the number of sequenced genomes becoming more and more abundant, it is impossible to perform functional and structural analyses on all individual genes. At this stage, comprehensive analyses of protein families using bioinformatic approaches to infer function and structure are more suitable.

The purpose of this research was to establish a framework for protein family analysis. Glycoside hydrolase family 10 was used as the template. This glycoside hydrolase family contains endo-1, 4-beta-xylanase that cleaves the backbone of xylan, the most abundant type of hemicellulose. Within the family, GH10 xylanases show considerable diversity, which is reflected by the structural complexity of xylan. By performing a phylogenetic analysis, I hoped to develop a standard procedure to classify sequences into subfamilies.

The phylogenetic analysis showed that 586 out of 626 (93.6%) analyzed GH10 sequences can be classified into 50 well-supported subfamilies (Figure 8; Table 11; Supplementary file 3). Among these, 46 subfamilies contain sequences that are restricted to a single Kingdom. The distribution analysis showed that GH10 genes are found in the Archaea, Bacteria and Eukaryotes domains, suggesting an ancient origin of the GH10 family. In addition, the Maximum Likelihood phylogeny of GH10 proteins does not reflect the previously established species tree. The complex topology of the family tree strongly argues that divergence of GH10 genes preceded the appearance of the eukaryotic lineage and the emergence of multiple subfamilies were resulted from duplication events followed by lineage specific gene loss.

To investigate the correlation between sequence similarity and biochemical properties, experimental data of biochemically characterized GH10 proteins were mapped onto the phylogenetic tree. The aim was to better understand the structure and the function of each subfamily. It is hoped that, by incorporating experimental data, a phylogenetic tree can be used as a prediction tool to annotate uncharacterized members of a protein family. To avoid the propagation of mis-annotation and to properly assign function to uncharacterized genes, a set of reference sequences with reliable experimental evidence is essential. Biochemically characterized fungal GH10 sequences were collected from the *myco*CLAP database [74]. This database only contains annotated fungal glycoside hydrolases with experimental evidence. In addition, a set of bacterial genes encoding biochemically characterized family 10 glycoside hydrolases as well as those from organisms of other Kingdoms were manually curated. This dataset will be incorporated into the *myco*CLAP database.

The mapping of proteins with functional data showed that 13 subfamilies display correlations to pH and/or temperature optima. Previous studies such as the analysis of the GH13 family showed that sequences with the same substrate specificity are clustered together [74,117,175]. This correlation is less clear to visualize on GH10 phylogenetic tree as the majority of the sequences of this family are endo-1, 4-beta-xylanases (EC 3.2.1.8) except for subfamily 4 which shows tomatinase activity. However, comparison of crystal structures of the enzymes from different subfamilies shows discernible difference. These observations suggest that xylanases from different subfamilies hydrolyze xylan differently and show preference towards different types of xylan substrate.

In conclusion, I have used different bioinformatic approaches to study glycoside hydrolase family 10 proteins. It is hoped that this project can be used as a framework to study other protein families. The phylogenetic tree can be used to classify sequences into subfamilies and further understand the evolution of the protein family. The mapping of experimental data onto the protein tree served to establish relationships between sequences and function. Finally, subfamily discriminating residue analyses allowed us to identify amino acids that might be responsible for different function between subfamilies.

REFERENCES

1. Cherubini F, Str mman AH. Principles of Biorefining. In: *Biofuels: Alternative Feedstocks and Conversion Processes*. Academic Press, 642 (2011).

2. McMillan JD. Bioethanol production: Status and prospects. *Renew. Energy*. , 295–302 (1997).

3. Fatih Demirbas M. Biorefineries for biofuel upgrading: A critical review. *Appl. Energy*. 86, Supplement 1, S151–S161 (2009).

4. Naik SN, Goud VV, Rout PK, Dalai AK. Production of first and second generation biofuels: A comprehensive review. *Renew. Sustain. Energy Rev.* 14(2), 578–597 (2010).

5. Carriquiry MA, Du X, Timilsina GR. Second generation biofuels: Economics and policies. *Energy Policy*. 39(7), 4222–4234 (2011).

6. Mabee WE, Saddler JN. Bioethanol from lignocellulosics: Status and perspectives in Canada. *Bioresour. Technol.* 101(13), 4806–4813 (2010).

7. Soccol CR, Faraco V, Karp S, *et al.* Lignocellulosic Bioethanol: Current Status and Future Perspectives. In: *Biofuels: Alternative Feedstocks and Conversion Processes*. Academic Press, 642 (2011).

8. Chandel AK, Singh OV. Weedy lignocellulosic feedstock and microbial metabolic engineering: advancing the generation of "Biofuel." *Appl. Microbiol. Biotechnol.* 89(5), 1289–1303 (2011).

9. Decker SR, Sheehan J, Dayton DC, *et al.* Biomass Conversion [Internet]. In: *Kent and Riegel's Handbook of Industrial Chemistry and Biotechnology*. Kent JA (Ed.). . Springer US, Boston, MA, 1449–1548 (2007) [cited 2014 Feb 18]. Available from: http://adsabs.harvard.edu/abs/2007karh.book.1449D.

10. Bhaskar T, Bhavya B, Singh R, Naik DV, Kumar A, Goyal HB. Thermochemical Conversion of Biomass to Biofuels. In: *Biofuels: Alternative Feedstocks and Conversion Processes*. Academic Press, 642 (2011).

11. Warren RA. Microbial hydrolysis of polysaccharides. *Annu. Rev. Microbiol.* 50, 183–212 (1996).

12. Henrissat B, Davies G. Structural and sequence-based classification of glycoside hydrolases. *Curr. Opin. Struct. Biol.* 7(5), 637–644 (1997).

13. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37(Database issue), D233–238 (2009).

14. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* 280 (Pt 2), 309–316 (1991).

15. International Union of Biochemistry, Biochemical Society (Great Britain). Biochemical nomenclature and related documents. William Clowes & Sons for the Biochemical Society, London.

16. Ebringerová A, Heinze T. Xylan and xylan derivatives – biopolymers with valuable properties, 1. Naturally occurring xylans structures, isolation procedures and properties. *Macromol. Rapid Commun.* 21(9), 542–556 (2000).

17. Ebringerová A. Structural Diversity and Application Potential of Hemicelluloses. *Macromol. Symp.* 232(1), 1–12 (2005).

18. Wong KK, Tan LU, Saddler JN. Multiplicity of beta-1,4-xylanase in microorganisms: functions and applications. *Microbiol. Rev.* 52(3), 305 (1988).

19. Polizeli MLTM, Rizzatti ACS, Monti R, Terenzi HF, Jorge JA, Amorim DS. Xylanases from fungi: properties and industrial applications. *Appl. Microbiol. Biotechnol.* 67(5), 577–591 (2005).

20. Sunna A, Antranikian G. Xylanolytic enzymes from fungi and bacteria. *Crit. Rev. Biotechnol.* 17(1), 39–67 (1997).

21. Deutschmann R, Dekker RFH. From plant biomass to bio-based chemicals: latest developments in xylan research. *Biotechnol. Adv.* 30(6), 1627–1640 (2012).

22. Henrissat B, Bairoch A. New families in the classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem. J.* 293 (Pt 3), 781–788 (1993).

23. Derewenda U, Swenson L, Green R, *et al.* Crystal structure, at 2.6-A resolution, of the Streptomyces lividans xylanase A, a member of the F family of beta-1,4-D-glycanases. *J. Biol. Chem.* 269(33), 20811–20814 (1994).

24. Törrönen A, Harkki A, Rouvinen J. Three-dimensional structure of endo-1,4-beta-xylanase II from Trichoderma reesei: two conformational states in the active site. *EMBO J.* 13(11), 2493–2501 (1994).

25. Biely P, Vršanská M, Tenkanen M, Kluepfel D. Endo-β-1,4-xylanase families: differences in catalytic properties. *J. Biotechnol.* 57(1–3), 151–166 (1997).

26. Dodd D, Cann IKO. Enzymatic deconstruction of xylan for biofuel production. *Glob. Change Biol. Bioenergy*. 1(1), 2–17 (2009).

27. Ryabova O, Vrsanská M, Kaneko S, van Zyl WH, Biely P. A novel family of hemicellulolytic alpha-glucuronidase. *FEBS Lett.* 583(9), 1457–1462 (2009).

28. Sharma M, Bhupinder SC. Production of Hemicellulolytic Enzymes for Hydrolysis of Lignocellulosic Biomass. In: *Biofuels: Alternative Feedstocks and Conversion Processes*. Academic Press, 642 (2011).

29. Tenkanen M, Siika-aho M. An alpha-glucuronidase of Schizophyllum commune acting on polymeric xylan. *J. Biotechnol.* 78(2), 149–161 (2000).

30. Biely P, MacKenzie CR, Puls J, Schneider H. Cooperativity of Esterases and Xylanases in the Enzymatic Degradation of Acetyl Xylan. *Nat. Biotechnol.* 4(8), 731–733 (1986).

31. Grohmann K, Mitchell DJ, Himmel ME, Dale BE, Schroeder HA. The role of ester groups in resistance of plant cell wall polysaccharides to enzymatic hydrolysis. *Appl. Biochem. Biotechnol.* 20-21(1), 45–61 (1989).

32. Li X-L, Skory CD, Cotta MA, Puchart V, Biely P. Novel family of carbohydrate esterases, based on identification of the Hypocrea jecorina acetyl esterase gene. *Appl. Environ. Microbiol.* 74(24), 7482–7489 (2008).

33. Andrade CMMC, Aguiar WB, Antranikian G. Physiological aspects involved in production of xylanolytic enzymes by deep-sea hyperthermophilic archaeon Pyrodictium abyssi. *Appl. Biochem. Biotechnol.* 91-93(1-9), 655–669 (2001).

34. Uhl AM, Daniel RM. The first description of an archaeal hemicellulase: the xylanase from Thermococcus zilligii strain AN1. *Extrem. Life Extreme Cond.* 3(4), 263–267 (1999).

35. Ahmed S, Riaz S, Jamil A. Molecular cloning of fungal xylanases: an overview. *Appl. Microbiol. Biotechnol.* 84(1), 19–35 (2009).

36. Harris GW, Jenkins JA, Connerton I, *et al.* Structure of the catalytic core of the family F xylanase from Pseudomonas fluorescens and identification of the xylopentaose-binding sites. *Struct. Lond. Engl. 1993*. 2(11), 1107–1116 (1994).

37. Henrissat B, Callebaut I, Fabrega S, Lehn P, Mornon JP, Davies G. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc. Natl. Acad. Sci. U. S. A.* 92(15), 7090–7094 (1995).

38. MacLeod AM, Lindhorst T, Withers SG, Warren RA. The acid/base catalyst in the exoglucanase/xylanase from Cellulomonas fimi is glutamic acid 127: evidence from detailed kinetic studies of mutants. *Biochemistry (Mosc.)*. 33(20), 6371–6376 (1994).

39. McCarter JD, Withers SG. Mechanisms of enzymatic glycoside hydrolysis. *Curr. Opin. Struct. Biol.* 4(6), 885–892 (1994).

40. Davies GJ, Wilson KS, Henrissat B. Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem. J.* 321(Pt 2), 557 (1997).

41. Schmidt A, Gubitz GM, Kratky C. Xylan binding subsite mapping in the xylanase from Penicillium simplicissimum using xylooligosaccharides as cryo-protectant. *Biochemistry (Mosc.)*. 38, 2403–2412 (1998).

42. Grigoriev IV, Nordberg H, Shabalov I, *et al.* The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res.* 40(Database issue), D26–32 (2012).

43. Grigoriev IV, Nikitin R, Haridas S, *et al.* MycoCosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Res.* 42(Database issue), D699–704 (2014).

44. Cuomo CA, Birren BW. The fungal genome initiative and lessons learned from genome sequencing. *Methods Enzymol.* 470, 833–855 (2010).

45. Markowitz VM, Ivanova NN, Szeto E, *et al.* IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* 36(Database issue), D534–538 (2008).

46. Goodstein DM, Shu S, Howson R, *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* 40(Database issue), D1178–1186 (2012).

47. Pevsner J. Multiple Sequence Alignment. In: *Bioinformatics and functional genomics*. Wiley-Blackwell, Hoboken, N.J. (2009).

48. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22(22), 4673–4680 (1994).

49. Notredame C. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*. 3(1), 131–144 (2002).

50. Pei J. Multiple protein sequence alignment. *Curr. Opin. Struct. Biol.* 18(3), 382–386 (2008).

51. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32(5), 1792–1797 (2004).

52. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30(14), 3059–3066 (2002).

53. Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302(1), 205–217 (2000).

54. Felsenstein J. Distance Methods for Inferring Phylogenies: A Justification. *Evolution*. 38(1), 16–24 (1984).

55. Van de Peer Y. Phylogenetic inference based on distance methods: theory [Internet]. In: *The phylogenetic handbook : a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press, 142–160 (2009) [cited 2014 Feb 11]. Available from: http://hdl.handle.net/1854/LU-593074.

56. Brinkman FSL, Leipe DD. Phylogenetic Analysis [Internet]. In: *Bioinformatics*. Baxevanis AD, Ouellette BFF (Eds.). . John Wiley & Sons, Inc., 323–358 (2002) [cited 2014 Feb 13]. Available from: http://onlinelibrary.wiley.com/doi/10.1002/0471223921.ch14/summary.

57. Pevsner J. Molecular Phylogeny and Evolution. In: *Bioinformatics and functional genomics*. Wiley-Blackwell, Hoboken, N.J. (2009).

58. Hillis DM, Moritz C, Mable BK, editors. Molecular systematics. 2nd ed. Sinauer Associates, Sunderland, Mass.

59. Bishop MJ, Rawlings CJ. DNA and protein sequence analysis: a practical approach. IRL Press at Oxford University Press, Oxford; New York.

60. Liò P, Goldman N. Models of molecular evolution and phylogeny. *Genome Res.* 8(12), 1233–1244 (1998).

61. Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet. TIG*. 17(5), 262–272 (2001).

62. Liò P, Goldman N. Models of molecular evolution and phylogeny. *Genome Res.* 8(12), 1233–1244 (1998).

63. Sneath PHA, Sokal RR. Numerical Taxonomy: The Principles and Practice of Numerical Classification. W H Freeman Limited.

64. Sokal RR, Michener CD, Kansas U of. A Statistical Method for Evaluating Systematic Relationships. University of Kansas.

65. Pevsner J. Bioinformatics and functional genomics. Wiley-Blackwell, Hoboken, N.J.

66. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4(4), 406–425 (1987).

67. Studier JA, Keppler KJ. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol. Biol. Evol.* 5(6), 729–731 (1988).

68. Yang Z. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* 42(2), 294–307 (1996).

69. Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends Genet. TIG*. 17(5), 262–272 (2001).

70. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17(6), 368–376 (1981).

71. Hasegawa M, Fujiwara M. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phylogenet. Evol.* 2(1), 1–5 (1993).

72. Huelsenbeck JP. Performance of Phylogenetic Methods in Simulation. *Syst. Biol.* 44(1), 17–48 (1995).

73. Kuhner MK, Felsenstein J. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.* 11(3), 459–468 (1994).

74. Murphy C, Powlowski J, Wu M, Butler G, Tsang A. Curation of characterized glycoside hydrolases of Fungal origin. *Database*. 2011(0), bar020–bar020 (2011).

75. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42(Database issue), D490–495 (2014).

76. Punta M, Coggill PC, Eberhardt RY, *et al.* The Pfam protein families database. *Nucleic Acids Res.* 40(D1), D290–D301 (2011).

77. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinforma. Oxf. Engl.* 21(4), 456–463 (2005).

78. Bernstein FC, Koetzle TF, Williams GJ, *et al.* The Protein Data Bank: a computer-based archival file for macromolecular structures. *Arch. Biochem. Biophys.* 185(2), 584–591 (1978).

79. The PyMOL Molecular Graphics System. Schrödinger, LLC.

80. Liu X. Comprehensive bioinformatic analysis of kinesin classification and prediction of structural changes from a closed to an open conformation of the motor domain. (2009).

81. Yuan J, Zhao Y. Evolutionary aspects of the synuclein super-family and sub-families based on large-scale phylogenetic and group-discrimination analysis. *Biochem. Biophys. Res. Commun.* 441(2), 308–317 (2013).

82. Lumbsch HT, Huhndorf SM. Outline of Ascomycota-2007. *Myconet*. 13, 1–58 (2007).

83. Sugiyama J, Hosaka K, Suh S-O. Early Diverging Ascomycota: Phylogenetic Divergence and Related Evolutionary Enigmas. *Mycologia*. 98(6), 996–1005 (2006).

84. Webster J. Introduction to fungi. 3rd ed. Cambridge University Press, Cambridge ; New York.

85. Suh S-O, Blackwell M, Kurtzman CP, Lachance M-A. Phylogenetics of Saccharomycetales, the Ascomycete Yeasts. *Mycologia*. 98(6), 1006–1017 (2006).

86. Gazis R, Miadlikowska J, Lutzoni F, Arnold AE, Chaverri P. Culture-based study of endophytes associated with rubber trees in Peru reveals a new class of Pezizomycotina: Xylonomycetes. *Mol. Phylogenet. Evol.* 65(1), 294–304 (2012).

87. Spatafora JW, Sung G-H, Johnson D, *et al.* A Five-Gene Phylogeny of Pezizomycotina. *Mycologia*. 98(6), 1018–1028 (2006).

88. Swann EC, Taylor JW. Higher Taxa of Basidiomycetes: An 18S rRNA Gene Perspective. *Mycologia*. 85(6), 923–936 (1993).

89. Ainsworth GC, CABI Bioscience. Ainsworth & Bisby's dictionary of the fungi / by P.M. Kirk ... [et al.] ; with the assistance of A. Aptroot ... [et al.]. 9th ed. CABI Pub, Wallingford, Oxon, UK ; New York, NY.

90. Hibbett DS. A Phylogenetic Overview of the Agaricomycotina. *Mycologia*. 98(6), 917–925 (2006).

91. Aime MC, Matheny PB, Henk DA, *et al.* An Overview of the Higher Level Classification of Pucciniomycotina Based on Combined Analyses of Nuclear Large and Small Subunit rDNA Sequences. *Mycologia*. 98(6), 896–905 (2006).

92. Begerow D, Stoll M, Bauer R. A Phylogenetic Hypothesis of Ustilaginomycotina Based on Multiple Gene Analyses and Morphological Data. *Mycologia*. 98(6), 906–916 (2006).

93. James TY, Kauff F, Schoch CL, *et al.* Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature*. 443(7113), 818–822 (2006).

94. White MM, James TY, O'Donnell K, Cafaro MJ, Tanabe Y, Sugiyama J. Phylogeny of the Zygomycota based on nuclear ribosomal sequence data. *Mycologia*. 98(6), 872–884 (2006).

95. Corradi N, Akiyoshi DE, Morrison HG, *et al.* Patterns of genome evolution among the microsporidian parasites Encephalitozoon cuniculi, Antonospora locustae and Enterocytozoon bieneusi. *PloS One*. 2(12), e1277 (2007).

96. James TY, Letcher PM, Longcore JE, *et al.* A molecular phylogeny of the flagellated fungi (Chytridiomycota) and description of a new phylum (Blastocladiomycota). *Mycologia*. 98(6), 860–871 (2006).

97. Fischer M. A new wood-decaying basidiomycete species associated with esca of grapevine: Fomitiporia mediterranea (Hymenochaetales). *Mycol. Prog.* 1(3), 315–324 (2002).

98. DiGuistini S, Wang Y, Liao NY, *et al.* Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont Grosmannia clavigera, a lodgepole pine pathogen. *Proc. Natl. Acad. Sci. U. S. A.* 108(6), 2504–2509 (2011).

99. Haridas S, Wang Y, Lim L, *et al.* The genome and transcriptome of the pine saprophyte Ophiostoma piceae, and a comparison with the bark beetle-associated pine pathogen Grosmannia clavigera. *BMC Genomics*. 14, 373 (2013).

100. Bremer K. Summary of Green Plant Phylogeny and Classification. *Cladistics*. 1(4), 369–385 (1985).

101. Edgecombe GD, Giribet G, Dunn CW, *et al.* Higher-level metazoan relationships: recent progress and remaining questions. *Org. Divers. Evol.* 11(2), 151–172 (2011).

102. Markowitz VM, Szeto E, Palaniappan K, *et al.* The integrated microbial genomes (IMG) system in 2007: data content and analysis tool extensions. *Nucleic Acids Res.* 36(Database issue), D528–533 (2008).

103. Baldauf SL. An overview of the phylogeny and diversity of eukaryotes. *Acta Phytotaxon. Sin.* 46(3), 263–273 (2008).

104. Keeling PJ, Burger G, Durnford DG, *et al.* The tree of eukaryotes. *Trends Ecol. Evol.* 20(12), 670–676 (2005).

105. Ventura M, Canchaya C, Tauch A, *et al.* Genomics of Actinobacteria: tracing the evolutionary history of an ancient phylum. *Microbiol. Mol. Biol. Rev. MMBR*. 71(3), 495–548 (2007).

106. Boone DR, Castenholz RW, Garrity GM, editors. Bergey's manual of systematic bacteriology. 2nd ed. Springer, New York.

107. Kersters K, De Vos P, Gillis M, Swings J, Vandamme P, Stackebrandt E. Introduction to the Proteobacteria [Internet]. In: *The Prokaryotes: A Handbook on the Biology of Bacteria*. Springer (2006) [cited 2014 Mar 17]. Available from: http://link.springer.com/referenceworkentry/10.1007%2F0-387-30745-1_1/fulltext.html.

108. Woese CR. Bacterial evolution. *Microbiol. Rev.* 51(2), 221–271 (1987).

109. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U. S. A.* 87(12), 4576–4579 (1990).

110. Barns SM, Delwiche CF, Palmer JD, Pace NR. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* 93(17), 9188–9193 (1996).

111. Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. *Nat. Rev. Microbiol.* 6(3), 245–252 (2008).

112. Huber H, Hohn MJ, Rachel R, Fuchs T, Wimmer VC, Stetter KO. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*. 417(6884), 63–67 (2002).

113. Roldán-Arjona T, Pérez-Espinosa A, Ruiz-Rubio M. Tomatinase from Fusarium oxysporum f. sp. lycopersici defines a new class of saponinases. *Mol. Plant-Microbe Interact. MPMI*. 12(10), 852–861 (1999).

114. Yin L-F, Wang F, Zhang Y, *et al.* Evolutionary analysis revealed the horizontal transfer of the Cyt b gene from Fungi to Chromista. *Mol. Phylogenet. Evol.* 76, 155–161 (2014).

115. Ospina-Giraldo MD, Griffith JG, Laird EW, Mingora C. The CAZyome of Phytophthora spp.: a comprehensive analysis of the gene complement coding for carbohydrate-active enzymes in species of the genus Phytophthora. *BMC Genomics*. 11, 525 (2010).

116. Adelsberger H, Hertel C, Glawischnig E, Zverlov VV, Schwarz WH. Enzyme system of Clostridium stercorarium for hydrolysis of arabinoxylan: reconstitution of the in vivo system from recombinant enzymes. *Microbiol. Read. Engl.* 150(Pt 7), 2257–2266 (2004).

117. Stam MR, Danchin EGJ, Rancurel C, Coutinho PM, Henrissat B. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng. Des. Sel. PEDS*. 19(12), 555–562 (2006).

118. Haas H, Herfurth E, Stöffler G, Redl B. Purification, characterization and partial amino acid sequences of a xylanase produced by Penicillium chrysogenum. *Biochim. Biophys. Acta*. 1117(3), 279–286 (1992).

119. Hou Y-H, Wang T-H, Long H, Zhu H-Y. Novel cold-adaptive Penicillium strain FS010 secreting thermo-labile xylanase isolated from Yellow Sea. *Acta Biochim. Biophys. Sin.* 38(2), 142–149 (2006).

120. Schmidt A, Schlacher A, Steiner W, Schwab H, Kratky C. Structure of the xylanase from Penicillium simplicissimum. *Protein Sci. Publ. Protein Soc.* 7(10), 2081–2088 (1998).

121. The PyMOL Molecular Graphics System. Schrödinger, LLC.

122. Van Gool MP, van Muiswinkel GCJ, Hinz SWA, Schols HA, Sinitsyn AP, Gruppen H. Two GH10 endo-xylanases from Myceliophthora thermophila C1 with and without cellulose binding module act differently towards soluble and insoluble xylans. *Bioresour. Technol.* 119, 123–132 (2012).

123. Vardakou M, Flint J, Christakopoulos P, Lewis RJ, Gilbert HJ, Murray JW. A family 10 Thermoascus aurantiacus xylanase utilizes arabinose decorations of xylan as significant substrate specificity determinants. *J. Mol. Biol.* 352(5), 1060–1067 (2005).

124. Lo Leggio L, Kalogiannis S, Eckert K, *et al.* Substrate specificity and subsite mobility in T. aurantiacus xylanase 10A. *FEBS Lett.* 509(2), 303–308 (2001).

125. Belancic A, Scarpa J, Peirano A, Díaz R, Steiner J, Eyzaguirre J. Penicillium purpurogenum produces several xylanases: purification and properties of two of the enzymes. *J. Biotechnol.* 41(1), 71–79 (1995).

126. Pollet A, Beliën T, Fierens K, Delcour JA, Courtin CM. Fusarium graminearum xylanases show different functional stabilities, substrate specificities and inhibition sensitivities. *Enzyme Microb. Technol.* 44(4), 189–195 (2009).

127. Furniss CS, Williamson G, Kroon PA. The substrate specificity and susceptibility to wheat inhibitor proteins of Penicillium funiculosum xylanases from a commercial enzyme preparation. *J. Sci. Food Agric.* 85(4), 574–582 (2005).

128. Decelle B, Tsang A, Storms RK. Cloning, functional expression and characterization of three Phanerochaete chrysosporium endo-1,4-beta-xylanases. *Curr. Genet.* 46(3), 166–175 (2004).

129. Tanaka H, Muguruma M, Ohta K. Purification and properties of a family-10 xylanase from Aureobasidium pullulans ATCC 20524 and characterization of the encoding gene. *Appl. Microbiol. Biotechnol.* 70(2), 202–211 (2006).

130. Luo H, Li J, Yang J, *et al.* A thermophilic and acid stable family-10 xylanase from the acidophilic fungus Bispora sp. MEY-1. *Extrem. Life Extreme Cond.* 13(5), 849–857 (2009).

131. Wang J, Bai Y, Shi P, *et al.* A novel xylanase, XynA4-2, from thermoacidophilic Alicyclobacillus sp. A4 with potential applications in the brewing industry. *World J. Microbiol. Biotechnol.* 27(2), 207–213 (2011).

132. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods*. 8(10), 785–786 (2011).

133. Bai Y, Wang J, Zhang Z, *et al.* A new xylanase from thermoacidophilic Alicyclobacillus sp. A4 with broad-range pH activity and pH stability. *J. Ind. Microbiol. Biotechnol.* 37(2), 187–194 (2010).

134. Solomon V, Teplitsky A, Shulami S, Zolotnitsky G, Shoham Y, Shoham G. Structure-specificity relationships of an intracellular xylanase from Geobacillus stearothermophilus. *Acta Crystallogr. D Biol. Crystallogr.* 63(Pt 8), 845–859 (2007).

135. Gallardo O, Diaz P, Pastor FIJ. Characterization of a Paenibacillus cell-associated xylanase with high activity on aryl-xylosides: a new subclass of family 10 xylanases. *Appl. Microbiol. Biotechnol.* 61(3), 226–233 (2003).

136. Solomon V, Teplitsky A, Shulami S, Zolotnitsky G, Shoham Y, Shoham G. Structure-specificity relationships of an intracellular xylanase from Geobacillus stearothermophilus. *Acta Crystallogr. D Biol. Crystallogr.* 63(Pt 8), 845–859 (2007).

137. Roy N, Okai N, Tomita T, Muramoto K, Kamio Y. Purification and some properties of high-molecular-weight xylanases, the xylanases 4 and 5 of Aeromonas caviae W-61. *Biosci. Biotechnol. Biochem.* 64(2), 408–413 (2000).

138. Stjohn FJ, Rice JD, Preston JF. Paenibacillus sp. strain JDR-2 and XynA1: a novel system for methylglucuronoxylan utilization. *Appl. Environ. Microbiol.* 72(2), 1496–1506 (2006).

139. Waeonukul R, Pason P, Kyu KL, *et al.* Cloning, sequencing, and expression of the gene encoding a multidomain endo-beta-1,4-xylanase from Paenibacillus curdlanolyticus B-6, and characterization of the recombinant enzyme. *J. Microbiol. Biotechnol.* 19(3), 277–285 (2009).

140. St John FJ, Preston JF, Pozharski E. Novel structural features of xylanase A1 from Paenibacillus sp. JDR-2. *J. Struct. Biol.* 180(2), 303–311 (2012).

141. Gallardo O, Pastor FIJ, Polaina J, *et al.* Structural insights into the specificity of Xyn10B from Paenibacillus barcinonensis and its improved stability by forced protein evolution. *J. Biol. Chem.* 285(4), 2721–2733 (2010).

142. Fontes CM, Gilbert HJ, Hazlewood GP, *et al.* A novel Cellvibrio mixtus family 10 xylanase that is both intracellular and expressed under non-inducing conditions. *Microbiol. Read. Engl.* 146 (Pt 8), 1959–1967 (2000).

143. Guo B, Chen X-L, Sun C-Y, Zhou B-C, Zhang Y-Z. Gene cloning, expression and characterization of a new cold-active and salt-tolerant endo-beta-1,4-xylanase from

marine Glaciecola mesophila KMM 241. *Appl. Microbiol. Biotechnol.* 84(6), 1107–1115 (2009).

144. Lee CC, Smith M, Kibblewhite-Accinelli RE, *et al.* Isolation and characterization of a cold-active xylanase enzyme from Flavobacterium sp. *Curr. Microbiol.* 52(2), 112–116 (2006).

145. Mirande C, Mosoni P, Béra-Maillet C, Bernalier-Donadille A, Forano E. Characterization of Xyn10A, a highly active xylanase from the human gut bacterium Bacteroides xylanisolvens XB1A. *Appl. Microbiol. Biotechnol.* 87(6), 2097–2105 (2010).

146. Zhou J, Huang H, Meng K, *et al.* Molecular and biochemical characterization of a novel xylanase from the symbiotic Sphingobacterium sp. TN19. *Appl. Microbiol. Biotechnol.* 85(2), 323–333 (2009).

147. Santos CR, Meza AN, Hoffmam ZB, *et al.* Thermal-induced conformational changes in the product release area drive the enzymatic activity of xylanases 10B: Crystal structure, conformational stability and functional characterization of the xylanase 10B from Thermotoga petrophila RKU-1. *Biochem. Biophys. Res. Commun.* 403(2), 214–219 (2010).

148. Saul DJ, Williams LC, Reeves RA, Gibbs MD, Bergquist PL. Sequence and expression of a xylanase gene from the hyperthermophile Thermotoga sp. strain FjSS3-B.1 and characterization of the recombinant enzyme and its activity on kraft pulp. *Appl. Environ. Microbiol.* 61(11), 4110–4113 (1995).

149. Zhengqiang J, Kobayashi A, Ahsan MM, Lite L, Kitaoka M, Hayashi K. Characterization of a thermostable family 10 endo-xylanase (XynB) from Thermotoga maritima that cleaves p-nitrophenyl-beta-D-xyloside. *J. Biosci. Bioeng.* 92(5), 423–428 (2001).

150. Szilágyi A, Závodszky P. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Struct. Lond. Engl. 1993*. 8(5), 493–504 (2000).

151. Zhou X-X, Wang Y-B, Pan Y-J, Li W-F. Differences in amino acids composition and coupling patterns between mesophilic and thermophilic proteins. *Amino Acids*. 34(1), 25–33 (2008).

152. Ihsanawati, Kumasaka T, Kaneko T, *et al.* Structural basis of the substrate subsite and the highly thermal stability of xylanase 10B from Thermotoga maritima MSB8. *Proteins*. 61(4), 999–1009 (2005).

153. Xie H, Flint J, Vardakou M, *et al.* Probing the structural basis for the difference in thermostability displayed by family 10 xylanases. *J. Mol. Biol.* 360(1), 157–167 (2006).

154. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J. Mol. Graph.* 14(1), 33–38, 27–28 (1996).

155. Pettersen EF, Goddard TD, Huang CC, *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J. Comput. Chem.* 25(13), 1605–1612 (2004).

156. Chang P, Tsai W-S, Tsai C-L, Tseng M-J. Cloning and characterization of two thermostable xylanases from an alkaliphilic Bacillus firmus. *Biochem. Biophys. Res. Commun.* 319(3), 1017–1025 (2004).

157. Zhang G, Mao L, Zhao Y, Xue Y, Ma Y. Characterization of a thermostable xylanase from an alkaliphilic Bacillus sp. *Biotechnol. Lett.* 32(12), 1915–1920 (2010).

158. Gupta N, Reddy VS, Maiti S, Ghosh A. Cloning, expression, and sequence analysis of the gene encoding the alkali-stable, thermostable endoxylanase from alkalophilic, mesophilic Bacillus sp. Strain NG-27. *Appl. Environ. Microbiol.* 66(6), 2631–2635 (2000).

159. Gerasimova J, Kuisiene N. Characterization of the novel xylanase from the thermophilic Geobacillus thermodenitrificans JK1. *Mikrobiologiia*. 81(4), 457–463 (2012).

160. Khasin A, Alchanati I, Shoham Y. Purification and characterization of a thermostable xylanase from Bacillus stearothermophilus T-6. *Appl. Environ. Microbiol.* 59(6), 1725–1730 (1993).

161. Mamo G, Delgado O, Martinez A, Mattiasson B, Hatti-Kaul R. Cloning, sequence analysis, and expression of a gene encoding an endoxylanase from Bacillus halodurans S7. *Mol. Biotechnol.* 33(2), 149–159 (2006).

162. Canakcı S, Cevher Z, Inan K, *et al.* Cloning, purification and characterization of an alkali-stable endoxylanase from thermophilic Geobacillus sp. 71. *World J. Microbiol. Biotechnol.* 28(5), 1981–1988 (2012).

163. Adelsberger H, Hertel C, Glawischnig E, Zverlov VV, Schwarz WH. Enzyme system of Clostridium stercorarium for hydrolysis of arabinoxylan: reconstitution of the in vivo system from recombinant enzymes. *Microbiol. Read. Engl.* 150(Pt 7), 2257–2266 (2004).

164. Lee YE, Lowe SE, Zeikus JG. Gene cloning, sequencing, and biochemical characterization of endoxylanase from Thermoanaerobacterium saccharolyticum B6A-RI. *Appl. Environ. Microbiol.* 59(9), 3134–3137 (1993).

165. Shao W, Deblois S, Wiegel J. A High-Molecular-Weight, Cell-Associated Xylanase Isolated from Exponentially Growing Thermoanaerobacterium sp. Strain JW/SL-YS485. *Appl. Environ. Microbiol.* 61(3), 937–940 (1995).

166. Hung K-S, Liu S-M, Fang T-Y, *et al.* Characterization of a salt-tolerant xylanase from Thermoanaerobacterium saccharolyticum NTOU1. *Biotechnol. Lett.* 33(7), 1441–1447 (2011).

167. Mamo G, Thunnissen M, Hatti-Kaul R, Mattiasson B. An alkaline active xylanase: insights into mechanisms of high pH catalytic adaptation. *Biochimie*. 91(9), 1187–1196 (2009).

168. Han X, Gao J, Shang N, *et al.* Structural and functional analyses of catalytic domain of GH10 xylanase from Thermoanaerobacterium saccharolyticum JW/SL-YS485. *Proteins*. 81(7), 1256–1265 (2013).

169. Verma D, Satyanarayana T. Cloning, expression and applicability of thermo-alkali-stable xylanase of Geobacillus thermoleovorans in generating xylooligosaccharides from agro-residues. *Bioresour. Technol.* 107, 333–338 (2012).

170. Lee YE, Lowe SE, Zeikus JG. Gene cloning, sequencing, and biochemical characterization of endoxylanase from Thermoanaerobacterium saccharolyticum B6A-RI. *Appl. Environ. Microbiol.* 59(9), 3134–3137 (1993).

171. Han X, Gao J, Shang N, *et al.* Structural and functional analyses of catalytic domain of GH10 xylanase from Thermoanaerobacterium saccharolyticum JW/SL-YS485. *Proteins*. 81(7), 1256–1265 (2013).

172. Zolotnitsky G, Cogan U, Adir N, Solomon V, Shoham G, Shoham Y. Mapping glycoside hydrolase substrate subsites by isothermal titration calorimetry. *Proc. Natl. Acad. Sci. U. S. A.* 101(31), 11275–11280 (2004).

173. Atkinson GC, Baldauf SL. Evolution of elongation factor G and the origins of mitochondrial and chloroplast forms. *Mol. Biol. Evol.* 28(3), 1281–1292 (2011).

174. Stam MR, Danchin EGJ, Rancurel C, Coutinho PM, Henrissat B. Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng. Des. Sel. PEDS*. 19(12), 555–562 (2006).

175. Aspeborg H, Coutinho PM, Wang Y, Brumer H, Henrissat B. Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evol. Biol.* 12, 186 (2012).