GH7 cellobiohydrolases structural, functional and evolutional aspects

Anna Borisova
Faculty of Natural Resources and Agricultural Sciences
Department of Molecular Sciences
Uppsala

Doctoral thesis
Swedish University of Agricultural Sciences
Uppsala 2017

Acta Universitatis agriculturae Sueciae
2017:57

Cover: Spot the difference! GH7 cellobiohydrolases *Tre*Cel7A (grey), *Tha*Cel7A (green), *Tat*Cel7A (blue), *Scy*Cel7A (red), *Ddi*Cel7A (orange) and *Dpu*Cel7A (magenta).

ISSN 1652-6880
ISBN (print version) 978-91-7760-004-6
ISBN (electronic version) 978-91-7760-005-3

Print: SLU Service/Repro, Uppsala 2017

GH7 cellobiohydrolases structural, functional and evolutional aspects

Abstract

Fungal glycoside hydrolase family 7 cellobiohydrolases (GH7 CBH) are the workhorses of cellulose degradation and, thus, play a key role in the recycling of biomass on Earth. As central as these enzymes are to biomass degradation, they have become the cornerstone of modern industrial enzyme cocktails for biofuels processes. By combination of X-ray cystallography structure studies and other approaches, this thesis provides essential impact into understanding of sequence, structure and function correlation concepts in the GH7 family.

Several new GH7 CBH structures were solved, ranging from the distant Amoebozoa to *Trichoderma atroviride* Cel7A, which is a close ortholog to the archetypal *Trichoderma reesei* Cel7A. Another ascomycete fungus, *Scytalidium* sp., exhibits new features of GH7 CBHs, never observed before. Namely, *Scytalidium* sp. Cel7A revealed *O*-glycosylation on the tunnel-enclosing B2 loop, and the loop can adopt different conformations, and even bend into the tunnel and obstruct cellulose binding. *Trichoderma atroviride* Cel7A ligand complex with thio-cellotriose represents a sliding intermediate during processive cellulose hydrolysis. Three new structures of GH7 CBHs labeled with a novel mechanism-based affinity tag confirm the proposed inactivation mechanism, presenting covalently bound enzyme-ligand complexes with *Trichoderma reesei* Cel7A and *Scytalidium* sp. Cel7A, and one complex with *Trichoderma reesei* Cel7A E217Q acid/base mutant where covalent bond is not formed. Two crystal structures of social amoeba GH7 CBHs appeared to be very similar to the well-studied fungal *Trichoderma reesei* Cel7A, despite the large evolutionary distance between these organisms. Phylogenetic analysis revealed high similarity between GH7 CBHs from different branches of the eukaryotic tree of life. Biochemical characterization and performance assays of the novel GH7 CBHs along with Molecular Dynamics (MD) simulations based on their structures, highlight important loop regions implicated in processivity, binding in the product sites, endo-/exo-initiation as well as thermostability and initiation of thermal unfolding. Reverse concervation analysis (RCA) identified potentially important evolutionary target sites in *Trichoderma* spp GH7 sequences. Loop dynamics and correlation with structure and function of GH7 CBH is thoroughly discussed in the thesis.

Keywords: cellulose, cellobiohydrolase, X-ray structure, molecular dynamics, Dictyostelium, Trichoderma, Scytalidium,

Author's address: Anna Borisova, SLU, Department of Molecular Sciences,
P.O. Box 7015, 750 07 Uppsala, Sweden

Felix, qui potuit rerum cognoscere causas.
Virgil

Contents

List of publications

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Hobdey, S.E.; Knott, B.C.; Momeni, M.H.; Taylor, L.E.; **Borisova, A.S.**; Podkaminer, K.K.; VanderWall, T.A.; Himmel, M.E.; Decker, S.R.; Beckham, G.T.; Ståhlberg, J. (2016). Biochemical and structural characterization of two *Dictyostelium* cellobiohydrolases from the *Amoebozoa* kingdom reveal a high conservation between distant phylogenetic trees of life. *Applied and Environmental Microbiology,* **82**(11):3395-409.

II. **Borisova, A.S**.; Eneyskaya, E.V.; Jana, S.; Badino S.F.; Kari, J.; Amore, A.; Karlsson, M.; Hansson, H.; Sandgren, M.; Himmel, M.E.; Westh, P.; Payne, C.M.; Kulminskaya, A.A.; Ståhlberg, J. Correlation of structure, function and protein dynamics in GH7 cellobiohydrolases from *Trichoderma atroviride, T. reesei and T. harzianum*. (Manuscript).

III. **Borisova, A.S**; Eneyskaya, E.V.; Bobrov, K.S.; Jana, S.; Logachev, A.; Polev, D.E.; Lapidus, A.L.; Ibatullin, F.M.; Saleem, U.; Sandgren, M.; Payne, C.M.; Kulminskaya, A.A.; Ståhlberg, J. (2015). Sequencing, biochemical characterization, crystal structure and molecular dynamics of cellobiohydrolase Cel7A from *Geotrichum candidum 3C*. *FEBS Journal*, **282**(23):4515-37.

IV. **Borisova, A.S**.; Hansson, H.; Rasmussen, T.S.; Zierke, M.; Withers, S.G.; Ståhlberg, J. Crystal structures of mechanism-based affinity labelled GH7 cellobiohydrolases. (Manuscript).

Other publications

1. **Borisova, AS***; Isaksen, T*; Dimarogona, M; Kognole, AA; Mathiesen, G; Varnai, A; Rohr, ÅK; Payne, CM; Sorlie, M; Sandgren, M. (2015). Structural and functional characterization of a lytic polysaccharide monooxygenase with broad substrate specificity. *Journal of Biological Chemistry*, **290(**38):22955-22969.

2. **Borisova, AS**; Ivanen, DR; Bobrov, KS; Eneyskaya, Elena V; Rychkov, GN; Sandgren, M; Kulminskaya, AA; Sinnott, ML; Shabalin, KA (2015). Alpha-Galactobiosyl units: thermodynamics and kinetics of their formation by transglycosylations catalysed by the GH36 alpha-galactosidase from *Thermotoga maritima*. *Carbohydrate research*, **401**:115-121.

3. **Borisova, AS**; Reddy, SK; Ivanen, DR; Bobrov, KS; Eneyskaya, EV; Rychkov, GN; Sandgren, M; Stålbrand, H; Sinnott, ML; Kulminskaya, AA (2015). The method of integrated kinetics and its applicability to the exo-glycosidase-catalyzed hydrolyses of p-nitrophenyl glycosides. *Carbohydrate research*, **412**: 43-49.

4. Bobrov, KS*; **Borisova, AS***; Eneyskaya, EV; Ivanen, DR; Shabalin, KA; Kulminskaya, AA; Rychkov, GN (2013). Improvement of the efficiency of transglycosylation catalyzed by alpha-galactosidase from *Thermotoga maritima* by protein engineering". *Biochemistry (Moscow)*, **78**(10):1112-1123.

*First authorship shared

The contribution of Anna Borisova to the papers included in this thesis was as follows:

I. Performed the phylogenetic analysis, and contributed with associated interpretation, figures and paper writing. Took part in writing together with other co-authors.

II. Planned the work together with other co-authors. Coordinated the experimental work. Conducted almost all experiments, including protein purification, biochemical characterization, crystallization, structure determination, and RCA analysis; except for performance assays on pretreated corn stover and MD simulations. Compiled all the results. Took leading role in interpretation and paper writing..

III. Planned the work together with other co-authors. Conducted enzyme kinetics experiments and all X-ray crystallography work. Compiled all the results. Took leading role in interpretation and paper writing.

IV. Planned the work together with other co-authors. Conducted all lab work and protein structure studies. Compiled all the results. Took active role in interpretation and paper writing together with other co-authors.

Abbreviations

BCA	Biocontrol agent
BMCC	Bacterial microcrystalline cellulose
CBH	Cellobiohydrolase
CBM	Carbohydrate binding module
CD	Catalytic domain
CMC	Carboxymethyl cellulose
*Ddi*Cel7A	*Dictyostelium discoideum* Cel7A
*Dpu*Cel7A	*Dictyostelium purpureum* Cel7A
EC	Enzyme classification
EG	Endo-1,4-β-D-glucanase
*Gca*Cel7A	*Geotrichum candidum/Scytalidium sp* Cel7A, identical to *Scy*Cel7A
GH	Glycoside hydrolase
HGT	Horizontal gene transfer
*Hje*Cel7A	*Hypocrea jecorina* Cel7A, identical to *Tre*Cel7A
MD	Molecular dynamics
PASC	Phosphoric acid swollen cellulose
PCS	Pretreated corn stover
*p*NP-Lac	*p*-nitrophenyl β-lactoside
RCA	Reverse conservation analysis
RMSD	Root mean square deviation
RMSF	Root mean square fluctuation
*Scy*Cel7A	*Scytalidium sp/Geotrichum candidum* Cel7A, identical to *Gca*Cel7A
*Tat*Cel7A	*Trichoderma atroviride* Cel7A
*Tha*Cel7A	*Trichoderma harzianum* Cel7A
*Tre*Cel7A	*Trichoderma reesei* Cel7A, identical to *Hje*Cel7A
TSRG14	Mechanism-based affinity tag

1 Introduction

There is an enormous amount of different enzymes in the world. These molecules are very powerful catalysers of myriads of biochemical reactions accompanying every biological process in nature. Humanity has always been striving to rule the world and one way to do that is to understand the way how the world is built. A lot of fermentative processes have been known and successfully used for ages (for example, brewery, bakery, cheese production), but the underlying molecular mechanisms were secret. Early enzyme discoveries started already in 1835 by the Swedish chemist Jon Jakob Berzelius who termed their chemical action catalytic. In the beginning of the 20th century enzymology was established as an independent scientific field and since then attracted a significant burst of interest from society, resulting in thousands of various proteins being characterized. All these data including sequence, structure and function of the enzymes is now organized in various databases, to shape our knowledge and to provide information for useful application of these molecules in industry, medicine and biotech.

Cellulose-degrading enzymes are widely used in modern technologies and constitute the third largest group of industrial biotech enzymes (Bischof et al, 2016). Fundamental studies of cellulases have been driven mostly by industrial needs, such as cotton processing, paper recycling, development of detergents and animal feed additives. However, the use cellulose-degrading enzymes for large-scale biomass-to-fuel conversion has been brought forward as the most promising prospect, which is now at the verge of being realized. The concept of sustainable development became a hot topic in political and social discussions since it was declared in 1992 at the United Nations Conference on Environment and Development held in Rio de Janeiro. As part of a program of developing renewable fuels, research on cellulases and other biomass-degrading enzymes expanded dramatically during the last 30 years.

Second generation biofuels are produced from plant biomass. The main component of plant biomass is cellulose, supported by essential amounts of

hemicellulose and lignin, which provide cell wall strength and flexibility. Bioethanol production from lignocellulosic biomass can be divided into three steps: (i) pretreatment of lignocellulosic biomass, (ii) depolymerisation of cellulose (and hemicellulose) to soluble sugars and (iii) fermentation of released sugars to ethanol. In plant cell walls, cellulose and hemicellulose are surrounded by lignin; thus, in order to make cellulose accessible for the cellulose-degrading enzymes, pretreatment is an absolute requirement. The most promising option for pretreatment is to combine chemical and physical manipulations, but there is a desire to use milder conditions to prevent formation of byproducts, that may be inhibitory in downstream processes. The most promising physicochemical methods to increase further enzymatic hydrolysis of cellulose can be divided as following: alkali pretreatment, organosolv pretreatment, acid pretreatment, hydrothermal pretreatment and pretreatment with ionic liquids. The main goal of alkaline and organosolv pretreatments is to remove lignin from the cell wall. During acid and hydrothermal pretreatment mainly hemicellulose is dissolved, while pretreatment with ionic liquids is directed to make cellulose more amorphous and porous.

In bioreactors, cellulose saccharification is conducted by enzyme cocktails, where the major component is glycoside hydrolase family 7 cellobiohydrolase, which can degrade crystalline cellulose in a processive manner. In nature these enzymes are expressed predominantly in fungi, which have developed efficient machinery for degradation of recalcitrant lignocellulosic material. Glycoside hydrolase family 7 cellobiohydrolase enzymes are thus key enzymes in the global carbon cycle.

The filamentous fungus *Trichoderma reesei* was discovered for its remarkable biomass degrading potential over 70 years ago and since then it was thoroughly studied, becoming the archetypal enzyme of the cellulose-degrading machinery in fungi. Nowadays industrial strains of *Trichoderma reesei* are widely used for batch production of enzymatic cocktails for saccharification process of pretreated biomass (Bischof et al, 2016). Optimization of these cocktails is still one of the key factors determining the cost performance of cellulosic ethanol processes and finding efficient enzyme formulations is an economically driven process. Site-directed mutagenesis and strain developments are done by researchers in undustry and academia in order to improve the degradation process. This research is accompanied by extensive search for better cellulases, already designed by nature. Understanding the diversity of these key enzymes is critical to engineer them for higher levels of activity and greater stability.

In this thesis, the key enzymes of biomass degradation, fungal cellobiohydrolases from glycoside hydrolase family 7, are studied in scope of

evolution. Novel structural data of these enzymes are combined with biochemical characterization and computational analysis, including molecular dynamics and molecular evolution analysis.

2 Background

2.1 Cellulose

Cellulose is a polysaccharide consisting of β-1,4-linked β-D-glucose units in a linear chain with a reducing and a non-reducing end (Figure 1). The cellulose chain can contain 100-15,000 glucose units depending on the plant tissue and species. Cellulose is known as a recalcitrant biopolymer with strong covalent bonds (β-1,4-*O*-glycosidic linkage) at atomic level, and is estimated to have a half-life of 4.7 billion years at room temperature (Wolfenden and Snider, 2001).

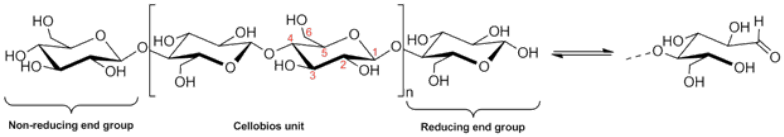


Figure 1. The molecular structure of a cellulose polymer where the cellobiose is the smallest repeating unit in the polymer. Carbon atoms in glucose residue are numbered from 1 to 6 (marked in red). The reducing end group can be either a free hemiacetal or an aldehyde. Figure was taken from Börjesson (2015).

Cellulose chains are bound together with hydrogen bonds and van der Waals interactions into crystalline microfibrils, which form the mechanical framework of cell walls (Somerville et al., 2004, Carpita and Gibeaut, 1993). There are several types of cellulose crystalline forms: cellulose I (Iβ and Iα polymorphs), cellulose II and cellulose III (Figure 2) (Payne et al., 2015). Natural systems, including plants, bacteria, algae and oomycetes produce cellulose type I, which is described as a regular crystalline structure comprised by parallel-oriented chains, with only intralayer hydrogen bonding (Gardner and Blackwell, 1975). It was

shown that native cellulose is a composition of polymorphs Iβ and Iα, and these are distinct in hydrogen bonding patterns, as well as interlayer stacking arrangement (Atalla and Vanderhart, 1984, Nishiyama et al., 2002, Nishiyama et al., 2003). Cellulose II and cellulose III are synthetic crystalline forms, obtained via certain chemical treatment from cellulose I. These types of cellulose represent significant difference in their stacking configuration, where hydrogen bonding occurs between sheets as well as across the layers (Nishiyama et al., 2010, Langan et al., 2001). Chemically pretreated cellulose is used in bioreactors, since it typically exhibits greater digestibility by cellulose enzymes (Chundawat et al., 2011a, Chundawat et al., 2011b).

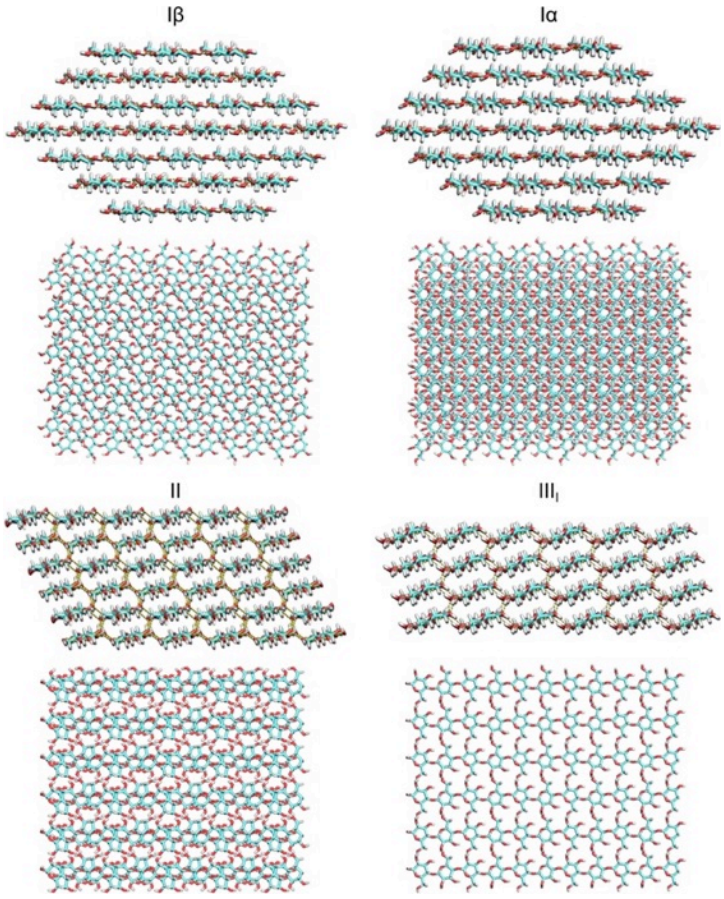


Figure 2. Natural and synthetic cellulose polymorphs. The differences between the two polymorphs Iα and Iβ are most easily observed from the "top down" view, which illustrates the subtle differences in interlayer chain stacking. Celluloses II an d III$_I$, the result of chemically pretreating cellulose I, are significantly different in their chain stacking arrangement. Hydrogen bonding, shown in yellow, occurs between sheets as across the layers. Figure was taken from Payne et al. (2015).

2.2 Cellulosic substrates

Fundamental research on cellulose-degrading enzymes is usually carried out with different types of cellulosic substrates, instead of raw biomass material. Cellulosic substrates differ in particle size, pore-size, accessible surface area and degree of polymerization (DP). Avicel and bacterial microcrystalline cellulose (BMCC) are mostly used in cellulose digestion experiments as representatives of crystalline cellulose. Avicel is commercially available microcrystalline cellulose derived from wood pulp. Cellulose produced by green algae and some bacteria is also available, but must be purified before use in enzyme assays. After chemical treatment bacterial microcrystalline cellulose (BMCC) demonstrates microfibrils around 100 times thinner than microfibrils found in plant cell walls, and lower DP, and can be used successfully in processivity experiments (Valjamae et al., 1999). Phosphoric acid swollen cellulose (PASC), sometimes referred to as amorphous cellulose, has undergone phosphoric acid treatment in order to make cellulose chains more easily accessible for the enzymes (Walseth, 1952). PASC is widely used as a test substrate for cellobiohydrolases, while chemically modified carbomethyl cellulose (CMC) is primarily used as substrate for endoglucanases.

2.3 Role of cellulose synthesis and deconstruction in evolution of cell walls

Plant biomass on Earth is mainly cell wall material, which in turn consist of 40-50% cellulose. Considering the enormous biochemical and phylogenetic diversity of living systems, it seems remarkable that a certain compound, cellulose, should contribute so significantly to most of the biomass on earth. Due to their functional importance cellulose microfibrils were proposed to be a critical factor in the evolution of modern plant life (Duchesne and Larson, 1989).

Present evidence indicates that primitive cell walls appeared about 3.5 billion years ago in ancient microbes, the ancestors of modern cyanobacteria, that were the first organisms producing O_2 via oxygenic photosynthesis. During evolution organisms with cell walls obtained increased survival ability due to development of cell osmosis, that led to greater metabolic rates and capacity for faster growth and colonization of hypotonic environments. The random mutations responsible for cell wall development probably resulted in an initial diversity in the chemical nature of these walls. Cell wall designs in

two prokaryotic domains, bacteria and archaea, are commonly considered to evolve from a common wall-less ancestor (Kandler, 1994), whereas cell walls in eukaryotes have evolved by lateral gene transfer from previously established cell wall-producing organisms during primary or secondary endosymbiosis (Niklas, 2004).

Besides clear differences in cell wall design between eukaryotes and prokaryotes, remarkable diversity can also be seen within the eukaryotic kingdoms (Niklas, 2004). In scope of cell wall evolution, the lateral gene transfer between ancestral cyanobacterial endosymbionts and the nuclei of their host cells may explain the distribution of cellulose biosynthesis among some of the most ancient algal lineages. Correspondingly, secondary endosymbiotic events may explain why some, more recently derived protist lineages possess the ability to synthesize cellulose. For example, amoebozoa, oomycetes and tunicates have diverse types of cell walls, with compositions similar to those of their respective plant or fungal descendants, while probable ancestors of animals, protozoans, lack cell walls. The land plants (embryophyta) are believed to have evolved from green algae and the closest ancestor is thought to be the group Charophyta, cell walls of these were based on cellulose (Lewis and McCourt, 2004, Popper and Fry, 2003). In these algae, cellulose is synthesized by cellulose synthases, but they lack xyloglucan-modifying enzymes, since no xyloglucan was found in charophyta (Sarkar et al., 2009). Later, bryophytes developed xyloglucan-modifying enzymes and some peroxidase enzymes, which are typically involved in lignin biosynthesis, even though lignin was not commonly present is bryophytes. In vascular plants, primary and secondary cell wall were differentiated, and pectinases and mannonases were developed, as well as lignin modifying enzymes. Tall woody plants invented high amounts of mannose-rich hemicelluloses and lignin, which provide not only mechanical strength, but also an advanced level of protection from cell wall-degrading enzymes secreted by pathogens (Sarkar et al., 2009). Through evolution, plants developed powerful mechanisms of cellulose synthesis and metabolism, which made them extremely successful in colonisation of land. Development of specific enzymes is an evolutionary process towards improvement of cell wall properties for survival needs of a certain plant. Enzymes are responsible for cellulose synthesis and utilization, and cell wall development, remodelling and breakdown. There is a huge variety of enzymes involved, for example, in Arabidopsis and poplar around 1000–2500 are estimated, including impressive numbers of glycoside hydrolases and glycosyl transferases, and the composition correlates directly with evolutionary relationships between certain species (Frankova and Fry, 2011).

In contrast to plants, fungal cell walls primarily consist of chitin and beta-1,3-glucans, which place them into different eukaryotic kingdom from plants and animals. Furthermore fungi are heterotrophic like animals, but they are stationary like plants. Chitin is thought to be weaker than cellulose, but provides enough mechanical support for the small fungal plant bodies; it comprises a cell wall, which encloses the mycellium (Sarkar et al., 2009). Cell walls of oomycetes contain cellulose and lack chitin while hyphochytrids have both. Slime molds lack a cell wall during the assimilative phase.

2.4 Evolution of cellulose-degrading organisms

There is a large diversity of organisms using cellulose in their cell walls across the kingdoms of life Cellulose is the most abundant food source on Earth and thus many organisms have evolved the necessary enzymatic machinery for degrading cellulose to soluble sugars as food and energy source. Cellulose-degrading organisms are wide spread among prokaryotes and eukaryotes, and represent diverse mechanisms of lignocellulose deconstruction (Figure 3). Among prokaryotes, bacteria and archaea from termites gut express a number of wood-degrading cellulases. Free-living wood degrading prokaryotes from marine sources (e.g. tunnelling bacteria) are efficient cellulose and hemicellulose degraders (Cragg et al., 2015). In eukaryotes endogenous cellulases are found in oomycetes and some free-living protists, such as slime molds, whose primary role may be in endogenous cellulose metabolism due to their cellulosic cell wall organisation. Genes of cellulose-degrading enzymes are found in plant-parasitic nematodes, but also unambiguously demonstrated in other taxa, such as insects, gastropoda, crustacea and annelida (Cragg et al., 2015). Tunicates are unusual among animals in that they produce a large fraction of their tunic and some other structures in the form of cellulose. However, cellulose processing was lost in the ancestor cells of animal kingdom. Through evolution animals developed a modified form of storage called glycogen that can quickly replace glucose, when it becomes depleted in the cell from activity.

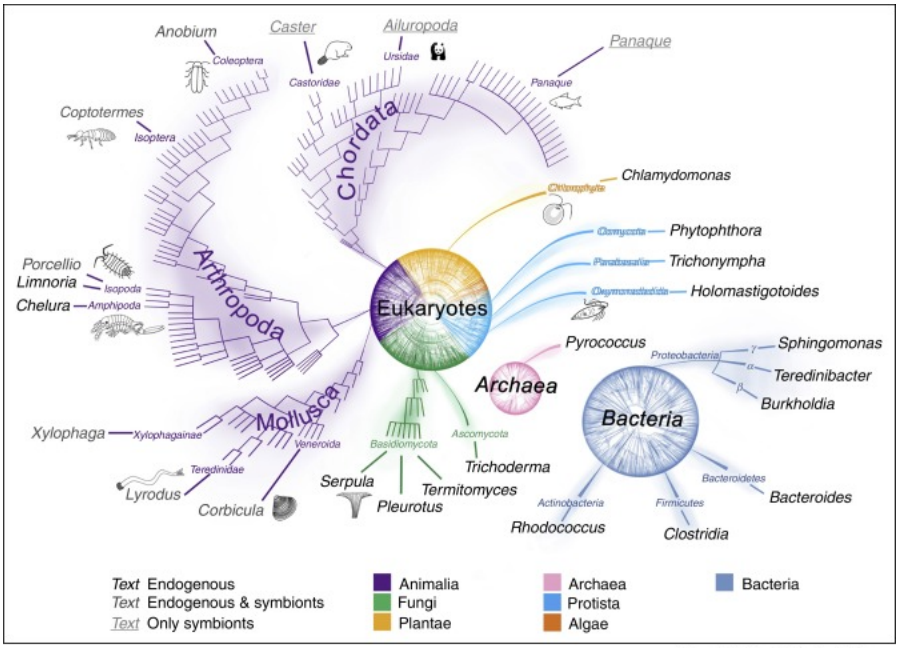


Figure 3. The sparse and localised distribution of selected organisms capable of lignocellulose or cellulose degradation mapped onto the Tree of Life, with highest taxonomic ranks colour-coded as shown in key. Genus names of organisms degrading lignocellulose using endogenous enzymes shown in bold, those with endogenous plus symbiont-derived enzymes shown printed pale and those with only symbiont-derived enzymes shown underlined. Figure was taken from Cragg et al. (2015).

2.5 Fungi are main lignocellulose degraders

Fungi are known to be predominant degraders of lignocellulose biomass in nature and play a key role in global carbon recycling on Earth. As decomposers, they play an essential role in nutrient cycling, degrading organic matter to inorganic molecules, for uptake by plants and other organisms (Lindahl et al., 2007). In general fungi employ two degradation modes: i) direct enzymatic depolymerization, for example, by cellobiohydrolases and ii) generation of oxidative species (e.g., radicals) that act on the biomass (Cragg et al., 2015).

By the type of lignocellulose decay, fungi can be classified into brown rot, soft rot and white rot. Each type produce different enzymes, degrade different plant materials, and colonise different environmental niches. Brown-rot fungi initially utilize Fenton chemistry to generate hydroxyl radicals, which attack plant cell walls via powerful oxidation reactions (Eastwood et al., 2011). Filamentous fungi characterized as soft-rots and white-rots have been long

known to primarily employ enzymatic means to break down biomass (Cragg et al., 2015).

Degradation of plant cell walls occurs outside of the fungus by extracellular enzymes, which cooperate synergistically in action on recalcitrant lignocellulosic material (Chundawat et al., 2011a). These enzymatic cocktails usually contain several types of enzymes involved in cellulose degradation. Copper-dependent lytic polysaccharide monooxygenases (LPMO) can act at the surface of insoluble cellulose and cleave *O*-glycosidic bonds by an oxidative mechanism. Endoglucanases (EGs) bind cellulose chains and hydrolyse internal bonds. Cellobiohydrolases (CBHs) bind preferentially to reducing or non-reducing ends of cellulose chains and cleave off cellobiose units in processive manner. Finally, β-glucosidases hydrolyse cellobiose and cellooligosaccharides to glucose.

2.6 Evolution of fungi

Fungi have colonized a vast range of terrestrial and marine environments, being very important components in most ecosystems. They are known for their symbiotic relationship with organisms from nearly all kingdoms, mutualistic, antagonistic or commensal nature. The fungal kingdom contains an enormous diversity of species; it was estimated to include from 1.5 to 5 million species, with only about 5% being formally classified so far. Fungi have been classified in major divisions mainly on the basis of characteristics of their sexual reproductive structures, including Microsporidia, Chytridiomycota, Blastocladiomycota, Neocallimastigomycota, Glomeromycota, Ascomycota, and Basidiomycota.

The main decomposers of plant biomass are Ascomycota and Basidiomycota, comprising the subkingdom Dikarya, often referred to as higher fungi. In contrast to Dikarya, the lineages that diverged early in fungal evolution use diverse sources of nutrition, being largely associated with animals or animal products. The question is when plant-associated fungi may have originated and how to trace evolution of fungi according to their nutritional modes (Chang et al., 2015).

Fungi diverged from metazoans around 800 Mya (million years ago), long before land plants diverged from green algae. However the age of the ancestor of terrestrial fungi was dated about 700 Mya and most of the diversification within terrestrial fungal phyla occurred within the last 500 Mya (Chang et al., 2015). Since the fossil record of the plant lineage is rich, evolution of terresial fungi can be linked to evolution of streptophytes and developments of the plant cell wall, particularly, pectin and lignin (see paragraph 2.3). Following plant

cell wall evolution, fungi developed efficient lignocellulose-degrading machinery. It is important to note that the presence of cellulases as such does not serve as a good marker for an association with land plant lineage, since cellulose is widely distributed across organisms including green, red, and brown algae. Pectinases are better in that sense; pectin-containing streptophytes are estimated to be no older than 750 Ma, so the pectin-degrading common ancestor of the Chytridiomycota and Dikarya is probably no older than 750 Ma (Chang et al., 2015). Subsequent development of lignin-degrading enzymes was a major invention in fungi and made a revolution in terrestrial life (Floudas et al., 2012). The only organisms capable of substantial lignin decay are white rot fungi in the Agaricomycetes, whose lignin-degrading peroxidases were shown to originate from a common ancestor. Molecular clock analyses suggest that the origin of lignin degradation correlates with the sharp decrease in the rate of organic carbon burial around the end of the Carboniferous period, around 300 Mya. It was also shown that white rot fungi genomes have about two-fold more genes encoding cellulose-degrading enzymes, than brown rot fungi (Floudas et al., 2012).

2.7 Organisms under investigation

2.7.1 Trichoderma spp

Trichoderma spp. are free-living fungi commonly found in soil and root ecosystems, representing wide range of functions and behaviour. They are opportunistic, avirulent plant symbionts, as well as being parasites of other fungi. At least some *Trichoderma* strains build long-lasting colonisations of root surfaces and penetrate into the epidermis, producing a variety of compounds that induce localized or systemic resistance responses, and this explains their lack of pathogenicity to plants. These root–microorganism associations cause substantial changes to the plant proteome and metabolism. Root colonization by *Trichoderma* spp. also frequently enhances root growth and development, crop productivity, resistance to abiotic stresses and the uptake and use of nutrients (Harman et al., 2004). *Trichoderma* spp. were also shown to have toxigenic potential (McMullin et al., 2017).

In the current study, closely related *Trichoderma* species were studied in scope of their expressed GH7 CBHs. *Trichoderma reesei* Cel7A (also known as CBH1 and CBH I) is probably the most extensively studied cellulase and was the first GH7 enzyme whose structure was solved and it is the archetype of GH7 CBHs (Divne et al., 1994). Thus, *T. reesei* (strain QM9414) was chosen

for investigation, together with the close relatives *T. harzianum* (strain IOC-3844) and *T. atroviride* (strain IOC 4503).

The discovery and history of the filamentous fungus *T. reesei* (syn. *Hypocrea jecorina*) started more than 70 years ago. The original *Trichoderma* strain was isolated from rotting US Army equipment on the Solomon Islands during World War II. By 1955, research at the Natick Army Research Laboratories led by M. Mandels and E.T. Reese turned the destructive potential of the fungi into a great advantage (Allen et al., 2009). The research initially driven by US Army turned into fundamental studies of the multicomponent nature of the cellulase enzyme complex and the concept of enzymatic saccharification of cellulose through synergy of different enzymes (Reese, 1956). The large-scale investigation involved already 14,000 strains in the Quartermaster Fungal Collection including *Trichoderma* sp QM6a (later *T. reesei* in honour of E.T. Reese). Already in the 1970s efficient strain mutagenesis was developed along with screening procedures towards industrial application of *T. reesei*. The strain QM6a is the one from which all the mutants used in industry today have been derived. The mutated strain RUT-C30 is among the most prolific protein producing organisms that are publicly available. RUT-C30 is still the prototype cellulase hyperproducer with protein expression reaching 30 g/L, although yields of over 100 g/L with recent industrial strains have been reported (Bischof et al., 2016).

The world-wide installed cellulosic biofuel production capacity is now 480.5 million liters per year (MMLY) of ethanol of which 380.5 MMLY (or roughly 80 %) are produced using *T. reesei* enzyme preparation such as Accellerase and Cellic (Figure 4). Nowadays, different high throughput methods, such as comparative genomics together with CRISPR/Cas9 system, provide efficient ways to rationally engineer and design this hyperproducing cell factory not only for regulation and expression of cellulases, but also for a broader range of enzymes and other proteins (Bischof et al., 2016).

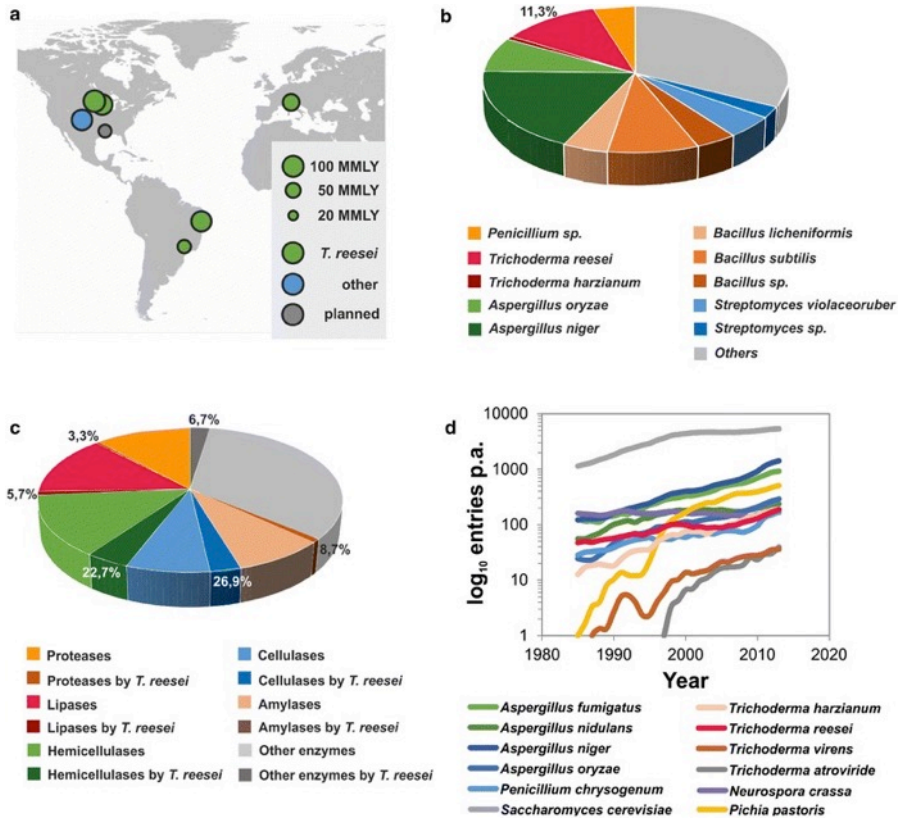


Figure 4. Industrial use of enzymes. a - Installed and planned cellulosic ethanol production as of April 2015 in million liter per year (MMLY). b - Number of different technical enzyme preparations produced by individual species. c Number of a given type of enzyme produced by *T. reesei* (darker color) or other fungi (lighter color). d Number of research papers per year for different fungi retrieved by a Scopus search with the species name as the entry. Results were averaged over 3 year intervals to reduce the effect of random fluctuation. Figure was taken from Bischof et al. (2016).

Trichoderma harzianum was shown to be an efficient antagonist active against pathogenic fungi already in the late 1970s and is nowadays widely used as a biological control agent, for foliar application, seed treatment and soil treatment for suppression of various disease causing fungal pathogens (Harel et al., 2014). However, quite recently *Trichoderma harzianum* has also revealed its potential for cellulase production and cellulosic ethanol applications (de Castro et al., 2010). Today it is used for industrial manufacturing of enzymes and biomass hydrolysis (Figure 4).

Trichoderma atroviride is a filamentous cosmopolitan fungus, commonly found in soil and isolated from both tropical as well as temperate climates. This and related mycoparasitic fungi have been widely studied for their capacity to

produce antibiotics, parasitize other fungi and compete with deleterious plant microorganisms (Liu et al., 2004). Though it has been shown that some *Trichoderma atroviride* mutant strains are capable of secreting high levels of cellulases and beta-glucosidases (Grigorevski-Lima et al., 2013, Kovacs et al., 2009), studies specifically addressing cellulolytic activity by enzymes from *Trichoderma atroviride* are rare.

2.7.2 *Scytalidium* spp

The ascomycete *Scytalidium* is a genus of anamorphic fungi in the Helotiales order, which contains 18 species and is considered to be widespread. These filamentous fungi occur predominantly in nature in soil and on decaying wood, but are sometimes described as opportunistic plant pathogens. Some species have been determined as causative agents of human dermatomycosis-like infections and foot infections predominantly in tropical areas.

The GH7 CBH described in paper III under the acronym *Gca*Cel7A, was isolated from a fungal strain named *Geotrichum candidum* strain 3C at that time. The organism was originally placed into *Geotrichum* based on morphological identification. However, more recent genomics analyses underway of this fungal strain rather point towards *Scytalidium* sp. (and not *Geotrichum*), but the exact *Scytalidium* species is not established yet (Kulminskaya et al, manuscript in preparation). The same GH7 CBH protein was later used in paper IV, where it was renamed accordingly, i.e. to *Scytalidium* sp Cel7A (*Scy*Cel7A). In the following, the new name *Scy*Cel7A will be used in most instances. In any case *Gca*Cel7A (paper III) is identical to *Scy*Cel7A (paper IV). The *Geotrichum candidum* strain 3C, from now *Scytalidium sp*, was studied and used in USSR in the 1980s, but was then undeservedly forgotten for decades. The filamentous yeast-like *Scytalidium* sp./*Geotrichum candidum* strain 3C was isolated from a rotting rope and was found to have high cellulolytic and xylanolytic activities already in early 1970s (Rodionova, 1988) The cellulases of *Scytalidium sp* were initially characterized and it was shown that the *Scytalidium sp* cellulase complex was more efficient than that of well-studied *Trichoderma* sp. An enzyme preparation from this fungus, 'Cellokandin G10x', has been patented and used in the pulp and paper industry for waste paper utilization, and applied research has been carried out on *Scytalidium sp* to improve the process of bleaching of softwood and hardwood kraft pulp (Paper III).

2.7.3 *Dictyostelium* spp

Dictyostelium is traditionally known as a slime mold, but rather belongs to the class of social amoebae, containing four different groups of terrestrial bacterivores from the kingdom Amoebozoa. They are present in most terrestrial ecosystems as a normal component of the soil microflora, and play an important role in the maintenance of balanced bacterial populations in soils. During vegetative growth, Dictyostelia exist as single-celled organisms; upon starvation, a lack of nutrients become preventive for vegetative growth, and the cells aggregate into a multi-cellular slug. Slugs have the ability to migrate, are sensitive to light and temperature, and exhibit an innate immune system. When conditions are sufficiently severe, the slug can form a fruiting body where cells differentiate into spore and stalk. During the formation of the slug and fruiting body, proteins and cellulose are deposited as an extracellular matrix, providing the organism with environmental protection and structural rigidity (Wang et al., 2001, Freeze and Loomis, 1977b, Freeze and Loomis, 1977a). Cellulose is also found in the sheath that surrounds the cell aggregates, and is deposited in the stalk, stalk cell walls, and spore coats. Thus, the deposition and reorganization of cellulose upon morphogenesis into the fruiting body is crucial to the development and propagation of the organism (Freeze and Loomis, 1977a, Zhang et al., 2001). The genomes of *D. discoideum* and *D. purpureum* exhibit >40 genes related to cellulose synthesis and hydrolysis, including a single GH7 CBH encoding gene. Expression of the GH7 gene coincides with cellulose synthesis, with the highest levels occuring during formation of the fruiting body. Thus, the function of the enzyme is believed to be in endogenous cellulose metabolism rather than food acquistion from plant material.

2.8 Carbohydrate active enzyme database

Since 1991 carbohydrate active enzymes have been organised into families based on sequence, structure and their catalytic activities in the Carbohydrate Active Enzymes database (CAZy, www.cazy.org). The CAZy database currently contains over 300 protein families and several hundred thousands of sequences and is continuously updated as new proteins are functionally characterized and structurally determined (Lombard et al., 2014). The database includes six classes of carbohydrate active enzymes: Glycoside hydrolases (GH), Glycosyl transferases (GT), Polysaccharide Lyases (PL), Carbohydrate Esterases (CE), Auxiliary activities (AA) and Non-catalytic carbohydrate-binding modules (CBM).

Glycoside hydrolases (EC 3.2.1.-) are enzymes that hydrolase the glycosidic bond between two or more carbohydrates or between a carbohydrate and non-carbohydrate moiety. GHs are widespread in nature and CAZy includes 144 GH families up to date. The vast majority employ either retaining or inverting hydrolytic mechanisms (Koshland, 1953), depending on their catalytic residues and active site configuration. Retaining mechanisms proceed via two-step, where first step is called glycosylation and the second step is deglycosylation; and the anomeric center of the reaction undergoes double-reversion, resulting in retaining of stereochemistry (Figure 5A). Inverting mechanisms proceed via a single catalytic step, so the stereochemistry at the anomeric center is reverted (Figure 5B).

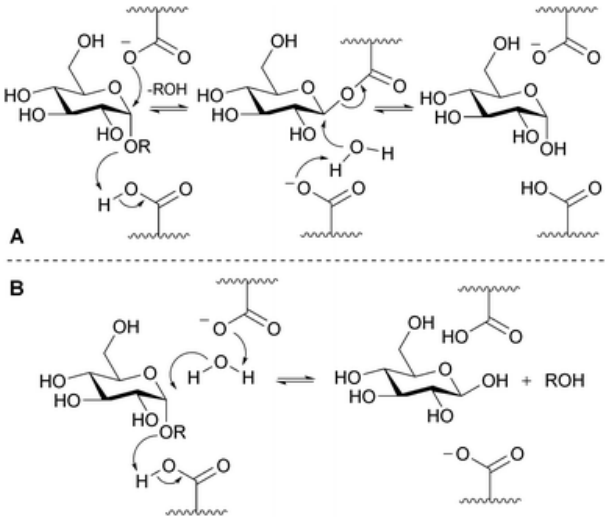


Figure 5. GH catalytic reaction mechanism. A - retaining stereochemistry of anomeric center; B – inverting stereochemistry of anomeric center.

Enzymes from the same GH family utilize the same mechanism. Otherwise GH families are based on protein sequence and structure and do not necessarily reflect specific activity. Thus, most of the families comprise proteins with different activities and the same enzymatic activity can be found in different GH families. For example, endoglucanases (EC 3.2.1.4) occur in 17 different GH families, while the largest family, GH5, includes 20 different enzymatic activities (Busk et al., 2014).

2.9 Fungal cellulases

Cellulases are glycoside hydrolase (GH) enzymes, besides recently discovered LPMOs, which are nowadays placed into the class Auxiliary activities (AA). The definition of cellulases as "enzymes active on cellulose" is quite rough and their classification is constantly updating; in general these enzymes are classified as acting on β-1,4-glycosidic bonds in cellulose. Practically, cellulases include two types of enzymes: EGs (EC 3.2.1.4) and CBHs (EC 3.2.1.91; EC 3.2.1.176), which differ by active site architecture and mode of action. EGs are predominantly endo-acting enzymes, while CBHs exhibit more exo-action, starting their hydrolytic action from the end of the cellulose chain. Cellulases are found in 17 GH families in the CAZy database. Notably, some ancestor enzymes were developed only in fungi and some only in bacteria. For example, GH family 8, 26, 44 and 124 cellulases are found only in bacteria, while GH7 cellulases are not found in prokaryotic organisms, but mostly in fungi. Primary components of fungal cellulolytic cocktails include GH family 5, 6, 7, 12, 45 cellulases (Payne et al., 2015).

Many fungal cellulases exhibit a bimodular architecture, where the catalytic domain (CD) of a certain GH family is connected to a binding module (CBM) via a highly glycosylated flexible linker. In fungi, the binding module usually belongs to CBM1 family and apparently helps the enzyme to bind and hold on to the cellulose surface (Stahlberg et al., 1991). The catalytic domains of CBHs and EGs can have similar overall structure, but have significant differences in the cellulose-binding region. EGs usually have their active site in an open cleft, which makes them able to access cellulose internal regions and cut the chain into shorter pieces, releasing more free ends, cellulose chains are further degraded by CBHs in a processive manner. EGs are mostly active on amorphous cellulose and less active on crystalline regions. CBHs due to their processive hydrolysis function have their active sites inside a cellulose-binding tunnel and attack cellulose preferentially from either the reducing (CBH1 type; e.g. GH7 CBHs) or non-reducing end (CBH2 type; e.g. GH6 CBHs), cleaving off a number of cellobiose units. Processive cellulases are found in GH family 6 and 7. CBHs are the most abundant enzymes expressed by fungi and are considered to be key enzymes in cellulose degradation. Gene knockout of CBHs in *T.reesei* have shown that the ability to degrade cellulose was significantly reduced and CBHs were proved to be essential in enzymatic cocktails (Suominen et al., 1993).

2.10 GH7 cellulases

2.10.1 Hydrolytic mechanism of GH7 enzymes

At the catalytic centre of GH7s, there is a catalytic triad with three carboxylic acid residues located on the same beta-strand inside the tunnel in the highly conserved motif EXDXXE. In the canonical *Tre*Cel7A, Glu212 and Glu217, were shown to act as nucleophile and acid/base, respectively (Figure 5). The hydrolytic activity was drastically reduced, by orders of magnitude, for the E212Q mutant and the E217Q mutant of TreCel7A.The D214N mutant also exhibited substantially reduced activity, showing that Asp214 is involved in catalysis, probably due to hydrogen bonding with Glu212 (Stahlberg et al., 1996).

GH7 enzymes have multiple carbohydrate binding sites in the catalytic domain and exhibit at least 9 subsites for cellulose binding. The subsites are numbered as *–n* for non reducing end and *+n* for reducing end (Figure 6). The glycosidic bond cleavage in GHs occurs between the −1 and +1 subsites, so for GH7s it takes place 2 glucose units from the reducing end of the chain bound in the tunnel (Payne et al., 2015).

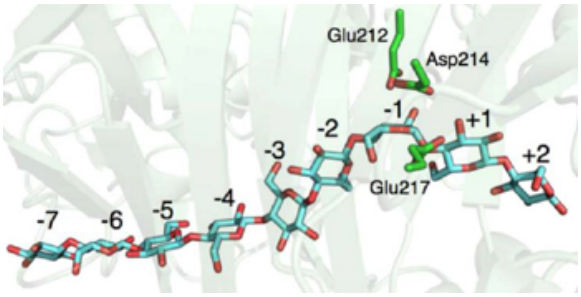


Figure 6. TreCel7A Michaelis complex (PDB code 4C4C) shows the standard numbering of the substrate binding sites (catalytic residues shown in green). A cellulose chain enters from the −7 site; 1/+2 sites are termed the "product sites". Figure was taken from Payne et al. (2015).

GH7 utilize a two-step retaining mechanism of hydrolysis, which may occasionally instead lead to transglycosylation. During the glycosylation step proton transfer occurs from the acid/base residue to the glycosidic oxygen. The nucleophile simultaneously makes nucleophilic attack at the anomeric carbon of the glycosyl unit in subsite -1. The glycosidic bond is broken and glycosyl-enzyme intermediate is formed, where the remaining cellulose chain is bound with a glycosyl-ester bond to nucleophile. During deglycosylation step, an activated water molecule makes a nucleophilic attack at the anomeric carbon,

which breaks the glycosyl-intermediate bond and simultaneously restores the acid-base and nucleophile residues.

In nature, GH7 act on insoluble substrates, but in vitro they easily hydrolyse glycosidic bond in short oligosaccharides and model substrates with chromophoric group (*p*-nitrophenyl β-lactoside, *p*Np-Lac; 2-chloro-4-nitrophenyl β-cellobioside, CNp-G2), which are commonly used for their initial characterisation.

2.10.2 Processive hydrolysis of cellulose

In general, processivity is defined as a multistep consecutive catalysis, where the enzyme stays attached to the substrate during several catalytic events before detachment from the substrate. For GH7 CBHs, processivity is crucial for the ability of the enzyme to act efficiently on a cellulose chain. There are several methods, direct and indirect, to measure processivity. P. Väljämäe and coworkers at Tartu University, Estonia, developed ^{14}C-labelling of cellulose (Kurasin and Valjamae, 2011) and the laboratory of P. Westh at Roskilde University, Denmark use amperometric cellobiose dehydrogenase (CDH) biosensor technology (Cruys-Bagger et al., 2012, Cruys-Bagger et al., 2013). These two methods are in good agreement and enable to obtain values for apparent processivity number n, enzyme-substrate association rate constant k_{on}, enzyme dissociation constant k_{off} and hydrolysis rate constant k_{cat}. Processivity parameter n represents the average number of sequential catalytic cycles. A three-step kinetic model of processive hydrolysis is represented by reaction scheme shown in Figure 7 (Praestgaard et al., 2011). The Väljamäe group has also been able to estimate the ratio between endo- and exo-initiation for GH7 CBHs, and found unexpectedly high endo-initiation probability of 40-50% for *Tre*Cel7A and 70-80% for *Pch*Cel7D on bacterial cellulose (Kurasin and Valjamae, 2011).

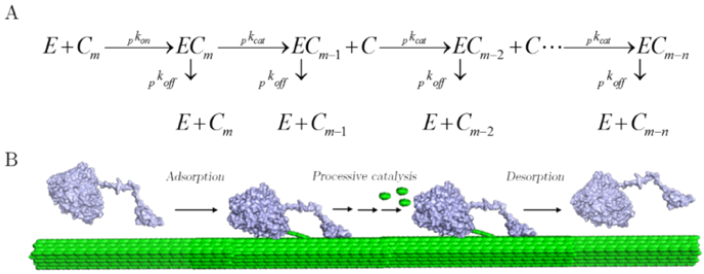


Figure 7. Simplified reaction scheme for a processive cellulase (A) and an illustration of the molecular steps involved in this scheme (B). Reaction scheme was taken from Praestgaard et al. (2011).

2.10.3 Structural features of GH7 enzymes

The first cellulases structures were published in the late 1980s and early 1990s, providing identification of substrate interactions and catalytic mechanisms, and since then number of structural studies has come up, followed by computational analysis and number of structure dynamics simulation. Nowadays there are 17 structures of GH7s available in the PDB database (Protein Data Bank, http://www.rcsb.org/pdb/), including mainly fungal cellulases from ascomycetes: *Trichoderma reesei* Cel7A (PBD code 1CEL) and Cel7B (PDB code 1EG1), *Trichoderma harzianum* Cel7A (PDB code 2YOK), *Fusarium oxysporum* Cel7B (PDB code 1OVW), *Aspergillus fumigatus* Cel7A (PDB code 4V1Z), *Humicola grisea* Cel7A (PDB code 4SCI), *Humicola insolens* Cel7B (PDB code 1A39), *Scytalidium sp* Cel7A (PDB code 4ZZT), *Melanocarpus albomyces* Cel7B (PDB code 2RFW), *Talaromyces funiculosus* Cel7A (PDB code 4XEB), *Rasamsonia emersonii* Cel7A (PDB code 1Q9H); basidiomycetes *Phanerochaete chrysosporium* Cel7D (PDB code 1GPI) and *Heterobasidion irregulare* Cel7A (PDB code 2XSP); and non-fungal from amoebozoa *Dictyostelium discoideum* Cel7A (PDB code 4ZZP) and *Dictyostelium purpureum* Cel7A (PDB code 4ZZP); crustacea *Limnoria quadripunctata* Cel7B (PDB code 4GWA) and water flea *Daphnia pulex* Cel7A (PDB code 4XMN). Despite the diversity of cellulase-expressing organisms the overall fold of GH7 enzymes is rather similar.

In general GH7 proteins share a β-jelly roll fold with two largely antiparallel β-sheets packing face-to-face into a curved β-sandwich. Loop regions *extend* the edges of the β-sandwich to form a 45 Å long groove along the entire catalytic domain. CBHs within GH7 are readily distinguished because several loops are further elongated, which effectively encloses the active site in a tunnel. This enables the CBHs to act processively along a cellulose chain and cleave off numerous cellobiose units before detachment from the substrate, which is believed to be key to their efficiency on highly crystalline cellulose (Payne et al., 2015). Deconstruction of cellulose by GH7 CBHs is a multi-step process that includes substrate binding, formation of the catalytically active complex, hydrolysis, product release and processive translation along the substrate chain.

GH7 CBHs work predominantly from the reducing towards the non-reducing ends of cellulose chains, while GH6 CBHs preferentially act in the opposite direction. However, CBHs are not true exo-enzymes in the sense that they do not seem to be exclusively restricted to chain initiation by threading of a chain end through the tunnel. Experiments with *Trichoderma reesei* Cel7A (*Tre*Cel7A) and *Phanerochaete chrysosporium* Cel7D (*Pch*Cel7D) point towards substantial ratios of endo-initiation (40-80 %; (Kurasin and Valjamae,

2011)). *Tre*Cel7A exhibits the most enclosed tunnel amongst known GH7 CBH structures, while *Pch*Cel7D displays the most open active site due to several loop deletions and residue size reductions on the tips of tunnel enclosing loops (von Ossowski et al., 2003).

Probability of loop opening and closing seems to be responsible for endo/exo initiation, especially loop contacts seem to effect processivity of the enzyme. In contrast to GH7 CBHs, EGs exhibit a more open binding cleft, due to absence and shortening of essential loops (Figure 8). This significant structural feature results consequently in higher propensity for internal bond cleavage. In *Trichoderma reesei* the GH7 enzymes CBH and EG (Cel7A and Cel7B, respectively) share 55% sequence identity and are quite different in performance of cellulose degradation. About one third of the known GH7 members are bimodular in nature, having a family 1 carbohydrate-binding module (CBM1) connected to the catalytic domain (CD) by a glycosylated, flexible linker comprised of about 30 amino acids (Stals et al., 2004, Beckham et al., 2010, Sammond et al., 2012).

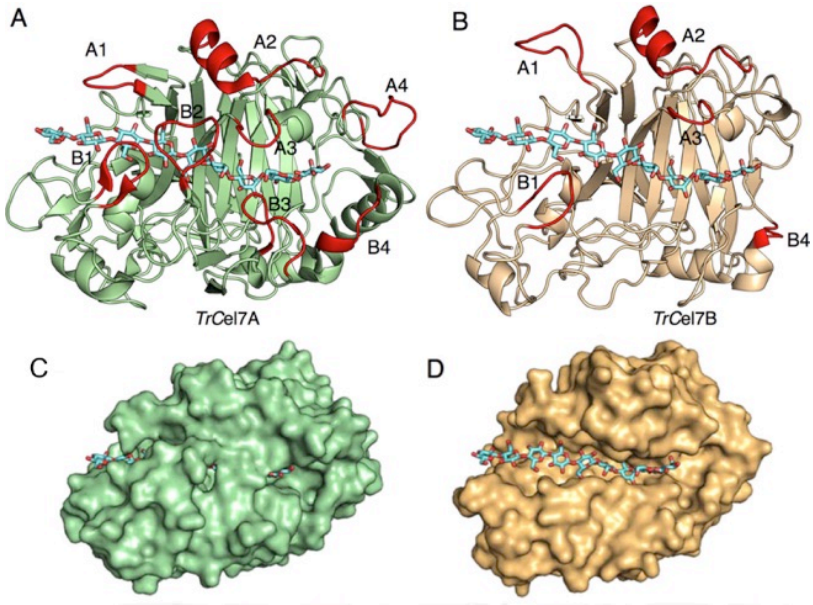


Figure 8. Structural features of GH7 cellulases from *Trichoderma reesei*: panel A – CBHI, *Tre*Cel7A ligand complex structure with cellonanose (PDB code 4C4C); panel B – EGI, *Tre*Cel7B (PDB code 1EG1); panel C - TreCel7A view from bottom showing the more closed substrate binding tunnel; panel D - TreCel7B view from the bottom showing the more open binding groove. Tunnel-enclosing loops are marked in red, according to nomenclature (Momeni et al., 2013b). The ligand from 4C4C is shown in all panels. Differences in loop regions: there is 13-15 residue insertion in B2 loop in CBHs and B3 loop is missing in EGs. Figure was taken from Payne et al. (2015).

Compared to other GH families, Differences are primarily found in loops and surface regions distant from the active site. However, there are also small variations in the length and sequence of loop regions along the cellulose-binding path that will affect the dynamics of loop regions and the accessibility of the active site (Borisova et al., 2015, Momeni et al., 2013, Textor et al., 2013). These variations may in turn influence key enzymatic properties such as processivity, product inhibition, endo-initiation and rate of substrate dissociation (Kurasin and Valjamae, 2011, Sorensen et al., 2017).

2.11 Phylogenetic analysis of fungal GH7 cellulases

GH7 cellulases are specifically designed by nature for cellulose degradation and processing cellulose chain. Simply one can propose that cellulase evolution always been following evolution of cellulose and plant cell walls. In addition to fungi, GH7 encoding genes are found in very distant branches of the eukaryotic tree of life such as e.g. Amoebozoa, Oomycetes, Dinoflagellates and Crustaceans, but not in any procaryote so far (King et al., 2010).

The degree of conservation of GH7 CBHs through evolution is remarkably high, even between organisms that diverged more than 1 billion years ago. This suggests that GH7 CBHs cannot accommodate a broad sequence space for primary function (Paper I). At the same time GH7 EGs appear to be represented by higher sequence diversity and therefore presumably larger structural variation (Sukharnikov et al., 2011). This indicates that EGs are either more evolutionary diverse or many novel CBHs are yet to be found (Gilbert, 2010). Phylogenetic analysis of protein sequences of fungal GH7 cellulases reveal that there are significant differences in evolutional relations within Ascomycetes and Basidiomycota. In Ascomycetes there are two distinct clades, one is formed by CBH sequences and another is formed by EG sequences. This classification is supported by biochemical and structural data available on Cel7s from Ascomycota (Figure 9A).

In contrast to Ascomycetes, GH7 sequences in Basidiomycota are very closely related and form a cluster (Figure 9B). No obvious EGs have been identified in Basidiomycetes so far. Nevertheless, phylogenetic analysis suggests some sequences as putative EGs, for example Cel7s from *Melampsora larici-populina.* This allows us to hypothesize that genomes of early diverging Basidiomycetes inherited a GH7 CBH gene from Ascomycetes and then the two divisions of fungi developed their cellulose-degrading complexes independently. In any case, a GH7 CBH gene seems to be the ancestor gene for GH7 enzymes, also for the EGs of Ascomycetes, which may have developed by promotion of endo-action.

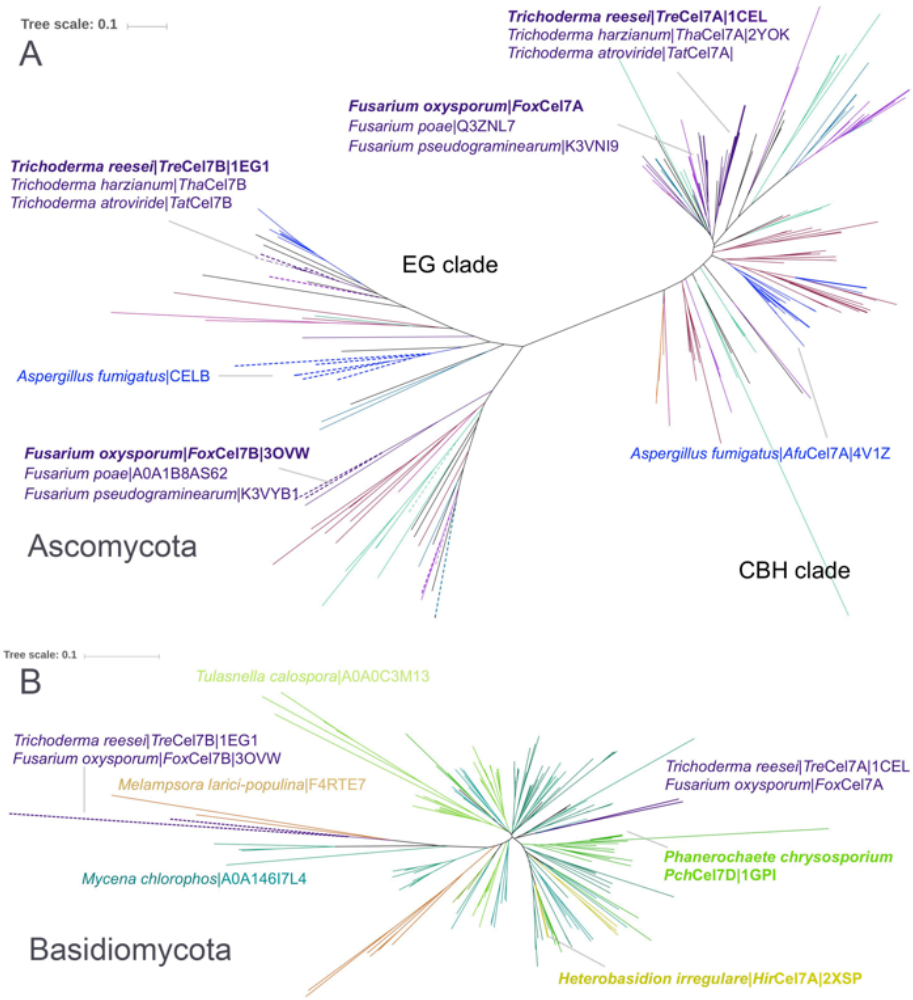


Figure 9. Phylogenetic analysis on fungal GH7. Panel A- Ascomycetes. The analysis was done on a set of 227 amino acid sequences, annotated as ascomycetes members of GH7 family in UniProt database (http://www.uniprot.org). For two-domain Cel7s the linker region and CBM were removed, and only catalytic domains without signal peptide were used in the alignment. The evolutionary history was inferred using the Minimum Evolution method (Rzhetsky and Nei, 1992) with 700 bootstrap replicates in MEGA7 (Kumar et al., 2016), from multiple alignment by ClustalW. Branch lengths are drawn to scale with evolutionary distances in units of number of amino acid substitutions per site, computed using the Dayhoff matrix based method. Two clades supported by biochemical and structural data: CBH and EG clade. Panel B - Basidiomycetes. The analysis was done on a set of 202 amino acid sequences, annotated as basidiomycetes members of GH7 family in UniProt database, and 4 reference sequences from Ascomycets (*Trichoderma reesei* Cel7A and Cel7B, and *Fusarium oxospirum* Cel7A and Cel7B), as described above. The tree represents a cluster. Both trees were visualised using online server iTOL (Interactive Tree of Life, http://itol.embl.de/).

3 Current investigation

3.1 Biochemical and structural characterization of two *Dictyostelium* cellobiohydrolases from the *Amoebozoa* kingdom reveal a high conservation between distant phylogenetic trees of life (Paper I).

Almost all examined GH7 CBHs are from fungi and most of them from ascomycetes due to their key role in cellulose degradation and extensive production by native hosts. Actually, it was long believed that GH7 CBHs only occur in fungi, until quite recently when members from non-fungal organisms started to emerge. The quest to develop detailed insights into structure-function relationships of GH7 enzymes, important in nature and for the growing biofuels industry, prompts for research on novel target CBHs from non-fungal organisms. The GH7 CBHs from the social amoeba *D. discoideum* and *D. purpureum (Ddi*Cel7A and *Dpu*Cel7A, respectively) were chosen for structural and biochemical characterization because Dictysotelia are phylogenetically distant from fungi. Furthermore, the fact that the enzymes may have a different functional role, in endogenous cellulose metabolism rather than for food acquisition from plant material, makes them even more interesting for comparison.

3.1.1 Enzyme expression and Biochemical characterization

The *Dictyostelium* enzymes were recombinantly expressed in *T. reesei* using the newly developed pTrEno expression system. A fungal enolase promoter allows consecutive expression with glucose as carbon source, without background of endogenous cellulases (Linger et al., 2015). The native sequences of *Ddi*Cel7A and *Dpu*Cel7A consist of a sole GH7 CD, and do not exhibit a CBM-linker domain, maybe for functional reasons such as environments of high cellulose density, where increasing the binding of the

enzyme to its crystalline cellulose substrate is not beneficial. However, in order to compare GH7 CBH conversion at low solids loadings, we have added the Family 1 CBM and linker from *Tre*Cel7A to the CD of native *Ddi*Cel7A and *Dpu*Cel7A, producing chimeric *Ddi*Cel7$_{CBM}$ and *Dpu*Cel7A$_{CBM.}$ To verify the functional significance of the unique product site motif, the biochemical characteristics and activities of *Ddi*Cel7A$_{CBM}$ and *Dpu*Cel7A$_{CBM}$ were investigated on soluble substrate *p*NP-Lac (Table 1).

Table 1. *Biochemical and kinetic characterization on pNP-Lac*

Enzyme	T_{opt} (°C)	pH opt.	T_m (°C)	k_{cat} (min^{-1})	K_m (mM)	K_i (μM)
*Ddi*Cel7A$_{CBM}$	45	5	53	16.1 ± 2.8	3.4 ± 0.1	205
*Dpu*Cel7A$_{CBM}$	55	5	63	36.0 ± 6.7	3.4 ± 0.3	130
*Tre*Cel7A	55	4	64	16.1 ± 3.0	1.2 ± 0.1	29

Both *Dictyostelium* Cel7s exhibit reduced cellobiose inhibition in comparison to *Tre*Cel7A; and *Dpu*Cel7A$_{CBM}$ shows lower catalytic activity on *p*NP-Lac, despite high sequence similarity with the homolog *Ddi*Cel7A$_{CBM}$. Enzymatic activity comparison on insoluble cellulosic substrates showed that *Ddi*Cel7$_{CBM}$ and *Dpu*Cel7A$_{CBM}$ hydrolysis parameters are strikingly similar to those of *Tre*Cel7A (Paper I).

3.1.2 Crystal structures of DdiCel7A and DpuCel7A

The *Ddi*Cel7A structure was solved in space group $P2_12_12_1$ with a single protein chain in the asymmetric unit; and the *Dpu*Cel7A structure was solved in $P2_12_12_1$ with two chains, A and B, in the asymmetric unit. *Ddi*Cel7A and *Dpu*Cel7A structures were refined at 2.1 and 2.7 Å resolution, with R/R_{free} factors of 0.25/0.35 and 0.23/0.31, respectively.

As expected from the high amino acid sequence identity, the folds of *Ddi*Cel7A and *Dpu*Cel7A are very similar to one another (80% identity; rmsd 0.45 Å) as well as to the catalytic module of *Tre*Cel7A (59% identity; rmsd 0.66 and 0.62 Å) (Figure 10).

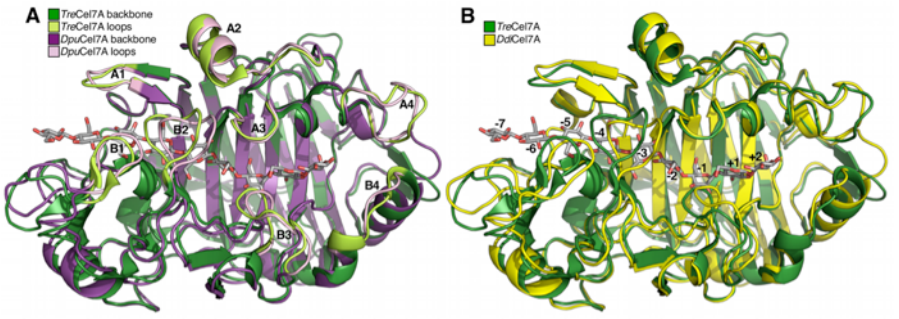


Figure 10. Superposition of *Dpu*Cel7A (A) and *Ddi*Cel7A (B) with the cellononaose Michaelis complex of *Tre*Cel7A (PDB code 4C4C). The key substrate-binding loops are labeled in (A) and the substrate binding sites are labeled in (B).

Structural alignments of *Ddi*Cel7A and *Dpu*Cel7A to other known GH7 CBH structures show that the conformation of tunnel-enclosing loops in *Ddi*Cel7A and *Dp*Cel7A are exceptionally similar to each other and to that of *Tre*Cel7A, revealing that the shape and degree of closure of the binding tunnel is most similar to *Tre*Cel7A among the GH7 homologs (Figure 11). Functionally, this indicates high processivity slow dissociation from substrate.

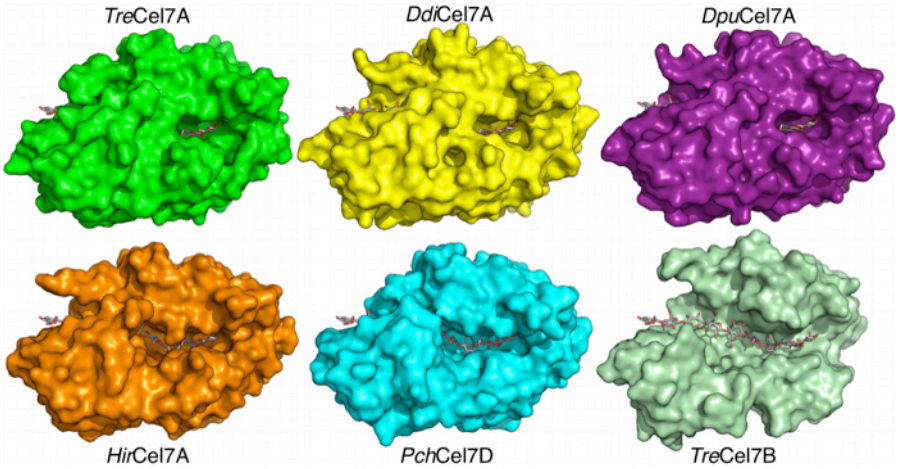


Figure 11. Space-fill GH7 structures comparing substrate tunnel enclosure of CBHs from T. reesei (PDB code 4C4C), D. discoideum, D. purpureum, H. irregulare (PDB code 2YG1), P. chrysosporium (PDB code 1GPI), and endoglucanase (EG) Cel7B from T. reesei (PDB code 1EG1). In all frames, the cellononaose ligand from the TreCel7A Michaelis complex is shown in gray 'sticks'.

However the loop contacts B2/A3/B3, essential for tunnel closure exhibit significant differences in comparison to *Tre*Cel7A. First one is the lack of

tyrosine in both *Ddi*Cel7A and *Dpu*Cel7A B3 and A3 loops whose side chains maintain some degree of flexibility (exemplified by their different positions in *Tre*Cel7A structures 1CEL and 4C4C) and partially enclose the binding tunnel at the catalytic center (Payne et al., 2015). To compensate, the *Dictyostelium* Cel7A enzymes have a unique tyrosine (Tyr202) that protrudes from the base of the B2 loop whose side chain resides in the same general space as that of the A3 and B3 tyrosines in the other GH7 structures (Tyr247 in *Tre*Cel7A) (Figure 12). This governs a relatively tight binding tunnel enclosure in *Dpu*Cel7A and *Ddi*Cel7A in this region and may suggest that the *Dictyostelium* Cel7A enzymes may have elevated probabilities of 'endo-initiation' as compared with *Tre*Cel7A (Kurasin and Valjamae, 2011). Another remarkable feature in loop contacts is Gln377 in *Ddi*Cel7A and Leu377 in *Dpu*Cel7A opposing residue on loop A3 (Tyr371 in *Tre*Cel7A). Glutamine at this position is a rare motif among GH7 CBHs and is found only in another protist, *Pseudotrichonympha grassii*, a parabasilian symbiont in the gut of termites. A few distantly related basidiomycetes have glutamate at this position, e.g., in *Hir*Cel7A, where MD simulations showed that the glutamate sidechain can interact with a bound cellulose chain (Momeni et al., 2013). Thr246 at the tip of loop B3 in *Tre*Cel7A is replaced by Ala245 in *Ddi*Cel7A and *Dpu*Cel7A, resulting in the loss of a hydrogen bond to the substrate at subsite +1 (Figure 12).

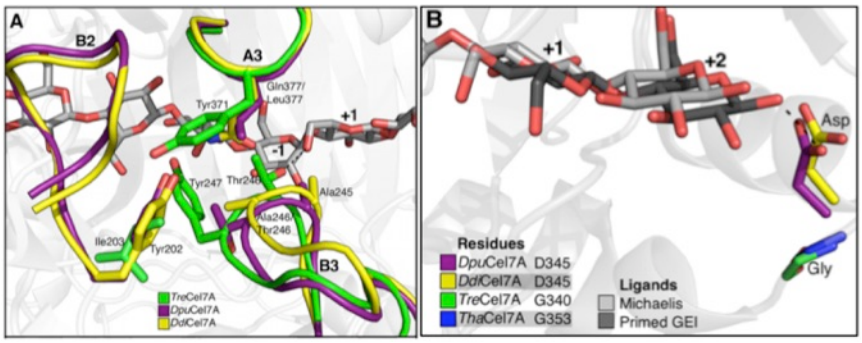


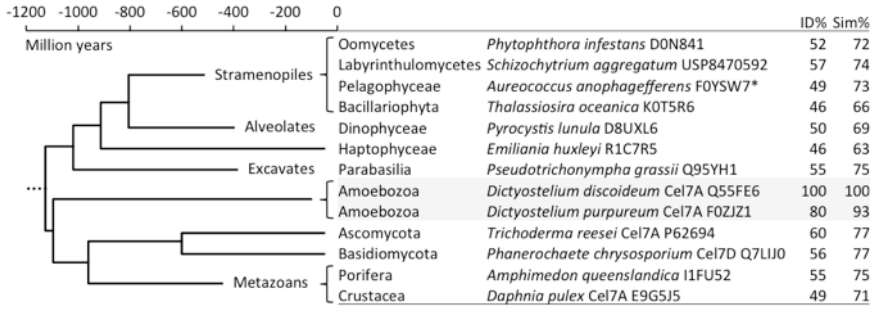
Figure 12. Structural alignments of *Ddi*Cel7A and *Dpu*Cel7A to other known GH7 CBH structures. A) Alignment around the active site illustrating interactions of the B2, B3, and A3 loops discussed in the main text. B) Product binding sites illustrating the relative positioning of the commonly found aspartate residue that interacts with the +2 glucosyl unit, but which is replaced by glycine in *Tre*Cel7A and *Tha*Cel7A. The ligands are both from *Tre*Cel7A structures: Michaelis (PDB code 4C4C) and Primed GEI (PDB code 3CEL). The aspartic acid side chains of *Pch*Cel7D, *Hir*Cel7A, *Tem*Cel7A, *Afu*Cel7A, *Lqu*Cel7B essentially overlay those of *Dpu*Cel7A and *Ddi*Cel7A and are not shown for clarity.

Reduced product inhibition by cellobiose in *Dictyostelium* enzymes might be due to that product binding sites of *Ddi*Cel7A and *Dpu*Cel7A are slightly

different from *Tre*Cel7A. One important difference is the single mutation on the tip of B3 loop, mentioned above, another is the insertion of an aspartate residue in loop B4 relative to *Tre*Cel7A (D345 in *Ddi* and *Dpu*Cel7A), making a hydrogen bond to the sugar in subsite +2 (Figure 12B).

3.1.3 Phylogenetic analysis of GH7 CBHs

Following rapid increase of metagenomics data, the number of annotated GH7 genes has dramatically expanded. These genes have been well characterized and identified in many cellulolytic fungi and more recently in animals such as marine wood borers (Crustacea, e.g., the gribble or *Limnoria quadripunctata*) (King et al., 2010). Moreover, they have been found in additional eukaryote branches such as amoeba (Eichinger et al., 2005), oomycetes (Stramenopiles, e.g., the potato-blight pathogen *Phytophthora infestans*), haptophytes (e.g., the phytoplankton *Emiliania huxleyi*), and parabasilids (Excavata, e.g., the termite hindgut symbiont *Pseudotrichonympha grassii*) (Parfrey et al., 2010, King et al., 2010). To visualise the evolutionary diversity, the evolution timeline of divergence of branches where GH7 CBHs have been found, was compared with sequence identities and similarites within the GH7 domain relative to *Ddi*Cel7A (Figure 13).



			ID%	Sim%
Stramenopiles	Oomycetes	*Phytophthora infestans* D0N841	52	72
	Labyrinthulomycetes	*Schizochytrium aggregatum* USP8470592	57	74
	Pelagophyceae	*Aureococcus anophagefferens* F0YSW7*	49	73
	Bacillariophyta	*Thalassiosira oceanica* K0T5R6	46	66
Alveolates	Dinophyceae	*Pyrocystis lunula* D8UXL6	50	69
	Haptophyceae	*Emiliania huxleyi* R1C7R5	46	63
Excavates	Parabasilia	*Pseudotrichonympha grassii* Q95YH1	55	75
...	Amoebozoa	*Dictyostelium discoideum* Cel7A Q55FE6	100	100
	Amoebozoa	*Dictyostelium purpureum* Cel7A F0ZJZ1	80	93
	Ascomycota	*Trichoderma reesei* Cel7A P62694	60	77
	Basidiomycota	*Phanerochaete chrysosporium* Cel7D Q7LIJ0	56	77
Metazoans	Porifera	*Amphimedon queenslandica* I1FU52	55	75
	Crustacea	*Daphnia pulex* Cel7A E9G5J5	49	71

Figure 13. Eukaryote evolution timeline of branches where GH7 genes have been found, and protein sequence identities (ID%) and similarities (Sim%) within the GH7 domain to *D. discoideum* Cel7A. Branch names and timepoints of divergence are from (Berney and Pawlowski, 2006). Only points of early divergence are included (<-600 Myr); later divergence among Stramenopiles, Amoebozoa and Metazoans is not resolved here. The multiple sequence alignment of selected GH7 sequences was done with MUSCLE, and flanking regions were trimmed off (e.g. signal peptide, CBM) before calculation of pairwise sequence identities and similarities using the Gonnet substitution matrix. Accession codes are from UniProt, except *S. aggregatum* from (Brevnova et al., 2013). *The *A. anophagefferens* F0YSW7 sequence appears to be a fragment containing only 135 residues of the C-terminal part of the GH7 domain in the alignment.

Noticeably GH7 genes occur in branches of the eukaryotic tree of life that diverged over 1 billion years ago (Berney and Pawlowski, 2006), suggesting either

an ancient ancestral GH7 gene or gene uptake via horizontal gene transfer (HGT). Compared to other GH families, GH7 CBHs represent remarkably high sequence identity (>40%; Figure 13), and given the early event of divergence, this suggests that GH7 CBHs cannot accommodate a broad sequence space for primary function. However, we cannot rule out that HGT has occurred on a more recent timescale, thus limiting the extent of sequence divergence.

Phylogenetic analyses of taxonomically diverse GH7 protein sequences of an extended set of 113 on the GH7 domain alone was conducted with the program MEGA7 (Kumar et al., 2016). Sequences were collected by pBLAST search of each non-fungal GH7 against NCBI and UniProt databases, and selection of taxonomically diverse hits among the many fungal entries retrieved. A 90% identity threshold was applied to remove redundancy, and EGs sequences, defined as lacking the B4 loop (marked in Figure 10A) were excluded. The evolutionary history was inferred using the Minimum Evolution method (Rzhetsky and Nei, 1992) and bootstrap phylogeny testing with 700 replicates (Felsenstein, 1985); and the optimal tree with the sum of branch length of 19.8 is shown on Figure 14.

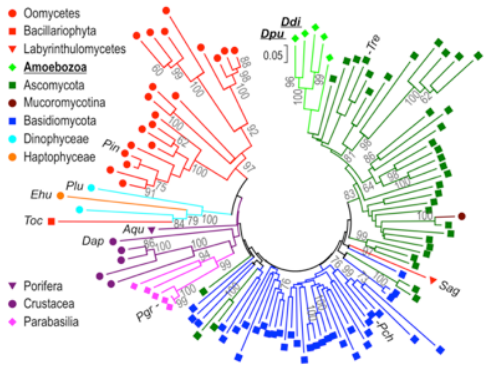


Figure 14. Phylogenetic tree of 113 GH7 CBH protein sequences. The evolutionary history was inferred using the Minimum Evolution method (Rzhetsky and Nei, 1992) with 700 bootstrap replicates in MEGA7 (Kumar et al., 2016), from multiple alignment by MUSCLE of the GH7 domain. Branch lengths are drawn to scale with evolutionary distances in units of number of amino acid substitutions per site, computed using the Dayhoff matrix based method. Grey numbers are bootstrap values in percentage. A key to colors and symbols is provided in the figure. Sequences listed in Figure 7 and Table 4 are marked with abbreviations: *Ddi, Dictyostelium discoideum* Cel7A Q55FE6; *Dpu, Dictyostelium purpureum* Cel7A F0ZJZ1; *Tre, Trichoderma reesei* Cel7A P62694; *Sag, Schizochytrium aggregatum* USP8470592; *Pch, Phanerochaete chrysosporium* Cel7D Q7LIJ0; *Pgr, Pseudotrichonympha grassii* Q95YH1; *Dap, Daphnia pulex* Cel7A E9G5J5; *Aqu, Amphimedon queenslandica* I1FU52; *Toc, Thalassiosira oceanica* K0T5R6; *Ehu, Emiliania huxleyi* R1C7R5; *Plu, Pyrocystis lunula* D8UXL6; *Pin, Phytophthora infestans* D0N841.

The amoeba are placed among ascomycete sequences, indeed suggesting HGT from an early ascomycete to a Dictyostelial ancestor (or vice versa). The event should then have occurred after the separation of Ascomycota and Basidiomycota (~600 Myr ago), but before the divergence of *D. discoideum* and *D. purpureum* estimated at around 400 million years ago (Sucgang et al., 2011). HGT has been hypothesized as a primary means of genome evolution in *D. discoideum* (Eichinger et al., 2005). Dictyostelia inhabits the forest soil and thus it seems plausible that HGT of GH7 CBHs from fungi may have been the primary mechanism for these enzymes being present in slime mold genomes. Similar arguments have been made for HGT into wood boring crustaceans, such as *L. quadripunctata* (King et al., 2010). Notably the genes in Dictyostelia acquired by HGT are all of bacterial origin (Sucgang et al., 2011), and prokaryote-eukaryote HGT is generally easier to detect than HGT of nuclear genes between eukaryotes (Keeling and Palmer, 2008).

3.1.4 Conclusions

This study reveals a remarkable conservation in GH7 enzymes across distant phylogenetic branches of the eukaryotic tree of life, suggesting that GH7 CBHs cannot accommodate a broad sequence space for primary function. Two crystal structures of social amoeba GH7 CBHs are presented, and it is shown that these enzymes are structurally and functionally very similar to the well-studied *Tre*Cel7A. GH7 CBHs function in Dictyostelia, being related to the entire cellulose metabolism during the slime mold life cycle (Kunii et al., 2013) or to their ability to digest cellulose-containing organisms in the forest soil (Eichinger et al., 2005), remains a question of particular interest to understand the evolutionary pressures on the activities of slime mold GH7 CBHs.

3.2 Correlation of structure, function and protein dynamics in GH7 cellobiohydrolases from *Trichoderma atroviride, T. reesei and T. harzianum* (Paper II).

The ascomycete fungus *T. reesei* is the predominant source of enzymes for industrial lignocellulosic ethanol production, while other *Trichoderma* species have attracted attention as alternative enzyme sources for different reasons (Jiang et al., 2011, van Wyk and Mohulatsi, 2003, Kovacs et al., 2009, Grigorevski-Lima et al., 2013). Most *Trichoderma* spp. are described as mycoparasitic fungi, and several have rendered interest as powerful biocontrol agents (BCA) against pathogenic fungi (Schmoll et al., 2016). Such BCA fungi include *T. harzianum* and *T. atroviride,* while *T. reesei* is a weak mycoparasite and is adapted to a saprotrophic lifestyle as a wood-degrader (Karlsson et al., 2017). Comparative studies of closely related GH7 CBHs from *T. atroviride*, *Tat*Cel7A, *T. harzianum, Tha*Cel7A**,** and from *T. reesei*, *Tre*Cel7A are extremely relevant for sequence-performance relation insight due to high similarities of these enzymes (80% sequence identity) and their functional properties. Together they represent three naturally occuring variants in terms of combination of tunnel-enclosing loop motifs: A1 loop and A3-B3 interations.

3.2.1 Preparation and biochemical characterization of Cel7A enzymes

The Cel7A enzymes were purified from culture filtrates of *T. atroviride* IOC 4503, *T. harzianum* IOC 3844 and *T. reesei* QM9414, grown in submerged culture under cellulase inducing conditions. In all cases, Cel7A was the major protein and the yield of purified enzyme per liter of culture was 70 mg for *Tat*Cel7A and 85 mg for *Tha*Cel7A, which is significantly lower than typical yields of 200-700 mg/L *Tre*Cel7A from *T. reesei* QM9414 (Stahlberg et al., 1996). Partial proteolysis with papain was used to remove the CBM-linker portion from the full length enzyme and prepare isolated catalytic domains, *Tat*Cel7A_CD, *Tha*Cel7A_CD and *Tre*Cel7A_CD.

Initial biochemical characterisation was done with Cel7A catalytic domains, following their activity on *p*NP-Lac. pH stability and pH optimum profiles for all three enzymes were rather similar, although *Tat*Cel7A_CD exhibited singificantly lower specific activity on pNP-Lac. A similar discrepancy between close homologs was observed for the *Dictyostelium* Cel7s in Paper I. For *Tat*Cel7A this can be rationalized by non-productive binding of *p*NP-Lac,

supported by structural features and high flexibility of A4 loop, predicted in MD simulations (Paper II). In pH stability experiments, all three enzymes showed loss of activity after incubation at pH 6 and higher, most pronounced with *Tha*Cel7A_CD. Temperature optimum and stability experiments were carried out at pH 4.5, defined as optimal pH for all three enzymes. *Tat*Cel7A_CD was found to be the most temperature sensitive, while *Tre*Cel7A_CD was the most thermostable.

The activity of the Cel7s on cellulosic substrates was studied on bacterial microcrystalline cellulose (BMCC) and on dilute acid-pretreated corn stover (PCS). The initial production of cellobiose from BMCC was monitored in real-time using an amperometric cellobiose dehydrogenase (CDH) enzyme biosensor (Cruys-Bagger et al., 2012, Cruys-Bagger et al., 2013). The kinetic parameters were derived from non-linear regression fit to progress curves of initial hydrolysis. *Tat*Cel7A was faster, than either *Tha*Cel7A or *Tre*Cel7A in initial hydrolysis of BMCC, while the processivity number was rather similar for all Cel7s, both full-length and catalytic domains Table 3.

Table 3. *Kinetic parameters derived from progress curves of initial BMCC hydrolysis: association (k_{on}), catalytic (k_{cat}) and dissociation (k_{off}) rate constants, and apparent processivity number (**n**). These parameters were derived for the 0-200s pre-steady-state time interval. The substrate load was 3.3 g/L and the enzyme concentration was 50 nM.*

Enzyme	k_{on} ($g^{-1}Ls^{-1}$)	k_{cat} (s^{-1})	k_{off} (s^{-1})	*n*
*Tre*Cel7A	0.0055 ±0.0002	4 ±0.2	0.0066 ±0.0005	89±5
*Tre*Cel7A$_{CD}$	0.0034 ±0.0005	4.9 ±0.7	0.0049 ±0.0012	88±5
*Tha*Cel7A	0.0056 ±0.0002	4.8 ±0.4	0.0061 ±0.0002	74±1
*Tha*Cel7A$_{CD}$	n/d	n/d	n/d	n/d
*Tat*Cel7A	0.0071 ±0.0001	8.3 ±0.3	0.0071 ±0.0001	97±1
*Tat*Cel7A$_{CD}$	0.0044 ±0	6.8 ±0.3	0.0066 ±0.0002	87±3

To estimate the efficiency of Cel7A enzymes in synergistic lignocellulose saccharification, performance assays on PCS were carried out for full-length GH7 CBHs together with a GH7 endoglucanase (*T. longibrachiatum* Cel7B/EG I) and a β-glucosidase. Along with the *Trichoderma* Cel7 enzymes under investigation, *Scy*Cel7A and *Pch*Cel7A were included in these experiments (Figure 15). Both *Tat*Cel7A and *Tha*Cel7A were more efficient than the canonical *Tre*Cel7A in synergistic saccharification of pretreated corn stover.

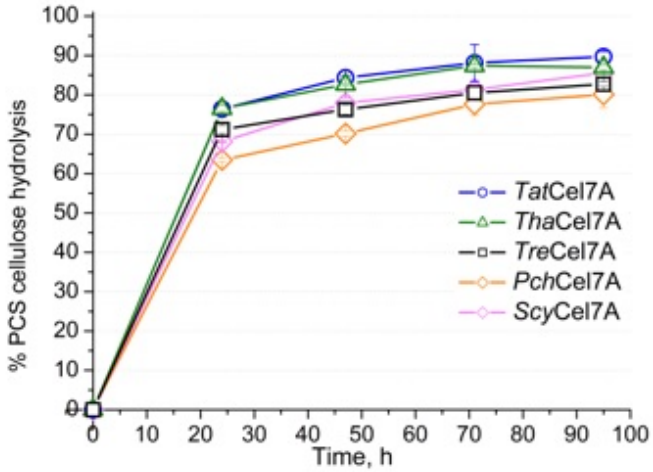


Figure 15. Conversion of pretreated corn stover (PCS) to soluble sugars at 40°C and pH 5.0 by full-length GH7 CBH enzymes together with a GH7 endoglucanase and a β-glucosidase (28, 1.9 and 0.5 mg enzyme per gram glucan, respectively). In addition to the three Trichoderma CBHs, Cel7A from Scytalidium sp (ScyCel7A, Paper III) and Cel7D from Phanerochaete chrysosporium (PchCel7D) were also analyzed at the same time; the results are shown for comparison. Experiments were performed in duplicate.

3.2.2 Crystal structures of TatCel7A and MD simulations

The *Tat*Cel7A_CD protein was successfully crystallized and two structures, apo structure (APO) and thio-cellotrioside complex (SG3), were solved in space group *P*21 with two protein chains, A and B, in the asymmetric unit. APO and SG3 were refined at 1.7 Å resolution, with R/R_{free} factors of 16.8/19.0 and 17.2/20.6, respectively. In the SG3 structure, from co-crystallization with thio-linked cellotriose, there are two cellotrioside molecules bound in each protein chain, in subsites -6/-5/-4 and +1/+2/+3, respectively (Figure 16). In the -6/-5/-4 position, the sugar ring at each site is flipped up-side-down compared to the orientation in the Michaelis complex of *Tre*Cel7A, thus representing a processive sliding intermediate.

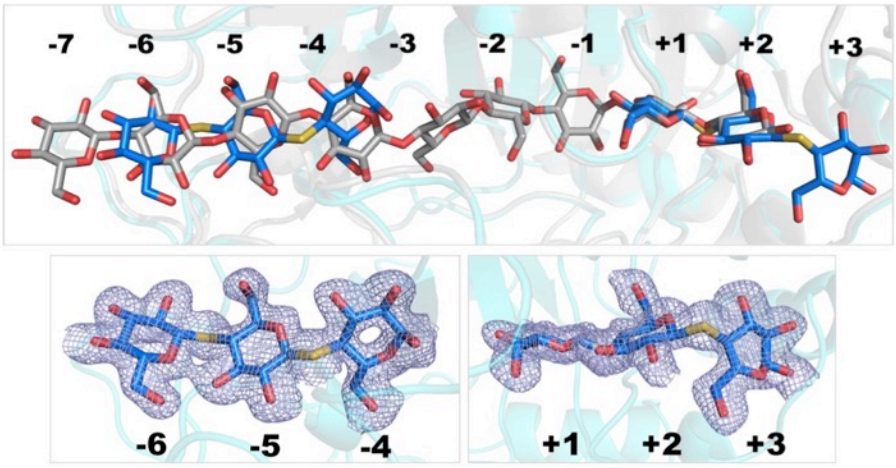


Figure 16. Sugar binding in the *Tat*Cel7A_CD thio-cellotrioside complex structure. (A) Superposition of ligand binding in the SG3 structure (blue) with cellononaose in the *Tre*Cel7A Michaelis complex (4C4C; grey). (B) Electron density for the ligand at subsites -6/-5/-4. (C) electron density for the ligand at the +1/+2/+3 position. The 2Fo-Fc electron density maps are contoured at 0.26 e/A^3.

Overall, the *Tat*Cel7A_CD structures are almost identical (0.18 Å root mean square deviation, RMSD) and the fold is very similar to that of *Tre*Cel7A_CD and *Tha*Cel7A_CD (RMSD 0.54 Å and 0.44 Å respectively), as expected from high sequence identity (80% and 82%, respectively) (Figure 18). *Tat*Cel7A appears to have fewer secondary structure interactions compared to *Tre*Cel7A and *Tha*Cel7A, with shorter β-strands and α-helixes at several locations. This is corroborated by a lower number of total native contacts found from the MD simulations (Paper II). Loop A4 carries an N-glycosylation site near the product sites that is conserved in all three enzymes. Glycosylation at this site has been observed in structures of *Tre*Cel7A (Asn384) and *Tha*Cel7A (Asn380), but was not visible in the *Tat*Cel7A structures (Asn384).

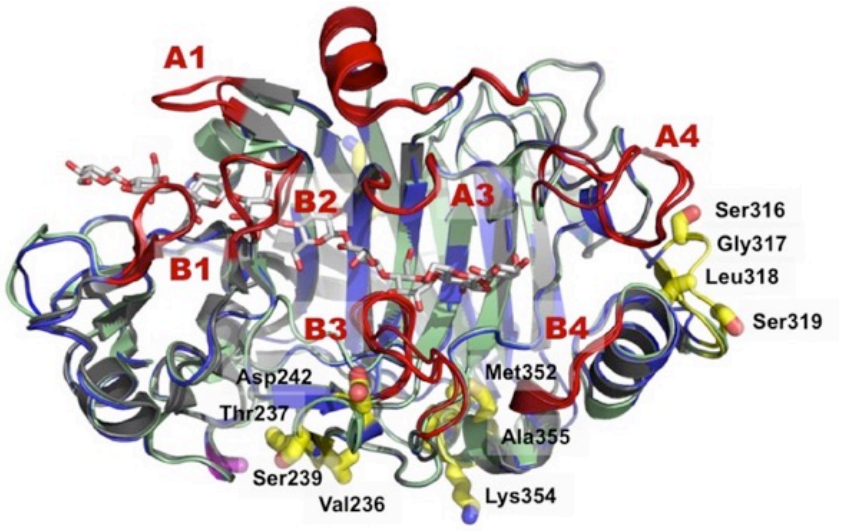


Figure 17. Overall structural alignment of apo *Tat*Cel7A_CD structures (APO), marked in blue, with *Tre*Cel7A_CD cellonanose complex (PDB code 4C4C), marked in grey, and *Tha*Cel7A_CD (PDB code 2YOK), marked in green. Loop regions are highlighted in red, cellonanose is colored in grey. RCA-defined sections I-IV are marked in yellow and amino acid residues in *Tat*Cel7A with high S-scores are shown as sticks.

The lining of the cellulose-binding path is identical in the three enzymes, except at two locations, loop A1 at the entrance to the tunnel, and loop A3 near the catalytic center. In loop A1, Glu101 in *Tre*Cel7A binds the 6-hydroxyl of the glucose unit in subsite -6. In *Tha*Cel7A, a corresponding interaction is completely absent due to shortening of the A1 loop (Figure 17). In *Tat*Cel7A there is Asn101 with a shorter sidechain instead of Glu101. In loop A3, Tyr371 in *Tre*Cel7A interacts with Tyr247 at the tip of the opposing B3 loop. Tyr371 is replaced by an alanine in both *Tat*Cel7A and *Tha*Cel7A, and there are no direct interactions between loops A3 and B3 across the tunnel. Based on the MD simulations, histograms of the distance probability between A2, B2 and B3 loops corroborate that B3 loop is the most flexible in *Tat*Cel7A among three TrCel7s (Paper II). Higher flexibility of the A4 loop may be due to the insertion of an extra residue nearby in *Tat*Cel7A (Gly317), which disrupts a β-strand-turn-strand motif that supports loop A4 in the other two enzymes. Protein dynamics and loop flexibility were calculated from MD simulations in terms of root mean square fluctuation (RMSF) for all three enzymes and the results are visualised in Figure 18. It shows that A4 loop is more flexible in *Tat*Cel7A, probably being responsible for its decreased thermostability (Paper II).

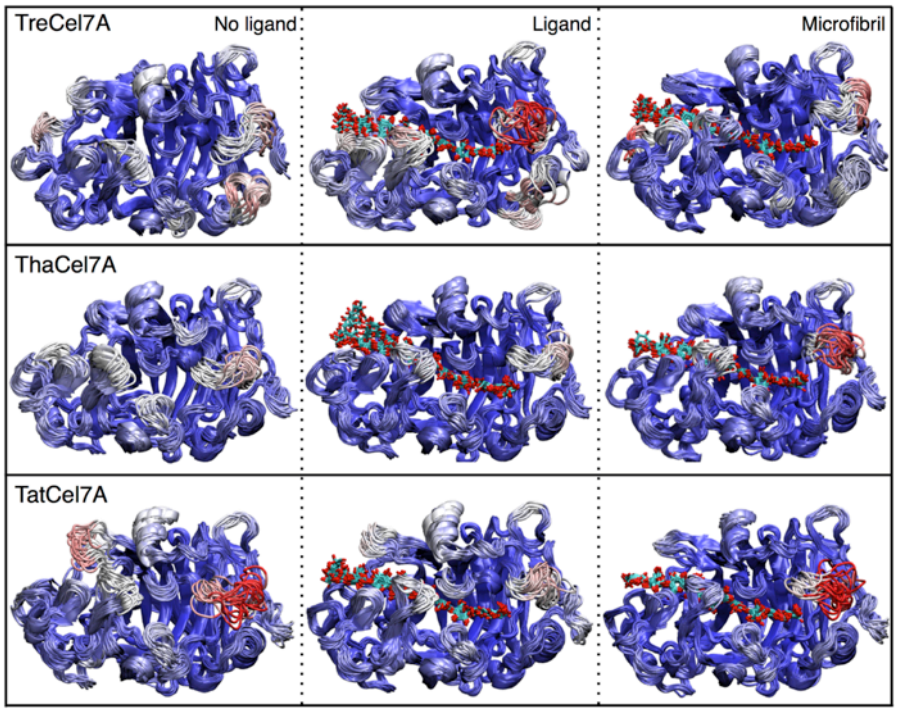


Figure 18. RMSF comparison of the protein backbone at 300 K. Red regions indicate larger fluctuations of the protein backbone (RMSF > 4 Å).

3.2.3 Molecular evolution

To highlight the specific regions in Cel7 sequences from evolutional point of view, the distribution of amino acid variation was analyzed using reverse conservation analysis (RCA; (Lee, 2008)) of GH7 CBH sequences from two groups of related fungi within the order Hypocreales, *Trichoderma* spp. (11 sequences) versus *Fusarium* spp. and *Clonostachys rosea* (6 sequences). When comparing closely related orthologs, amino acid residues that determine functional properties of an enzyme are expected to display higher diversity than other positions due to adaptation (Lee, 2008). Four regions were identified (Figure 19) displaying signs of type 1 functional divergence (i.e. site conserved in one lineage but variable in the other (Cole and Gaucher, 2011)) in *Trichoderma* spp. All sites in sections I, II and III are located at the surface of the protein. All catalytic amino acid residues and the ones, involved into substrate binding, are located in conserved regions with low W mean score (Figure 17). Temperature factors plotted versus residue numbers show the highest fluctuations for A4 loop in *Tat*Cel7A and *Tha*Cel7A, in good agreement with the MD simulations, Figure 18.

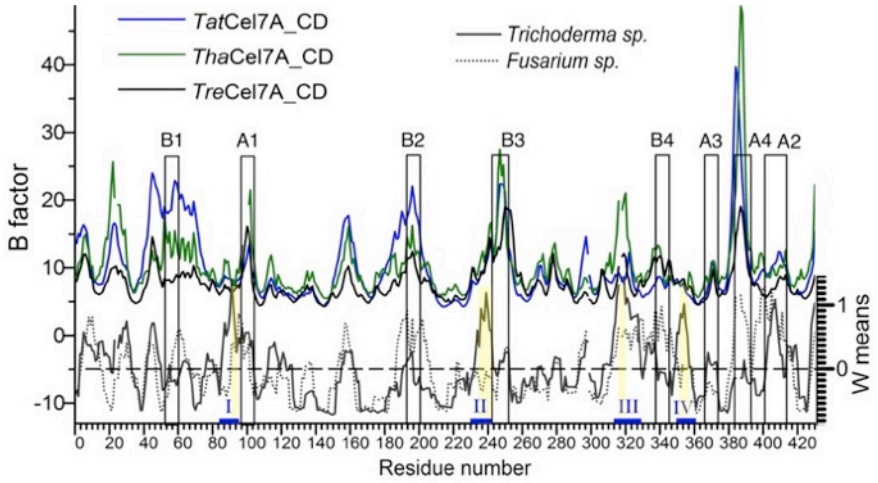


Figure 19. Temperature factors (B factors) for Cel7A structures plotted over W mean scores from RCA analysis versus residue number. The B factors for amino acid Cα atoms of chain A in *Tre*Cel7A (4C4C), *Tha*Cel7A (2YOK) and *Tat*Cel7A (XXX) structures are plotted against residue number in *Tat*Cel7A aligned with the corresponding GH7 sequences (loops are numbered accordingly Figure 18). The scale for W mean scores from RCA analysis is on the right side of the graph.

3.2.4 Conclusions

We hypothesize that combination of the A1 and A3 loop motifs is the main determinant for the observed differences in activity on cellulosic substrates. A single mutation in loop A3 weakens the interaction with the opposing loop B3, which becomes much more flexible. At the base of loop B3, RCA identifies a region showing signs of type 1 functional divergence. Together this indicates that fine-tuning of B3 loop flexibility represents an important evolutionary target in *Trichoderma* spp. Cel7 proteins. Loop B3 has been defined as beneficial in facilitating processive crystalline substrate degradation (von Ossowski et al., 2003) and its absence in GH7 EGs makes endo-initiation more probable than exo-initiation. A4 loop flexibility might be a driving force for the difference in interactions at the product site. Also A2-A4 region, defined as a hotspot for initiation of unfolding in MD simulations, seems to unfold faster in *Tat*Cel7A. GH7 CBHs are highly conserved within the family and functionally important loops seem to be no more a target for development, rather regions supporting the loops can be improved during further evolution of enzymes.

3.3 Sequencing, biochemical characterization, crystal structure and molecular dynamics of cellobiohydrolase Cel7A from *Geotrichum candidum 3C* (Paper III)

After publication of this article, new data have emerged, which show that the fungal strain previously called *Geotrichum candidum 3C* rather belongs to the genus *Scytalidium* (Kulminskaya et al, manuscript in preparation). Therefore, in the thesis and in paper IV, the name *Scytalidium sp.* is used instead, and cellobiohydrolase Cel7A from *Geotrichum candidum 3C*, *Gca*Cel7A*,* is named *Scy*Cel7A, accordingly. Furthermore, in Paper III the fungus *Trichoderma reesei* was called by its teleomorph name *Hypocrea jecorina*. and thus the *Hje*Cel7A enzyme in Paper III is the same protein as *Tre*Cel7A.

Optimization and development of enzyme cocktails applied in biofuel production is a challenging task in biotechnology. Discovery of new efficient GH7 CBHs and their utilization in bioreactors enlarge significantly a potential of cellulose-based cocktails. Novel GH7 CBH from *Scytalidium sp* (*Scy*Cel7A), (former *Geotrichum candidum 3C, Gca*Cel7A) shares 64% sequence identity with the archetypal industrial GH7 CBH of *Trichoderma reesei* (*Tre*Cel7A). The bimodular cellulase consists of a CBM1 carbohydrate binding module and long linker connected to a GH7 catalytic domain. Five crystal structures, with and without bound thio-oligosaccharides were solved and examined for conformational diversity of tunnel-enclosing loops and substrate binding. Detailed structural-functional studies were carried out, followed by MD simulations in order to predict processivity and endo-initiation probability of the enzyme.

3.3.1 Isolation and identification of *Scy*Cel7A

The major protein was isolated from culture filtrate of *Scytalidium sp* (former *Geotrichum candidum 3C*), grown on filter paper as sole carbon source. Purification steps included cellulose affinity, ion-exchange and hydrophobic interaction chromatograpy. The yield was 32 mg of purified protein per litre of culture with a specific activity against crystalline cellulose of $2.15 \cdot 10^{-3}$ U/mg. Trypsin digestion and peptide mapping by MALDI-TOF (data not shown) identified the enzyme as a member of GH family 7, further called *Gca*Cel7A and, in current thesis, *Scy*Cel7A. Isolation of catalytic domain (*Scy*Cel7A_CD)

was done using size-exclusion chromatography, after the full-length enzyme was proteolytically cleaved with papaine. In terms of modularity, the linker peptide connecting the catalytic domain with the CBM in *Scy*Cel7A is significantly longer than that of *Tre*Cel7A, and is comprised by 47 residues with an unusually large number of positively charged amino acids.

3.3.2 Initial biochemical characterization

General biochemical characterization of *Scy*Cel7A included pH and temperature optimum, substrate specificity, activity on crystalline cellulose and inhibition studies. From a functional standpoint, the activity pH profile of *Scy*Cel7A is broader and slightly shifted in the alkaline direction, in comparison to *Tre*Cel7A (Boer and Koivula, 2003, Becker et al., 2001) (Paper III). Activity comparisons of *Scy*Cel7A and *Tre*Cel7A were done with Avicel as substrate, both with the CBH acting alone and added to a commercial cellulase cocktail (Accellerase®) from which *Tre*Cel7A had been selectively removed. In both cases, *Scy*Cel7A yielded similar amounts of soluble sugar as *Tre*Cel7A, within the standard error of the experiment (Table 4).

Table 4. *Comparison of hydrolysis of 5 mg/mL Avicel cellulose by ScyCel7A and TreCel7A, acting alone (50 ug/mL) or added to a Cel7A-depleted T. reesei enzyme cocktail (25 ug/mL Cel7A + 25 ug/mL TreCel7A-free Accellerase 1500™), during 2 hours at 40 °C, pH 5.0.*[a)]

Enzyme	Cel7A alone		Cel7A + TreCel7A-free Accellerase	
	[Glc] mg/ml [b)]	Conversion %	[Glc] mg/ml [b)]	Conversion %
*Scy*Cel7A	0.39 ±0.021	6.9	0.65 ±0.072	11.7
*Tre*Cel7A	0.42 ±0.025	7.6	0.75 ±0.039	13.5

[a)] Excess β-glucosidase was subsequently added to convert all soluble sugars to glucose prior to the reducing sugar assay.
[b)] Average and standard deviation of five replicates.

Enzyme kinetics data show that cellobiose product inhibition is about two-fold weaker for *Scy*Cel7A than *Tre*Cel7A, which is somewhat surprising given the similarity of the product binding sites. Difference in inhibition/binding at the product site were observed in Paper I and II as well. Otherwise *Scy*Cel7A is biochemically similar to the well-characterised *Tre*Cel7A and shows comparable activity on crystalline cellulose.

3.3.3 ScyCel7A structures and MD simulations

The isolated catalytic domain, *Scy*Cel7A_CD, was successfully crystallised and we report five crystal structures: two apo structures (referred to as APO1 and APO2) and three ligand complexes, with cellobioside (G2), cellotrioside (G3) and cellotetraoside (G4) bound at the active site. All the structures were solved in the monoclinic space group $P2_1$ with one protein chain per asymmetric unit, but representing two discrete crystal packings differing by ~10 Å in length of the unit cell b-axis. The longer b-axis is observed in the APO1 and G4 structures. There is distinct electron density for cello-oligosaccharides bound in the active site of the *Scy*Cel7A _CD ligand complexes (Figure 20).

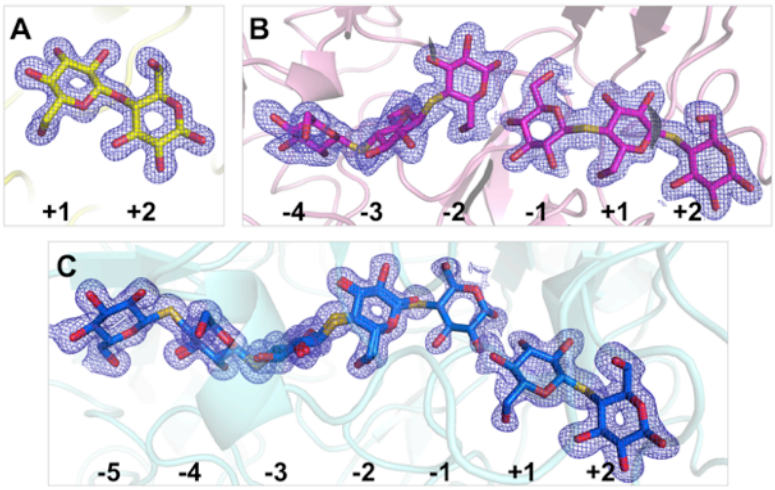


Figure 20. Electron density for the ligands in the *Scy*Cel7A G2, G3 and G4 structures. (A) The G2 structure shows cellobiose bound in the product binding subsites +1/+2. (B) The G3 complex shows two molecules of thio-cellotriose, at subsites -4/-3/-2 and -1/+1/+2, respectively. (C) In the G4 complex, overlapping thio-cellotetraose at partial occupancy were refined at subsites -5 to -2 and -4 to -1, respectively, and thio-cellobiose at +1/+2. At subsite -1, the density indicates the presence of multiple binding modes. All glucosyl units adopt the $^4C_1$ chair conformation. Sigma-averaged $2F_o$-F_c electron density maps are contoured at 0.53 e/A^3 in (A) and at 0.26 e/A^3 in (B) and (C).

Overall, the *Scy*Cel7A structures are very similar, as reflected by low pairwise RMSD values (0.16-0.35 Å). Notable differences occur at loop B2, which adopts a new conformation in APO2 structure, not previously seen in any Cel7 structure (Figure 21A). There are differences in B2/A3/B3 contacts across the active site, driven by Tyr374 flipped conformation (Figure 21B).

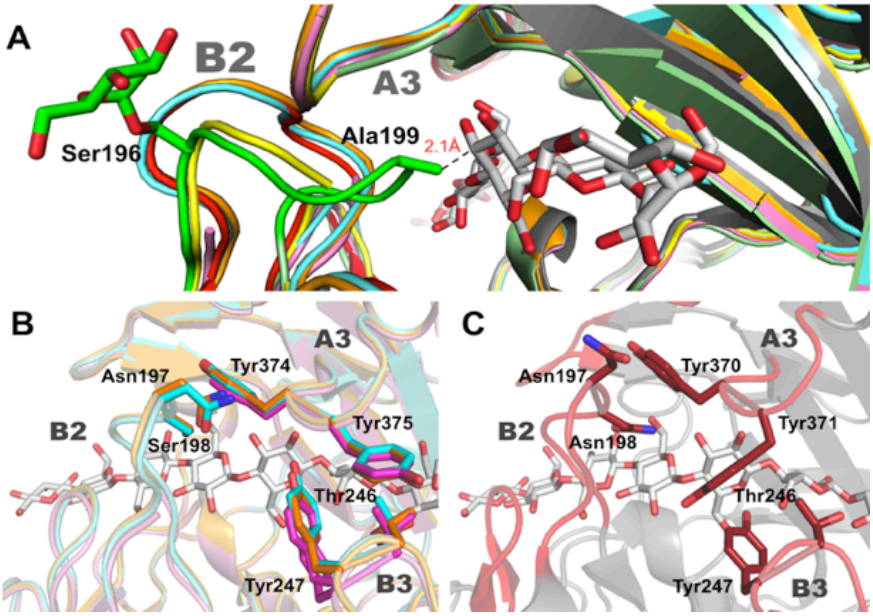


Figure 21. Tunnel-enclosing loop contacts. (A) In the *Scy*Cel7A APO2 structure (green), loop B2 bends into the tunnel and Ala-199 at the tip of the loop would clash with a glucosyl binding at subsite -4. (B) B2/A3/B3 loop contact region in the *Scy*Cel7A APO1, G3 and G4 structures. (C) B2/A3/B3 loop contacts in *Tre*Cel7A (G9 complex; PDB code 4C4C; (Knott et al., 2013)). Two tyrosines at the tip of loop A3 play an important role for tunnel-enclosing contacts with loop B2 and B3 across the active site. In *Scy*Cel7, the Tyr-375 side chain is flipped towards the +1/+2 product sites, relative to the position of Tyr-371 in *Tre*Cel7A above subsite -1, accompanied with a slight shift outwards of loop B3 and Tyr-247 at its tip. The coloring scheme is the same as in Figure 5: *Tre*Cel7A protein and cellononaose ligand, light grey, with loops in red; *Scy*Cel7A APO1, orange; APO2, green; G2, yellow; G3, magenta; G4, cyan.

The *Scy*Cel7A structure represents the first report of *O*-glycosylation of a GH7 catalytic domain, where a mannose residue is attached to Ser-196 near the tip of loop B2. The *O*-glycosylation is visible in the structures with a short unit cell b-axis (APO2 and G2) probably due to tight crystal contacts in this region, which obstructs the B2 loop conformation and restricts the mobility of the attached mannose residue. Exposed Ser and Thr as potential *O*-glycosylation sites are quite common in this region of loop B2 throughout GH7 CBHs, suggesting that *O*-glycosylation here may not be unique to ScyCel7A, but may also occur in other Cel7s, although not visilble in any other crystal structure. It should be noted that analysis of glycosylation in GH7 has only been reported for a very limited number of GH7 enzymes and almost exclusively concerns N-glycosylation (Andey, 2009, Gao et al., 2012, Garcia-Viloca et al., 2004, Gusakov et al., 2017, Dotsenko et al., 2016).

The ability to conduct endo-initiated attack of crystalline substrate is related to both flexibility and the length of this loop (B2), along with that of the nearby loop B3. Both of these loops must open sufficiently to allow the entry into the active site of an internal part of a cellulose chain. For example, *Pch*Cel7D, which exhibits relatively high probability of endo- inititiation, and *Hir*Cel7A, display a broad range of loop fluctuations in MD simulations (Momeni et al., 2013). Examination of the minimum distance between loops B2 and A3 allows to distinguish between endo- and exo- mode of action in GH7 CBHs. The MD simulations suggest that *Scy*Cel7A demonstrates similar relative behavior as *Tre*Cel7A, with comparatively high preference for exo-initiated attack of crystalline cellulose substrates (Figure 22).

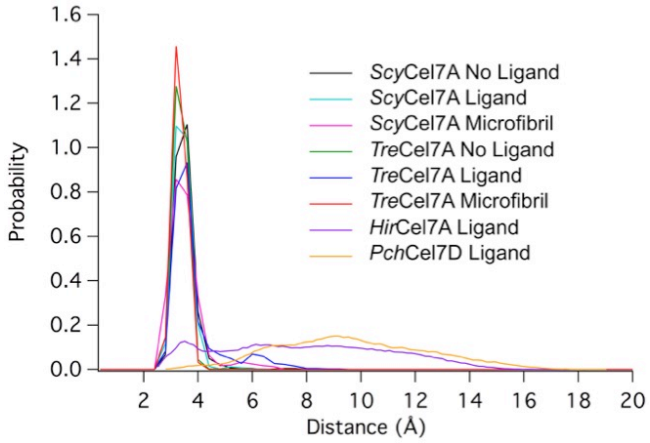


Figure 22. Histograms of the minimum distance between loops B2 and A3 from 100-ns MD simulations of *Scy*Cel7A, *Tre*Cel7A, *Hir*Cel7A, and *Pch*Cel7D. These two loops are thought to be critical in endo-initiated catalysis. In the case of *Scy*Cel7A and *Tre*Cel7A, the distances have been measured in the absence of a ligand, bound to cellononaose, and bound to a cellulose microfibril. The simulations of *Hir*Cel7A and *Pch*Cel7D examine the behavior of the enzymes bound to cellononaose and have previously been reported by Momeni et al. (Momeni et al., 2013). In all cases, the distances have been measured based on the minimum distance between each loop.

3.3.4 Conclusions

We find that *Scy*Cel7A exhibits similar structural and functional characteristics to the industrially relevant *Tre*Cel7A. Nevertheless, reduced product inhibition and a broader optimal pH range may be valuable industrial advantages over *Tre*Cel7A in the development of cellulase products. The first *O*-glycosylation in GH7 structures and remarkable conformational diversity of loop B2 was reported in the study.

3.4 Crystal structures of mechanism-based affinity labelled GH7 cellobiohydrolases (Paper IV)

High similarity of the cellulose-binding tunnel architecture of GH7 CBHs allows for design of universal affinity tags for specific labeling and quantification of these enzymes in complex samples. An affinity tag (TSRG14) containing a cellobiose core structure modified by a C-glycosidic bromoketone warhead, a triazolyl butyl selectivity element and a fluorescence probe has been designed based on substitution group library screening (Zierke M, Rasmussen TS, Withers SG, unpublished results). The fluorophore allows to profile and quantitate active CBHs in complex mixtures in order to monitor biomass conversion in a bioreactor.

Co-crystallisation of GH7 CBH catalytic domains with such a mechanism-based affinity label provides deep insight into the mechanism of inactivation and interactions within the cellulose-binding tunnel. Three structures, one of *Scytalidium* sp Cel7A and two of *Trichoderma reesei* Cel7A, wildtype (*Tre*Cel7A-WT) and acid/base-crippled E217Q mutant (*Tre*Cel7A-E217Q), were solved and affinity label positioning and interactions were described.

3.4.1 Mechanism of inactivation

According to the proposed mechanism for inactivation, the cellobiosyl unit directs productive binding at subsites -2/-1 and positions the bromoketone for reaction with the catalytic acid/base glutamic acid residue of GH7 CBHs (Figure 23). Nucleophilic substitution replaces the bromine with a covalent bond to the carboxyl group of the glutamate sidechain. The tag remains covalently attached to the enzyme, which is thus effectively inactivated.

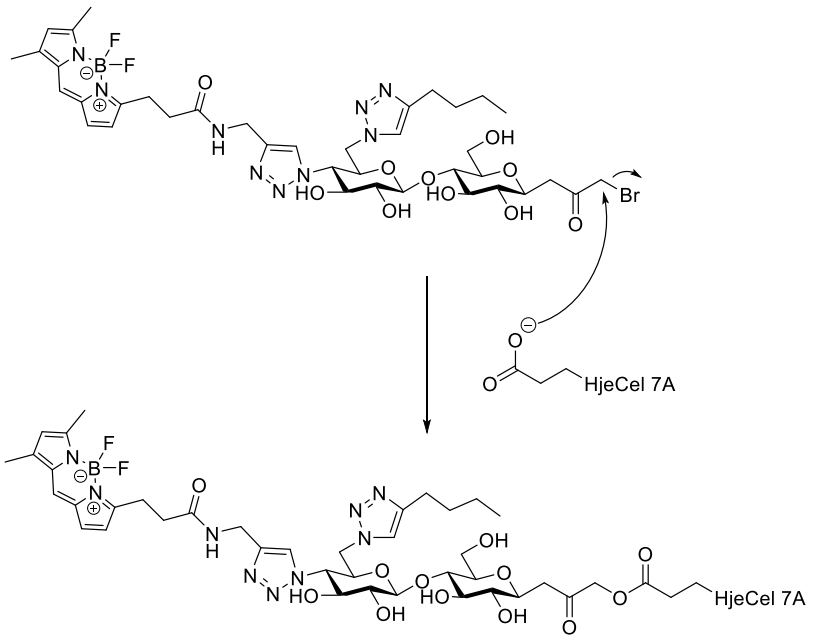


Figure 23. Mechanism-based inactivation of Cel7A. Chemical structure of affinity tag TSRG14 and proposed reaction mechanism for specific inactivation and mechanism-based affinity labelling of GH7 enzymes.

3.4.2 Crystal structures

Here we report the crystal structures of three different ligand complexes with TSRG14 bound at the active site. In the structures with *Scy*Cel7A and *Tre*Cel7A-WT, the ligand is covalently linked to the catalytic acid/base, but not with *Tre*Cel7A-E217Q, thereby illustrating the binding both before and after the inactivation reaction.

The SCY structure (*Scy*Cel7A/TSRG14 complex) is refined at 1.25 Å resolution and R_{work}/R_{free} of 0.194/0.206, and no convincing density is observed for *O*-glycosylation at Ser196, reported in Paper III. The TRE structure (*Tre*Cel7A-WT/TSRG14 complex) is solved in a new space group for this protein, P21, with 4 chains, A-D, in the asymmetric unit. The unit cell dimensions are similar to those for the 217Q structure and previous *Tre*Cel7A structures solved in orthorhombic space groups, but the β-angle is 101° instead of 90°. The structure is refined at 2.0 Å resolution and R_{work}/R_{free} of 0.204/0.259. The 217Q structure (*Tre*Cel7A-E217Q/TSRG14 complex) is solved in space group I222 with one chain per asymmetric unit, which is the

most common among previous *Tre*Cel7A structures. The structure is refined at 1.56 Å resolution and R_{work}/R_{free} of 0.221/0.234.

For the cellobiosyl moiety, the position and conformation at subsites -2/-1 is practically identical in all the TSRG14 complexes, including all four chains of the TRE structure (Figure 24). Also, the binding is very similar to the -2/-1 Glc units of the *Tre*Cel7A Michaelis complex (4C4C; (Knott et al., 2014); Figure 24) with <0.5 Å distance between the atoms that deviate most (C4 and C3 of the -2 glucoside). Thus, it seems indeed that the cellobiosyl unit can direct proper binding and positioning of TSRG14 at the active site.

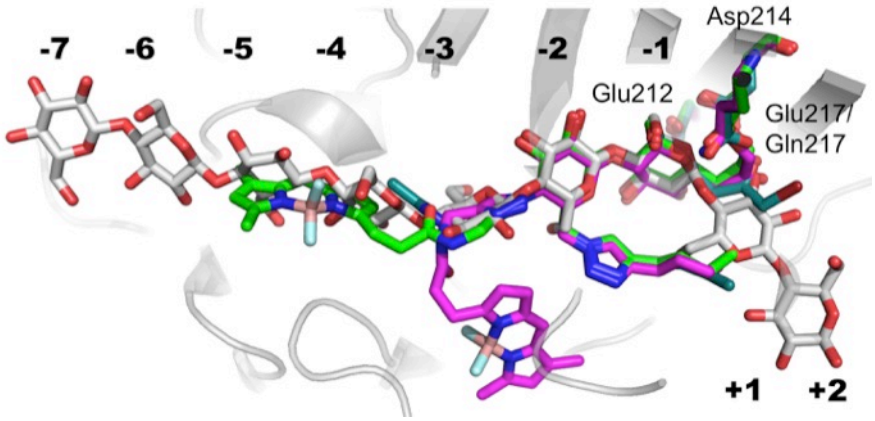


Figure 24. Comparison of TSRG14 and cellulose binding. The cellononaose chain from the *Tre*Cel7A Michaelis complex (light-grey; PDB code 4C4C; (Knott et al., 2014)) is superposed with TSRG14 bound in *Scy*Cel7A (magenta), *Tre*Cel7A-WT (green) and *Tre*Cel7A-E217Q (cyan).

The main difference between *Scy*Cel7A and *Tre*Cel7A is the binding of the fluorophore arm. In SCY it bends outwards from the tunnel and the BODIPY unit is bound between loops A3 and B3 above the active site, in contact with Tyr374 and Tyr375 on A3 and Tyr247 on B3 (Figure 25A). This is in contrast to TRE (chain A) where the arm instead extends inside the tunnel towards the entrance, with BODIPY at subsites -4/-5 and partly stacking with the -4 platform Trp38 (Figure 25B). Although the position of the fluorophore is not defined in E217Q and TRE chains B-D, it is likely to reside inside the tunnel in these structures as well.

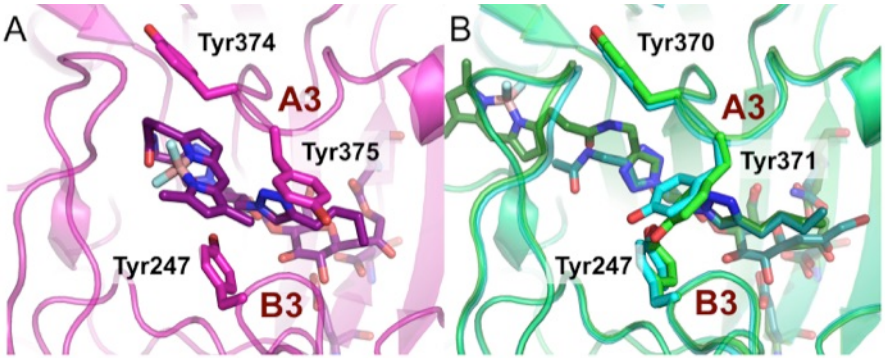


Figure 25. Region of contact between loops A3 and B3 above the catalytic centre in Cel7A TSRG14 ligand complexes. (A) In SCY the BODIPY group can point outwards and bind between the A3 and B3 loops, because Tyr375 adopts a conformation above the product site. The same Tyr375 conformation was observed in all previous *Scy*Cel7A structures (Borisova et al., 2015). (B) The corresponding Tyr371 in *Tre*Cel7A is positioned above subsite -2 in the TRE (green) and 217Q (cyan) structures, in direct contact with Tyr247 on loop B3. There is not room for the BODIPY group between loops A3 and B3, which instead extends within the cellulose-binding tunnel towards the entrance.

3.4.3 Conclusions

The cellobiosyl unit obviously plays a central role for the positioning of the affinity ligand. All hydrogen bonds of the ligand directly to protein are formed by the sugar hydroxyls, except for interactions of the ketone oxygen with Asp214 and His228. In addition there are hydrophobic interactions with the Trp platforms at -1 and +1 (and -4 in TRE), as well as numerous van der Waals interactions with surrounding protein atoms (Paper IV). Two possible binding of BODIPY are driven by tyrosine conformations on the tip of loop A3; even though the configuration of the fluorophore arm was designed to follow the substrate-binding tunnel, it apparently can position itself in between the loops, as seen in SCY structure.

4 Conclusions and future perspectives

The thesis work has been focused on exploration of structure/function relationship in GH7 CBHs, which are the cornerstones of lignocellulose saccharification in nature and in industry. Several new GH7 CBH structures were solved, both close orthologs to the archetypal *T. reesei* Cel7A, as well as from the distant Amoebozoa, and another ascomycete fungus *Scytalidium* sp. These data gave a substantial increase in the number of known GH7 CBH structures and their functional diversity. Structural and computational approaches were combined to elucidate structure/function correlations in different GH7 CBHs. An important conclusion is that the structures are highly similar, but MD simulations revealed that essential differences can be observed in loop dynamics. We hypothesize that combination of the A1 and A3 loop motifs is the main determinant for the efficiency of Cel7s in cellulose degradation, while A4 loop flexibility might affect binding in the product site. Also, the A2-A4 region was defined as a hotspot for initiation of unfolding in MD simulations. Five new structures of *Scy*Cel7A show that the B2 loop can adopt different more or less closed conformations, and even bend into the tunnel and obstruct cellulose binding. *O*-glycosylation was observed on loop B2, that may influence both the loop itself and interactions with the cellulose surface. The sugar is only visible in some of the ScyCel7A structures and we hypothesize that other GH7 CBHs may also be *O*-glycosylated in similar positions. We propose that probability for opening of loop B2 responsible for endo-initiation differs significantly depending on the length of loop B2, and the sequence at the tip of the loop, as well as the sequence at the tip of the opposing A3 loop across the tunnel.

Compared to other GH families the sequence similarity within GH7 family is very high, despite 1 billion years of evolution. It seems that GH7 CBH structure cannot tolerate large sequence diversity and still maintain the function. Differences related to function are found primarily in small sequence

variations of loop regions around the active site, resulting in local variations in structure and dynamics. Based on results of RCA analysis, B3 loop flexibility is suggested to represent an important evolutionary target in *Trichoderma* spp. The role of GH7 CBHs in Dictyostelia, whether it is related to endogenous cellulose metabolism or to the ability to feed on cellulose from other organisms, remains a question for further investigation of evolutionary pressures on the activities of slime mold GH7 CBHs.

Three new structures of labelled GH7 CBHs revealed that the cellobiosyl unit obviously plays a central role for the positioning of the affinity ligand, and that the substrate-binding tunnel can accommodate the triazolyl butyl selectivity element as well as the bulky fluorescence label. Such an affinity tag allows specific labelling and quantification of different GH7 CBHs in bioreactors.

Based on results of the thesis work, future perspectives should include processivity measurements on a larger set of enzymes, combined with MD simulations data, RCA analysis of a larger set of enzymes, phylogenetic analysis of GH7s together with GH16 and more detailed investigations of glycosylation in GH7s. Studies of *O*- and *N*-glycosylations in Cel7s should be carried out in order to reveal their role in loop dynamics and processivity of enzymes. Measurements of the effect of differences in sequence motifs on elementary kinetic parameters of cellulose hydrolysis (k_{off}, processivity, endo/exo initiation ratio) as well as performance on lignocellulose in the context of a synergistic enzyme cocktail should be done in order to find sequence-performance correlations with sequence motifs. The influence of characteristic loop and sequence motifs should be systematically investigated, for example by grafting into a common GH7 CBH framework, for example *Tre*Cel7A. RCA analysis on sufficient number of GH7 sequences should be carried out to understand the process of evolution within GH7 family.

References

ALLEN, F., ANDREOTTI, R., EVELEIGH, D. E. & NYSTROM, J. 2009. Mary Elizabeth Hickox Mandels, 90, bioenergy leader. *Biotechnology for Biofuels,* 2.

ANDEY, W. S., JEOH, T., BECKHAM, G.T., CHOU, Y.-C., BAKER, J.O., MICHENER, W., BRUNECKY, R., HIMMEL, M.E. 2009. Probing the role of N-linked glycans in the stability and activity of fungal cellobiohydrolases by mutational analysis. *Cellulose,* 16**,** 669-709.

ATALLA, R. H. & VANDERHART, D. L. 1984. Native cellulose: a composite of two distinct crystalline forms. *Science,* 223**,** 283-5.

BECKER, D., BRAET, C., BRUMER, H., 3RD, CLAEYSSENS, M., DIVNE, C., FAGERSTROM, B. R., HARRIS, M., JONES, T. A., KLEYWEGT, G. J., KOIVULA, A., MAHDI, S., PIENS, K., SINNOTT, M. L., STAHLBERG, J., TEERI, T. T., UNDERWOOD, M. & WOHLFAHRT, G. 2001. Engineering of a glycosidase Family 7 cellobiohydrolase to more alkaline pH optimum: the pH behaviour of Trichoderma reesei Cel7A and its E223S/ A224H/L225V/T226A/D262G mutant. *Biochem J,* 356**,** 19-30.

BECKHAM, G. T., BOMBLE, Y. J., MATTHEWS, J. F., TAYLOR, C. B., RESCH, M. G., YARBROUGH, J. M., DECKER, S. R., BU, L., ZHAO, X., MCCABE, C., WOHLERT, J., BERGENSTRAHLE, M., BRADY, J. W., ADNEY, W. S., HIMMEL, M. E. & CROWLEY, M. F. 2010. The O-glycosylated linker from the Trichoderma reesei Family 7 cellulase is a flexible, disordered protein. *Biophys J,* 99**,** 3773-81.

BERNEY, C. & PAWLOWSKI, J. 2006. A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proceedings of the Royal Society B: Biological Sciences,* 273**,** 1867-1872.

BISCHOF, R. H., RAMONI, J. & SEIBOTH, B. 2016. Cellulases and beyond: the first 70 years of the enzyme producer Trichoderma reesei. *Microb Cell Fact,* 15**,** 106.

BOER, H. & KOIVULA, A. 2003. The relationship between thermal stability and pH optimum studied with wild-type and mutant Trichoderma reesei cellobiohydrolase Cel7A. *Eur J Biochem,* 270**,** 841-8.

BORISOVA, A. S., ENEYSKAYA, E. V., BOBROV, K. S., JANA, S., LOGACHEV, A., POLEV, D. E., LAPIDUS, A. L., IBATULLIN, F. M., SALEEM, U., SANDGREN, M., PAYNE, C. M., KULMINSKAYA, A.

A. & STAHLBERG, J. 2015. Sequencing, biochemical characterization, crystal structure and molecular dynamics of cellobiohydrolase Cel7A from Geotrichum candidum 3C. *FEBS J,* 282**,** 4515-37.
BÖRJESSON, M. W., G. 2015. Crystalline Nanocellulose — Preparation, Modification, and Properties. *In:* POLETTO, M. (ed.) *Cellulose - Fundamental Aspects and Current Trends.* InTech.
BREVNOVA, E. E., FLATT, J., GANDHI, C., RAJGARHIA, V., MCBRIDE, J. & WARNER, A. 2013. *Isolation and characterization of Schizochytrium aggregatum cellobiohydrolase I (Cbh 1), US Patent 8,470,592 B2*.
BUSK, P. K., LANGE, M., PILGAARD, B. & LANGE, L. 2014. Several genes encoding enzymes with the same activity are necessary for aerobic fungal degradation of cellulose in nature. *PLoS One,* 9**,** e114138.
CARPITA, N. C. & GIBEAUT, D. M. 1993. Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J,* 3**,** 1-30.
CHANG, Y., WANG, S., SEKIMOTO, S., AERTS, A. L., CHOI, C., CLUM, A., LABUTTI, K. M., LINDQUIST, E. A., YEE NGAN, C., OHM, R. A., SALAMOV, A. A., GRIGORIEV, I. V., SPATAFORA, J. W. & BERBEE, M. L. 2015. Phylogenomic Analyses Indicate that Early Fungi Evolved Digesting Cell Walls of Algal Ancestors of Land Plants. *Genome Biol Evol,* 7**,** 1590-601.
CHUNDAWAT, S. P., BECKHAM, G. T., HIMMEL, M. E. & DALE, B. E. 2011a. Deconstruction of lignocellulosic biomass to fuels and chemicals. *Annu Rev Chem Biomol Eng,* 2**,** 121-45.
CHUNDAWAT, S. P., BELLESIA, G., UPPUGUNDLA, N., DA COSTA SOUSA, L., GAO, D., CHEH, A. M., AGARWAL, U. P., BIANCHETTI, C. M., PHILLIPS, G. N., JR., LANGAN, P., BALAN, V., GNANAKARAN, S. & DALE, B. E. 2011b. Restructuring the crystalline cellulose hydrogen bond network enhances its depolymerization rate. *J Am Chem Soc,* 133**,** 11163-74.
COLE, M. F. & GAUCHER, E. A. 2011. Utilizing natural diversity to evolve protein function: applications towards thermostability. *Curr Opin Chem Biol,* 15**,** 399-406.
CRAGG, S. M., BECKHAM, G. T., BRUCE, N. C., BUGG, T. D. H., DISTEL, D. L., DUPREE, P., ETXABE, A. G., GOODELL, B. S., JELLISON, J., MCGEEHAN, J. E., MCQUEEN-MASON, S. J., SCHNORR, K., WALTON, P. H., WATTS, J. E. M. & ZIMMER, M. 2015. Lignocellulose degradation mechanisms across the Tree of Life. *Current Opinion in Chemical Biology,* 29**,** 108-119.
CRUYS-BAGGER, N., REN, G., TATSUMI, H., BAUMANN, M. J., SPODSBERG, N., ANDERSEN, H. D., GORTON, L., BORCH, K. & WESTH, P. 2012. An amperometric enzyme biosensor for real-time measurements of cellobiohydrolase activity on insoluble cellulose. *Biotechnol Bioeng,* 109**,** 3199-204.
CRUYS-BAGGER, N., TATSUMI, H., REN, G. R., BORCH, K. & WESTH, P. 2013. Transient kinetics and rate-limiting steps for the processive

cellobiohydrolase Cel7A: effects of substrate structure and carbohydrate binding domain. *Biochemistry,* 52**,** 8938-48.

DE CASTRO, A. M., PEDRO, K. C., DA CRUZ, J. C., FERREIRA, M. C., LEITE, S. G. & PEREIRA, N., JR. 2010. Trichoderma harzianum IOC-4038: A promising strain for the production of a cellulolytic complex with significant beta-glucosidase activity from sugarcane bagasse cellulignin. *Appl Biochem Biotechnol,* 162**,** 2111-22.

DIVNE, C., STAHLBERG, J., REINIKAINEN, T., RUOHONEN, L., PETTERSSON, G., KNOWLES, J. K., TEERI, T. T. & JONES, T. A. 1994. The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma reesei. *Science,* 265**,** 524-8.

DOTSENKO, A. S., GUSAKOV, A. V., VOLKOV, P. V., ROZHKOVA, A. M. & SINITSYN, A. P. 2016. N-linked glycosylation of recombinant cellobiohydrolase I (Cel7A) from Penicillium verruculosum and its effect on the enzyme activity. *Biotechnol Bioeng,* 113**,** 283-91.

DUCHESNE, L. C. & LARSON, D. W. 1989. Cellulose and the Evolution of Plant Life. *Bioscience,* 39**,** 238-241.

EASTWOOD, D. C., FLOUDAS, D., BINDER, M., MAJCHERCZYK, A., SCHNEIDER, P., AERTS, A., ASIEGBU, F. O., BAKER, S. E., BARRY, K., BENDIKSBY, M., BLUMENTRITT, M., COUTINHO, P. M., CULLEN, D., DE VRIES, R. P., GATHMAN, A., GOODELL, B., HENRISSAT, B., IHRMARK, K., KAUSERUD, H., KOHLER, A., LABUTTI, K., LAPIDUS, A., LAVIN, J. L., LEE, Y. H., LINDQUIST, E., LILLY, W., LUCAS, S., MORIN, E., MURAT, C., OGUIZA, J. A., PARK, J., PISABARRO, A. G., RILEY, R., ROSLING, A., SALAMOV, A., SCHMIDT, O., SCHMUTZ, J., SKREDE, I., STENLID, J., WIEBENGA, A., XIE, X. F., KUES, U., HIBBETT, D. S., HOFFMEISTER, D., HOGBERG, N., MARTIN, F., GRIGORIEV, I. V. & WATKINSON, S. C. 2011. The Plant Cell Wall-Decomposing Machinery Underlies the Functional Diversity of Forest Fungi. *Science,* 333**,** 762-765.

EICHINGER, L., PACHEBAT, J. A., GLOCKNER, G., RAJANDREAM, M. A., SUCGANG, R., BERRIMAN, M., SONG, J., OLSEN, R., SZAFRANSKI, K., XU, Q., TUNGGAL, B., KUMMERFELD, S., MADERA, M., KONFORTOV, B. A., RIVERO, F., BANKIER, A. T., LEHMANN, R., HAMLIN, N., DAVIES, R., GAUDET, P., FEY, P., PILCHER, K., CHEN, G., SAUNDERS, D., SODERGREN, E., DAVIS, P., KERHORNOU, A., NIE, X., HALL, N., ANJARD, C., HEMPHILL, L., BASON, N., FARBROTHER, P., DESANY, B., JUST, E., MORIO, T., ROST, R., CHURCHER, C., COOPER, J., HAYDOCK, S., VAN DRIESSCHE, N., CRONIN, A., GOODHEAD, I., MUZNY, D., MOURIER, T., PAIN, A., LU, M., HARPER, D., LINDSAY, R., HAUSER, H., JAMES, K., QUILES, M., MADAN BABU, M., SAITO, T., BUCHRIESER, C., WARDROPER, A., FELDER, M., THANGAVELU, M., JOHNSON, D., KNIGHTS, A., LOULSEGED, H., MUNGALL, K., OLIVER, K., PRICE, C., QUAIL, M. A., URUSHIHARA, H., HERNANDEZ, J., RABBINOWITSCH, E.,

STEFFEN, D., SANDERS, M., MA, J., KOHARA, Y., SHARP, S., SIMMONDS, M., SPIEGLER, S., TIVEY, A., SUGANO, S., WHITE, B., WALKER, D., WOODWARD, J., WINCKLER, T., TANAKA, Y., SHAULSKY, G., SCHLEICHER, M., WEINSTOCK, G., ROSENTHAL, A., COX, E. C., CHISHOLM, R. L., GIBBS, R., LOOMIS, W. F., PLATZER, M., KAY, R. R., WILLIAMS, J., DEAR, P. H., NOEGEL, A. A., BARRELL, B. & KUSPA, A. 2005. The genome of the social amoeba *Dictyostelium discoideum*. *Nature,* 435**,** 43-57.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution***,** 783-791.

FLOUDAS, D., BINDER, M., RILEY, R., BARRY, K., BLANCHETTE, R. A., HENRISSAT, B., MARTINEZ, A. T., OTILLAR, R., SPATAFORA, J. W., YADAV, J. S., AERTS, A., BENOIT, I., BOYD, A., CARLSON, A., COPELAND, A., COUTINHO, P. M., DE VRIES, R. P., FERREIRA, P., FINDLEY, K., FOSTER, B., GASKELL, J., GLOTZER, D., GORECKI, P., HEITMAN, J., HESSE, C., HORI, C., IGARASHI, K., JURGENS, J. A., KALLEN, N., KERSTEN, P., KOHLER, A., KUES, U., KUMAR, T. K., KUO, A., LABUTTI, K., LARRONDO, L. F., LINDQUIST, E., LING, A., LOMBARD, V., LUCAS, S., LUNDELL, T., MARTIN, R., MCLAUGHLIN, D. J., MORGENSTERN, I., MORIN, E., MURAT, C., NAGY, L. G., NOLAN, M., OHM, R. A., PATYSHAKULIYEVA, A., ROKAS, A., RUIZ-DUENAS, F. J., SABAT, G., SALAMOV, A., SAMEJIMA, M., SCHMUTZ, J., SLOT, J. C., ST JOHN, F., STENLID, J., SUN, H., SUN, S., SYED, K., TSANG, A., WIEBENGA, A., YOUNG, D., PISABARRO, A., EASTWOOD, D. C., MARTIN, F., CULLEN, D., GRIGORIEV, I. V. & HIBBETT, D. S. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. *Science,* 336**,** 1715-9.

FRANKOVA, L. & FRY, S. C. 2011. Phylogenetic variation in glycosidases and glycanases acting on plant cell wall polysaccharides, and the detection of transglycosidase and trans-beta-xylanase activities. *Plant J,* 67**,** 662-81.

FREEZE, H. & LOOMIS, W. F. 1977a. Isolation and characterization of a component of the surface sheath of Dictyostelium discoideum. *J Biol Chem,* 252**,** 820-4.

FREEZE, H. & LOOMIS, W. F. 1977b. The role of the fibrillar component of the surface sheath in the morphogenesis of Dictyostelium discoideum. *Dev Biol,* 56**,** 184-94.

GAO, L., GAO, F., WANG, L., GENG, C., CHI, L., ZHAO, J. & QU, Y. 2012. N-glycoform diversity of cellobiohydrolase I from Penicillium decumbens and synergism of nonhydrolytic glycoform in cellulose degradation. *J Biol Chem,* 287**,** 15906-15.

GARCIA-VILOCA, M., GAO, J., KARPLUS, M. & TRUHLAR, D. G. 2004. How enzymes work: analysis by modern rate theory and computer simulations. *Science,* 303**,** 186-95.

GARDNER, K. H. & BLACKWELL, J. 1975. Refinement of the structure of beta-chitin. *Biopolymers,* 14**,** 1581-95.

GILBERT, H. J. 2010. The biochemistry and structural biology of plant cell wall deconstruction. *Plant Physiol,* 153**,** 444-55.
GRIGOREVSKI-LIMA, A. L., DE OLIVEIRA, M. M., DO NASCIMENTO, R. P., BON, E. P. & COELHO, R. R. 2013. Production and partial characterization of cellulases and Xylanases from Trichoderma atroviride 676 using lignocellulosic residual biomass. *Appl Biochem Biotechnol,* 169**,** 1373-85.
GUSAKOV, A. V., DOTSENKO, A. S., ROZHKOVA, A. M. & SINITSYN, A. P. 2017. N-Linked glycans are an important component of the processive machinery of cellobiohydrolases. *Biochimie,* 132**,** 102-108.
HAREL, Y. M., MEHARI, Z. H., RAV-DAVID, D. & ELAD, Y. 2014. Systemic resistance to gray mold induced in tomato by benzothiadiazole and Trichoderma harzianum T39. *Phytopathology,* 104**,** 150-7.
HARMAN, G. E., HOWELL, C. R., VITERBO, A., CHET, I. & LORITO, M. 2004. Trichoderma species--opportunistic, avirulent plant symbionts. *Nat Rev Microbiol,* 2**,** 43-56.
JIANG, X., GENG, A., HE, N. & LI, Q. 2011. New isolate of Trichoderma viride strain for enhanced cellulolytic enzyme complex production. *J Biosci Bioeng,* 111**,** 121-7.
KANDLER, O. 1994. Cell-Wall Biochemistry and 3-Domain Concept of Life. *Systematic and Applied Microbiology,* 16**,** 501-509.
KARLSSON, M., ATANASOVA, L., JENSEN, D. F. & ZEILINGER, S. 2017. Necrotrophic Mycoparasites and Their Genomes. *Microbiol Spectr,* 5.
KEELING, P. J. & PALMER, J. D. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics,* 9**,** 605-618.
KING, A. J., CRAGG, S. M., LI, Y., DYMOND, J., GUILLE, M. J., BOWLES, D. J., BRUCE, N. C., GRAHAM, I. A. & MCQUEEN-MASON, S. J. 2010. Molecular insight into lignocellulose digestion by a marine isopod in the absence of gut microbes. *Proceedings of the National Academy of Sciences,* 107**,** 5345-5350.
KNOTT, B. C., CROWLEY, M. F., HIMMEL, M. E., STAHLBERG, J. & BECKHAM, G. T. 2014. Carbohydrate-protein interactions that drive processive polysaccharide translocation in enzymes revealed from a computational study of cellobiohydrolase processivity. *J Am Chem Soc,* 136**,** 8810-9.
KNOTT, B. C., HADDAD MOMENI, M., CROWLEY, M. F., MACKENZIE, L. F., GÖTZ, A. W., SANDGREN, M., WITHERS, S. G., STÅHLBERG, J. & BECKHAM, G. T. 2013. The mechanism of cellulose hydrolysis by a two-step, retaining cellobiohydrolase elucidated by structural and transition path sampling studies. *Journal of the American Chemical Society,* 136**,** 321-329.
KOSHLAND, D. E. 1953. Stereochemistry and the Mechanism of Enzymatic Reactions. *Biological Reviews of the Cambridge Philosophical Society,* 28**,** 416-436.
KOVACS, K., SZAKACS, G. & ZACCHI, G. 2009. Comparative enzymatic hydrolysis of pretreated spruce by supernatants, whole fermentation

broths and washed mycelia of Trichoderma reesei and Trichoderma atroviride. *Bioresour Technol,* 100**,** 1350-7.
KUMAR, S., STECHER, G. & TAMURA, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution,* accepted.
KUNII, M., YASUNO, M., SHINDO, Y. & KAWATA, T. 2013. A *Dictyostelium* cellobiohydrolase orthologue that affects developmental timing. *Dev Genes Evol,* 224**,** 25-35.
KURASIN, M. & VALJAMAE, P. 2011. Processivity of cellobiohydrolases is limited by the substrate. *J Biol Chem,* 286**,** 169-77.
LANGAN, P., NISHIYAMA, Y. & CHANZY, H. 2001. X-ray structure of mercerized cellulose II at 1 a resolution. *Biomacromolecules,* 2**,** 410-6.
LEE, T. S. 2008. Reverse conservation analysis reveals the specificity determining residues of cytochrome P450 family 2 (CYP 2). *Evol Bioinform Online,* 4**,** 7-16.
LEWIS, L. A. & MCCOURT, R. M. 2004. Green algae and the origin of land plants. *American Journal of Botany,* 91**,** 1535-1556.
LINDAHL, B. D., IHRMARK, K., BOBERG, J., TRUMBORE, S. E., HOGBERG, P., STENLID, J. & FINLAY, R. D. 2007. Spatial separation of litter decomposition and mycorrhizal nitrogen uptake in a boreal forest. *New Phytol,* 173**,** 611-20.
LINGER, J. G., TAYLOR, L. E., 2ND, BAKER, J. O., VANDER WALL, T., HOBDEY, S. E., PODKAMINER, K., HIMMEL, M. E. & DECKER, S. R. 2015. A constitutive expression system for glycosyl hydrolase family 7 cellobiohydrolases in Hypocrea jecorina. *Biotechnol Biofuels,* 8**,** 45.
LIU, M., SUN, Z. X., ZHU, J., XU, T., HARMAN, G. E. & LORITO, M. 2004. Enhancing rice resistance to fungal pathogens by transformation with cell wall degrading enzyme genes from Trichoderma atroviride. *J Zhejiang Univ Sci,* 5**,** 133-6.
LOMBARD, V., RAMULU, H. G., DRULA, E., COUTINHO, P. M. & HENRISSAT, B. 2014. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research,* 42**,** D490-D495.
MCMULLIN, D. R., RENAUD, J. B., BARASUBIYE, T., SUMARAH, M. W. & MILLER, J. D. 2017. Metabolites of Trichoderma species isolated from damp building materials. *Can J Microbiol*.
MOMENI, M. H., PAYNE, C. M., HANSSON, H., MIKKELSEN, N. E., SVEDBERG, J., ENGSTROM, A., SANDGREN, M., BECKHAM, G. T. & STAHLBERG, J. 2013. Structural, biochemical, and computational characterization of the glycoside hydrolase family 7 cellobiohydrolase of the tree-killing fungus *Heterobasidion irregulare*. *J Biol Chem,* 288**,** 5861-72.
NIKLAS, K. J. 2004. The cell walls that bind the tree of life. *Bioscience,* 54**,** 831-841.
NISHIYAMA, Y., LANGAN, P. & CHANZY, H. 2002. Crystal structure and hydrogen-bonding system in cellulose Ibeta from synchrotron X-ray and neutron fiber diffraction. *J Am Chem Soc,* 124**,** 9074-82.

NISHIYAMA, Y., LANGAN, P., WADA, M. & FORSYTH, V. T. 2010. Looking at hydrogen bonds in cellulose. *Acta Crystallographica Section D-Biological Crystallography,* 66**,** 1172-1177.
NISHIYAMA, Y., SUGIYAMA, J., CHANZY, H. & LANGAN, P. 2003. Crystal structure and hydrogen bonding system in cellulose I(alpha) from synchrotron X-ray and neutron fiber diffraction. *J Am Chem Soc,* 125**,** 14300-6.
PARFREY, L. W., GRANT, J., TEKLE, Y. I., LASEK-NESSELQUIST, E., MORRISON, H. G., SOGIN, M. L., PATTERSON, D. J. & KATZ, L. A. 2010. Broadly Sampled Multigene Analyses Yield a Well-Resolved Eukaryotic Tree of Life. *Systematic Biology,* 59**,** 518-533.
PAYNE, C. M., KNOTT, B. C., MAYES, H. B., HANSSON, H., HIMMEL, M. E., SANDGREN, M., STAHLBERG, J. & BECKHAM, G. T. 2015. Fungal cellulases. *Chem Rev,* 115**,** 1308-448.
POPPER, Z. A. & FRY, S. C. 2003. Primary cell wall composition of bryophytes and charophytes. *Annals of Botany,* 91**,** 1-12.
PRAESTGAARD, E., ELMERDAHL, J., MURPHY, L., NYMAND, S., MCFARLAND, K. C., BORCH, K. & WESTH, P. 2011. A kinetic model for the burst phase of processive cellulases. *FEBS J,* 278**,** 1547-60.
REESE, E. T. 1956. A microbiological process report; enzymatic hydrolysis of cellulose. *Appl Microbiol,* 4**,** 39-45.
RODIONOVA, N. A., MARTINOVICH, L.I., GUKASYAN, G.S. AMELINA, D.S. 1988. Purification of cellulases from Geotrichum candidum and Trichoderma longibrachiatum by chromatography on microcrystalline cellulose. *Prikl. Biokhim. Mikrobiol.,* 24**,** 370-379.
RZHETSKY, A. & NEI, M. 1992. A simple method for estimating and testing minimum-evolution trees. *Mol. Biol. Evol,* 9**,** 945-967.
SAMMOND, D. W., PAYNE, C. M., BRUNECKY, R., HIMMEL, M. E., CROWLEY, M. F. & BECKHAM, G. T. 2012. Cellulase linkers are optimized based on domain type and function: insights from sequence analysis, biophysical measurements, and molecular simulation. *PLoS One,* 7**,** e48615.
SARKAR, P., BOSNEAGA, E. & AUER, M. 2009. Plant cell walls throughout evolution: towards a molecular understanding of their design principles. *J Exp Bot,* 60**,** 3615-35.
SCHMOLL, M., DATTENBOCK, C., CARRERAS-VILLASENOR, N., MENDOZA-MENDOZA, A., TISCH, D., ALEMAN, M. I., BAKER, S. E., BROWN, C., CERVANTES-BADILLO, M. G., CETZ-CHEL, J., CRISTOBAL-MONDRAGON, G. R., DELAYE, L., ESQUIVEL-NARANJO, E. U., FRISCHMANN, A., GALLARDO-NEGRETE JDE, J., GARCIA-ESQUIVEL, M., GOMEZ-RODRIGUEZ, E. Y., GREENWOOD, D. R., HERNANDEZ-ONATE, M., KRUSZEWSKA, J. S., LAWRY, R., MORA-MONTES, H. M., MUNOZ-CENTENO, T., NIETO-JACOBO, M. F., NOGUEIRA LOPEZ, G., OLMEDO-MONFIL, V., OSORIO-CONCEPCION, M., PILSYK, S., POMRANING, K. R., RODRIGUEZ-IGLESIAS, A., ROSALES-SAAVEDRA, M. T., SANCHEZ-ARREGUIN, J. A., SEIDL-SEIBOTH, V., STEWART, A.,

URESTI-RIVERA, E. E., WANG, C. L., WANG, T. F., ZEILINGER, S., CASAS-FLORES, S. & HERRERA-ESTRELLA, A. 2016. The Genomes of Three Uneven Siblings: Footprints of the Lifestyles of Three Trichoderma Species. *Microbiol Mol Biol Rev,* 80**,** 205-327.

SOMERVILLE, C., BAUER, S., BRININSTOOL, G., FACETTE, M., HAMANN, T., MILNE, J., OSBORNE, E., PAREDEZ, A., PERSSON, S., RAAB, T., VORWERK, S. & YOUNGS, H. 2004. Toward a systems approach to understanding plant cell walls. *Science,* 306**,** 2206-11.

SORENSEN, T. H., WINDAHL, M. S., MCBRAYER, B., KARI, J., OLSEN, J. P., BORCH, K. & WESTH, P. 2017. Loop variants of the thermophile Rasamsonia emersonii Cel7A with improved activity against cellulose. *Biotechnol Bioeng,* 114**,** 53-62.

STAHLBERG, J., DIVNE, C., KOIVULA, A., PIENS, K., CLAEYSSENS, M., TEERI, T. T. & JONES, T. A. 1996. Activity studies and crystal structures of catalytically deficient mutants of cellobiohydrolase I from Trichoderma reesei. *J Mol Biol,* 264**,** 337-49.

STAHLBERG, J., JONSSON, B. & HORVATH, C. 1991. Theory for electrostatic interaction chromatography of proteins. *Anal Chem,* 63**,** 1867-74.

STALS, I., SANDRA, K., GEYSENS, S., CONTRERAS, R., VAN BEEUMEN, J. & CLAEYSSENS, M. 2004. Factors influencing glycosylation of Trichoderma reesei cellulases. I: Postsecretorial changes of the O- and N-glycosylation pattern of Cel7A. *Glycobiology,* 14**,** 713-24.

SUCGANG, R., KUO, A., TIAN, X., SALERNO, W., PARIKH, A., FEASLEY, C. L., DALIN, E., TU, H., HUANG, E., BARRY, K., LINDQUIST, E., SHAPIRO, H., BRUCE, D., SCHMUTZ, J., SALAMOV, A., FEY, P., GAUDET, P., ANJARD, C., BABU, M. M., BASU, S., BUSHMANOVA, Y., VAN DER WEL, H., KATOH-KURASAWA, M., DINH, C., COUTINHO, P. M., SAITO, T., ELIAS, M., SCHAAP, P., KAY, R. R., HENRISSAT, B., EICHINGER, L., RIVERO, F., PUTNAM, N. H., WEST, C. M., LOOMIS, W. F., CHISHOLM, R. L., SHAULSKY, G., STRASSMANN, J. E., QUELLER, D. C., KUSPA, A. & GRIGORIEV, I. V. 2011. Comparative genomics of the social amoebae Dictyostelium discoideum and Dictyostelium purpureum. *Genome Biol,* 12**,** R20.

SUKHARNIKOV, L. O., CANTWELL, B. J., PODAR, M. & ZHULIN, I. B. 2011. Cellulases: ambiguous nonhomologous enzymes in a genomic perspective. *Trends Biotechnol,* 29**,** 473-9.

SUOMINEN, P. L., MANTYLA, A. L., KARHUNEN, T., HAKOLA, S. & NEVALAINEN, H. 1993. High frequency one-step gene replacement in Trichoderma reesei. II. Effects of deletions of individual cellulase genes. *Mol Gen Genet,* 241**,** 523-30.

TEXTOR, L. C., COLUSSI, F., SILVEIRA, R. L., SERPA, V., DE MELLO, B. L., MUNIZ, J. R., SQUINA, F. M., PEREIRA, N., JR., SKAF, M. S. & POLIKARPOV, I. 2013. Joint X-ray crystallographic and molecular dynamics study of cellobiohydrolase I from Trichoderma harzianum: deciphering the structural features of cellobiohydrolase catalytic activity. *FEBS J,* 280**,** 56-69.

VALJAMAE, P., SILD, V., NUTT, A., PETTERSSON, G. & JOHANSSON, G. 1999. Acid hydrolysis of bacterial cellulose reveals different modes of synergistic action between cellobiohydrolase I and endoglucanase I. *Eur J Biochem,* 266**,** 327-34.

VAN WYK, J. P. & MOHULATSI, M. 2003. Biodegradation of wastepaper by cellulase from Trichoderma viride. *Bioresour Technol,* 86**,** 21-3.

VON OSSOWSKI, I., STÅHLBERG, J., KOIVULA, A., PIENS, K., BECKER, D., BOER, H., HARLE, R., HARRIS, M., DIVNE, C., MAHDI, S., ZHAO, Y., DRIGUEZ, H., CLAEYSSENS, M., SINNOTT, M. L. & TEERI, T. T. 2003. Engineering the Exo-loop of *Trichoderma reesei* Cellobiohydrolase, Cel7A. A comparison with Phanerochaete chrysosporium Cel7D. *Journal of Molecular Biology,* 333**,** 817-829.

WALSETH, C. S. 1952. Occurrence of cellulases in enzyme preparations from microorganisms. *TAPPI J,* 35**,** 228-233.

WANG, Y., SLADE, M. B., GOOLEY, A. A., ATWELL, B. J. & WILLIAMS, K. L. 2001. Cellulose-binding modules from extracellular matrix proteins of Dictyostelium discoideum stalk and sheath. *Eur J Biochem,* 268**,** 4334-45.

WOLFENDEN, R. & SNIDER, M. J. 2001. The depth of chemical time and the power of enzymes as catalysts. *Accounts of Chemical Research,* 34**,** 938-945.

ZHANG, P., MCGLYNN, A. C., LOOMIS, W. F., BLANTON, R. L. & WEST, C. M. 2001. Spore coat formation and timely sporulation depend on cellulose in Dictyostelium. *Differentiation,* 67**,** 72-9.

Popular science summary

This thesis work is devoted to expansive studies of GH7 cellobiohydrolases, key enzymes of biomass degradation. It is a remarkable group of enzymes, which are specifically designed by Nature for processive hydrolysis of cellulose chains. We find cellulose everywhere, it is the most abundant polymer in the world. By peeling off impressively long cellulose chains from the surface of insoluble cellulose and cut them into small, soluble cellobiose units, cellobiohydrolases play a key role in the carbon cycle on Earth. The same principle of biomass degradation is used in bioreactors. To be able to apply these enzymes efficiently for sustainable development we have to get a clever insight in their structure, the unique molecular architecture which enables them to be so efficient degraders. A number of 3D structures are presented in the current work, showing interesting features of GH7 cellobiohydrolases. Computational simulations were done to examine the dynamics of these structures and their behaviour in enzymatic cellulose degradation. Real tests of enzymatic activities, as well as thermostability and pH stability properties of the enzymes, were also carried out. Together the data allowed to summarize important information about structural features of these enzymes, responsible for certain functions. For a long time it was thought that only fungi can produce GH7 cellobiohydrolases, but recently these enzymes were found in genomes of distantly related organisms, such as social amoebae, for example. Structures of GH7 cellobiohydrolases presented in this study showed that they are very similar to fungal GH7 CBH. It means that during evolution, these enzymes have not changed much, while the whole world has changed significantly. Why is that? Probably because cellulose was always there. In this sense GH7 cellobiohydrolase is like a bicycle, once invented, no need to change radically. During the evolution they maintained fantastic ability to do their main job – cellulose degradation in a processive manner. Nevertheless smart tuning is very useful and therefore further studies are absolutely required.

Acknowledgements

Here I want to acknowledge all the people who helped me on my long way to defence. Without you this couldn't happen.

First of all I would like to thank my principal supervisor **Jerry Ståhlberg**, for his easy and intelligent supervision, for all the deep knowledge I gained from him and for his optimistic attitude on many things in science and life. I want to thank Jerry for his enormous help with all the aspects of my research project, and teaching me, at last, how to make my thoughts more "scientific". I also want to thank Jerry and Kerstin for amazing skating tours in winter and for lovely choir concerts, good company and great fun!

I would like to thank **Mats Sandgren**, who actually brought me to Sweden, for his intuitive management of the projects, providing me great experience and help with expanding my network. I want to thank **Henrik Hansson** for extremely useful advices on biochemical and structural studies, and fruitful discussions. I want to thank **Saeid Karkehabadi** inspiring me as a student, sharing his experience in crystallography and crystallization room, as well. Besides science I'm very thankful to Saeid, Faranak, Vida and Behrad for being my "second family" in Sweden, even though I believe it was not easy. I want to thank **Nils Mikkelsen**, who was always there with IT support, even when there was nothing to support, after a double crash of my hard-drive during my PhD.

I want to thank former PhD students, who are now postdocs in our group and elsewhere, **Mikael Gudmundsson** for his brilliant ideas for solving structural problems, and trying to make me better software user (still quite hopeless); **Miao Wu** for being an example of incredible efficiency in work and private life; **Saumendra Roy** for his support and friendship, and giving me insights into Bangladesh culture; **Majid Momeni** for his cheerful talks and inspiration. I want to thank former postdoc **Maria Dimarogona** for her help with crystallography and all these data-collection trips and conferences, work was real fun, despite all the failures (or maybe because of them), when you

were around! I want to thank **Bing Liu** for being my mate in protein purifications, and notes of music he occasionally brought into my PhD life. I want to thank **Riin Kont** for fruitful cellulase discussions and great time in Estonia; **Sumitha Reddy** for her help, collaboration and great sense of humor. I want to thank current PhD students **Benjamin**, **Jule**, **Mahfuz, Topi** for being excellent co-wokers and office mates, now it is time to say that**.** Good luck with your Thesis! Also I want to thank my russian supervisor **Anna Kulminskaya**, and other colleagues, **Lena, Kirill, Sveta, Konstantin.** I would like to thank **Farid Ibatullin** for ligand synthesis and useful discussions, and **Alexander Golubev** for advices on crystallography.

I want to thank my dear friends, here and there, for being supportive and understanding, and motivating me one way or another to go through my PhD studies. **Nastya, Yulia, Ksenia, Alyona, Lyola, Alia, Marina, Anya, Ragnar, Bora, Patrick** you somehow made my way more bright and happy. I'm thankful to **Pasha** for IT support for the phylogenetic analysis and being such a great friend for all these years; and to **Nikita** for encouraging me to become what I am now.

At last, I would like to thank my dear parents, **Olga** and **Sergei,** for their love, support and understanding, for their sense of humour, which always makes me feel better about any situation; and for they always let me do whatever I want to do, and believe everything is possible. Дорогие мама и папа, большое вам спасибо! Вы мои лучшие друзья, и я вас очень люблю.