Ancestral sequence reconstruction for protein engineering: improving celulases for biomass hydrolysis

Nerea Barruetabeña Garate

- PhD Thesis -

Thesis supervisor
Prof. Raul Perez Jimenez

2017
This PhD thesis has been conducted in CIC NanoGUNE

by

Nerea Barruetabeña Garate
Eziña, ekiñe

egiña.
Table of contents

Laburpena ................................................................................................. 7
Summary .................................................................................................... 13
Chapter 1: Introduction ............................................................................ 15

Chapter 2: Methods for phylogenetic analysis ..................................... 36
2.1. Introduction .......................................................................................... 36
2.2. Theory .................................................................................................. 38
  2.2.1. Methods ............................................................................................ 39
    2.2.1.1. Parsimony .................................................................................... 40
    2.2.1.2 Maximum likelihood ....................................................................... 42
    2.2.1.3. Bayesian inference ........................................................................ 44
2.3. Methodology ......................................................................................... 45
  2.3.1. Selection of extant sequences: Uniprot ........................................... 46
  2.3.2. Creation of a multiple alignment: MUSCLE .................................. 50
  2.3.3. Computing a phylogenetic tree ....................................................... 54
    2.3.3.1. Beauti ......................................................................................... 54
# Table of Contents

3.1.7.4. Extant *T.maritima*..........................75
3.1.8. *T.reesei* cocktail protein determination........75

**3.2. Biochemical assays**.................................76
3.2.1. CMC..................................................76
3.2.1.1. Residual and long-term activity measurements..78
3.2.1.2. Inactivation constant (Kin) determination........78
3.2.2. CellG3.................................................79
3.2.2. Filter-paper............................................80
3.2.2.1. Lignocellulosic substrate hydrolysis.............82
3.2.3. Avicel................................................82
3.2.4. Thermal stability of the ancestral endoglucanase: Circular Dichroism..83
3.2.5. Ancestral endoglucanases Kinetic parameters determination........83

**3.3. Cellulosome**.........................................84
3.3.1. Minicellulosome.....................................84
3.3.1.1. Minicellulosome construct.........................84
3.3.1.2. Minicellulosome activity assays....................86
3.3.2. Cellulosome.........................................86
3.3.2.1. Cellulosome design................................86
3.3.2.2. Cellulosome construction..........................87
3.3.2.2.1. Cloning of the new designs in the expression plasmid by PCR.....87
3.3.2.2.2. Recombinant protein expression..................88
3.3.2.2.3. Gel electrophoresis................................88
3.3.2.2.4. Affinity based ELISA.............................89
Chapter 4: Phylogenetic results .................................................. 90

4.1. Reconstruction of an ancestral bacterial endoglucanase ............. 91
  4.1.1. Ancestral endoglucanase sequence analysis .......................... 95

4.2.1 Reconstruction of an ancestral bacterial exoglucanase ............. 96
  4.2.2. Ancestral exoglucanase sequence analysis .......................... 98

4.3.1. Reconstruction of an ancestral bacterial beta-glucosidase ....... 98
  4.3.2. Ancestral beta-glucosidase sequence analysis ....................... 101

Chapter 5: Experimental results ................................................. 102

5.1. Ancestral cellulases production ........................................... 103

5.2. Mass spectroscopy and protein concentration determination ......... 106

5.3. Endoglucanase assays ..................................................... 107
  5.3.1. Lignocellulosic substrates hydrolysis .............................. 117

5.3.2. Thermal stability of the ancestral endoglucanase: Circular Dichroism 119

5.3.3. Ancestral endoglucanases Kinetic parameters determination .... 121

5.4. Enzyme Cocktail assays .................................................. 123
  5.4.1. Ancestral enzyme cocktail .......................................... 123

5.4.2. Lignocellulosic substrates hydrolysis ................................ 130

5.5. Minicellulosome .......................................................... 133
  5.5.1. Minicellulosome construction .................................... 134

5.5.2. Minicellulosome activity assays .................................... 135

5.6. Cellulosome .............................................................. 139
## Table of Contents

5.6.1. Recombinant protein expression .............................................. 141
5.6.2. Cellulosome construction ....................................................... 141

**Chapter 6: Discussion** ............................................................... 146

**Bibliography** ............................................................................. 152

**Appendix I** ................................................................................ 169

**Appendix II** ................................................................................ 178

**Appendix III** .............................................................................. 193

**Appendix IV** ............................................................................... 213

**Acknowledgment** .................................................................... 216
Laburpena

Laburpena


1 irudia. **Endozelulazarentzat egindako azterketa. a)** Endozelulasaren kronograma edo arbola filogenetikoa. Bakterioen hiru erreinutako (aktinobakteria, firmikutes eta proteobakteria) 32 sekuentzia erabili egin da zuhaitza. Bertako arbaso bat hartuta egin da berpizketa, **b)** Temperaturaren menpe (30-90) pH 4.8-an egindako aktibitate espezifikoaren azterketa c) pH-rekiko (4-10) menpekoitasun diagrama 50ºC-tan eginikoa.

Irudian ikus daitekeenez, baldintza guztietan emaitza hobeak lortu dira antzinako entzima erabili eta. Horretaz gain, antzinako koktelaren jardueraren azterketak egin dira (hiru zelulasa: endozelulasa, exozelulasa eta beta-glukosidasa), hainbat temperaturatan substratu lignozelulosikoen degradazioa egin dut.

Substratu idealak (CMC eta iragaz papera) erabiltzeaz gain, material lignozelulosiko desberdinekin azterketak egin ditut, hala nola, kartoia, batzeko papera eta egunkari-papera. Jarraian dagoen irudian (2.irudia) ikus daiteke.
2 irudia. Zelulas koktelarentzat eginiko azterketa Koktel zaharreko jarduerak (CKA), koktel komertziala (CKC), koktel zaharrean T. pubescens laccase (CKA + L) eta komertzial koktelaren presentzia T. pubescens laccase (CKC + L) ikusgai 50-70 ºC-ko tenperaturan eta pH 4.8. Azterketak hiru substratu desberdinetan burutu ziren: kartoia (a), egunkaria (b) eta paperean (c).

Azterketa guztiak antzinako zelulazen errendimendu hobea neurtu da, bai temperaturnu altuetan, bai eta pH ezberdinetan ere. Biokonbertsio prozesuen (pH azidoa eta temperaturnu altua) muturreko baldintzei aurre egitako gaitasun hobea dutenaren
hipotesia frogatu da, beraz. Horrela, prozesuaren kostua murritztu daiteke eta bere bideragarritasun industriala hobetuko litzateke. Bain a ez hori bakarrik, gure antzinako koktelak material lignozelulosikoa degradatzeko erabiltzen diren beste entzima batzuekin daukan sinergia ere erakutsi dugu. Izan ere, material lignozelulosikoa zelulosaz gain, lignina eta hemizelulosaz osatua dago. Horiek hidrolizatzeko asmoz, lakasak eta xilan asak erabiltzen dira. Guk, azterketa honen bitartez, frogatu dugu gure entzimek beraien kabuz lan egiteaz gain, beste entzima horiekin lan egitean errendimendua asko hobetzen dela.


Summary

The use of enzymes in industrial applications requires an improvement thereof with various techniques of protein engineering. This is necessary since in many industrial conditions there are extreme conditions that inactivate enzymes. In many cases the use of enzymes from extremophiles produce satisfactory results, but there are situations in which even these extremophile enzymes should be improved. The method most used by biotech companies is directed evolution in vitro. However, this protocol is very expensive since the generation of mutant library and subsequent screening has to be done experimentally and often, thousands of variants are created. In addition, due to the inability to test all variants generated, this process does not guarantee that the best enzymes are selected. So there is a need of another technique for improving enzymes.

For this purpose the novel technique of the resurrection of ancient proteins is used. In this technique, sequences of enzymes from modern organisms are used and their phylogenetic relationship is studied. Once a phylogenetic tree is obtained, statistical techniques can be applied to obtain the sequence of ancestors and those can be studied in the laboratory. Nowadays, there are relatively few studies of resurrection of ancient proteins and possible industrial and biomedical applications. Ancestral proteins and enzymes often show exceptional properties related to its thermal stability, chemical or kinetic, in addition to its activity. This is due to the fact that our planet has been subjected to all
kinds of environmental conditions throughout history. Just the protein sequences of organisms that lived in those conditions have to be found. High activities are expected since it has been shown that ancestral enzymes are promiscuous, they are able to work with lower selectivity and therefore more effective with different types of substrates.

Cellulases enzymes are required to work with high efficiency under industrial conditions that limit their capabilities. Cellulase is the name of a group of enzymes responsible for the hydrolyization of cellulose. We have brought back to life ancestral cellulases up to 3000 million years. In all the assays we have measured a better performance of the ancestral cellulases, both in high temperatures and in a range of pH. Cellulases are quite limited in their properties being difficult to use at temperatures above 60 °C and extreme pH, improving these properties is a goal in biotechnology research.

Apart from soluble cellulases, it has been suggested that bacterial cellulases forming the macromolecular cellulosome complex may be of potential interest for industry due to the increased stability and cellulase activity of the complex. Efforts to develop cellulosomes for industrial applications have been focused on so-called Designer Cellulosomes, where individual components of different cellulosomes are combined to produce defined cellulosome-like complexes. Another intention is to redesign and create a new cellulosome composed of ancestral cellulases (endoglucanase, exoglucanase and beta-glucosidase) to function in the relevant industrial environment of high temperature, mechanical stress and acidic pH.
Chapter 1: Introduction

Enzymes are widely used in the chemical and biotechnological industry, being essential biocatalysts in diverse areas such as bioenergy, cosmetics, food industry, detergents and textile [1]. Natural enzymes are suited for their biological function, but when these enzymes are used as industrial catalysts they present significant limitations for the specific requirements for industrial application. In the past decades, research has been focused on the improvement of enzymes properties, paying special attention to the enhancement of the thermal stability, the increase of the specific activity, the improvement of their substrate promiscuity and the increase of the production rate [2-8]. In some cases the use of enzymes from extremophiles produce satisfactory results, but there are situations in which even these extremophile enzymes should be improved. Such non-natural conditions often result in poor enzyme activity, or complete deactivation due to denaturation or chemical modifications. Developments in protein engineering over the past ten years have enabled enzymes to be evolved in vitro for properties that favor the required process conditions, and also to obtain enzyme variants with altered substrate specificity or enantioselectivity [9, 10].
In order to obtain enhanced enzymes with improved efficiency several methods such as directed mutagenesis, DNA shuffling, error-prone PCR, directed evolution and rational design have been implemented [3, 5, 8, 11, 12]. The most classical method in enzyme engineering is rational design, which involves site directed mutagenesis introducing a specific amino acid into a target gene [13, 14]. The amino acid is chosen considering the structure and function of the enzyme. Its major drawback is that detailed structural knowledge of an enzyme is often unavailable. In addition, rational design does not allow the introduction of numerous mutations without perturbing the structure or function of the protein. DNA shuffling consists on the fragmentation of gene parents, followed by some PCR (polymerase chain reaction) cycles to obtain different mutants of a gene. Regarding error-prone PCR, this is a method by which random mutants maybe inserted into any piece of DNA. The technique is based on the well-founded PCR [15]. Nevertheless, the method that the biotechnological companies use the most is directed evolution in vitro [16, 17]. This is a method used in protein engineering that mimics the process of natural selection to evolve proteins or nucleic acids toward a user-defined condition. Using this method, thousands of variants of the enzyme of interest are created introducing random mutations, generating a library of mutants. These mutant enzymes are exposed to the desired conditions and the variants that best perform under these conditions are identified and selected for commercial exploitation. The likelihood of success in a directed evolution experiment is directly related to the total library size, as evaluating more mutants increases the chances of finding one with the desired properties [18]. However, this protocol is very expensive since the generation of a mutant library and subsequent screening has to be done experimentally and often, thousands of variants are created.
In addition, due to the inability to test all variants generated, this process does not guarantee that the best enzymes are selected [19].

Despite these advances, the limitation of these engineered enzymes is still a serious barrier in the chemical industry. Nowadays no methodology seems to be able to enhance, for instance, the temperature and pH operability, the expression level or the specific activity of enzymes, all at once. Developing a strategy capable of vastly improving the catalytic properties of enzymes in a cost efficient manner may revolutionize the biotechnology and chemical industries. In the past years, the so-called Ancestral Sequence Reconstruction technique (ASR) has been used to study the evolution of genes, proteins and enzymes [20-23]. Surprisingly, reconstructed ancestral enzymes displayed enhanced thermal stability, better pH response, improved activity and expression level, chemical promiscuity and in some cases, all of this at once [21, 22, 24-28] [29]. This technique is based on the evolution theory, which states that groups of organisms change over time so that descendants differ structurally and functionally from their ancestor. Today, combined with biophysical and biochemical state-of-the-art techniques, ancestral sequence reconstruction allows to study and compare features of extinct proteins and genes that are otherwise inaccessible [20-22, 30].

Phylogenetic methods applied to genomic information have made it possible to establish evolutionary relationships among different living organisms, including the possibility of inferring the putative sequences of the genes of their already extinct ancestors [31, 32]. Since Charles Robert Darwin (Fig 1.1) sketched an evolutionary tree in 1837 for the first time, to the current Time Tree of Life for all known living organisms a lot has been learnt in organismal and molecular evolution.
The two primary components of evolutionary history are the relationship of organisms (phylogeny) and their times of divergence which together form a phylogenetic tree scaled to time, i.e., a chronogram. Therefore, the two main goals of phylogenetic analysis are to reconstruct the correct genealogical relationships between organisms and estimate the time of divergence between organisms since they shared a common ancestor.

Ancestral Sequence Reconstruction allows determining the sequences of genes and proteins of the ancestor of modern species. The process comprises various steps (Fig 1.2). Protein or gene sequences of different species are acquired from the available databases. These sequences are then processed by bioinformatics tools to infer the phylogenetic tree. The genetic
code of the ancestor is obtained from the inferred tree and the protein can be resurrected in the laboratory using molecular cloning. This methodology is deeply explained in Chapter 3.

**Figure 1.2. Schematic explanation of ancestral reconstruction and resurrection process.** First, the sequences are retrieved from online databases, and then the tree is inferred by bioinformatics tools. After that, the selected sequence is cloned in the lab and brought back to life by molecular biology techniques.

The first reconstruction of an ancestral protein was carried in 1994 by Shindyalov et al [33]. They predicted the tertiary structure of….Since then numerous studies have been carried out using this technique, providing information not only related to physiological and metabolic features [34], but also information about the environmental conditions that hosted ancestral organisms [19, 20]. Precisely, the deduction of the environmental conditions of different geological eras is another important application of this technique. Several studies carried out with enzymes have reported the conditions of the earth at the
Precambrian era [19, 20, [35-37]. On od the first studies reporting environmental information was carried out by Gaucher and collaborators [21]. They reconstructed the translation elongation factors of species that lived in the range of 3.5-0.5 Gyr ago and calculated their melting temperature. According to this study, the temperature on earth cooled down over 30 ºC during that period, matching previous work related to the temperature of ancient oceans calculated from silicon isotopes [38]. Some years later, some new studies, such as, the one developed by Perez-Jimenez et al. studied the thermochemical evolution of thioredoxins from 4 to 1.4 Gyr with single-molecule force spectroscopy [22] confirming this cooling trend. The obtained results showed that ancestral enzymes have higher thermal stability than the extant ones.

Ancestral proteins and enzymes often show exceptional properties related to its thermal stability, chemical or kinetic, in addition to its activity. This is due to the fact that our planet has been subjected to extreme environmental conditions throughout history. Organisms and their molecules were adapted to these conditions. In addition, it has been shown that ancestral enzymes are promiscuous [24, 27], they are able to work with lower selectivity and therefore more effective with different types of substrates.

Nowadays, there are relatively few studies of resurrection of ancient enzymes with possible industrial and biomedical applications. These include most notably thioredoxin, or the lactamases [24]. These studies have shown that the study of ancestral enzymes not only has a great value from an evolutionary point of view, but can also have multiple applications in areas such as bioengineering and biomedicine.
Following the mentioned assumptions, we are going to explore the possible applications of ancestral enzymes for biotechnological and industrial applications. In this sense, we have decided to reconstruction ancestral variants of a group of enzymes widely used in industry, i.e., cellulases. These enzymes are among the most utilized enzymes in chemical industry [39] because of their ability to completely hydrolyze cellulose into glucose monomers (Fig 1.3).

![Cellulase reaction diagram](image)

**Figure 1.3. Hydrolysis of cellulose chain into glucose by cellulase enzymes.**

Cellulases are quite limited in their properties being difficult to use at temperatures above 60 °C and extreme pH, improving these properties is a goal in biotechnology research. Increasing
the thermal operability and activity of cellulases is perhaps the most investigated aspect for modifying them for industrial implementation [3, 11, 40-42] as well as other lignocellulosic enzymes. As I have explained before, several methods, such as directed mutagenesis, DNA shuffling, error-prone PCR, and directed evolution, have been implemented to obtain enhanced cellulases [3,[12, 40] that considerably improve the efficiency of cellulose bioconversion. Despite these improvements, the limitations of the enzymes is still a serious barrier in their industrial applications and we hypothesize that this methodology will help to enhance the temperature and pH resistance, the expression level or the specific activity of cellulases, hopefully all at once.

Cellulose is one of the major components in plant cell walls and is the most abundant organic polymer on the planet [43]. It is a homo-polysaccharide composed entirely of D-glucose monomers linked together by β-1,4-glucosidic bonds. There is an enormous variety of raw materials rich in cellulose, such as agricultural, industrial, and urban waste that can be used as sources for fermentable sugars [44].

Lignocellulose can be obtained from a variety of sources, such as agricultural residues, forestry residues, energy crops and bio-waste streams (Table 1.1).
<table>
<thead>
<tr>
<th>Lignocellulosic materials</th>
<th>Cellulose (%)</th>
<th>Hemicellulose (%)</th>
<th>Lignin (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardwood stems</td>
<td>40-50</td>
<td>24-40</td>
<td>18-25</td>
</tr>
<tr>
<td>Softwood stems</td>
<td>45-50</td>
<td>25-35</td>
<td>25-35</td>
</tr>
<tr>
<td>Paper</td>
<td>85-99</td>
<td>0</td>
<td>0-15</td>
</tr>
<tr>
<td>Newspaper</td>
<td>40-55</td>
<td>25-40</td>
<td>18-30</td>
</tr>
<tr>
<td>Leaves</td>
<td>15-20</td>
<td>80-85</td>
<td>0</td>
</tr>
<tr>
<td>Switch grass</td>
<td>45</td>
<td>31.4</td>
<td>12</td>
</tr>
<tr>
<td>Corn cobs</td>
<td>45</td>
<td>35</td>
<td>15</td>
</tr>
<tr>
<td>Waste paper from chemical pulps</td>
<td>60-70</td>
<td>10-20</td>
<td>5-10</td>
</tr>
<tr>
<td>Primary wastewater solids</td>
<td>8-15</td>
<td>NA</td>
<td>24-29</td>
</tr>
</tbody>
</table>

Table 1.1. Lignocellulose contents of common agricultural residues and wastes [45].

In lignocellulosic biomass, lignin and hemicellulose form an amorphous matrix in which, crystalline cellulose fibrils are embedded (Fig 1.4). The combination of hemicellulose and lignin provides a protective sheath around cellulose and the crystalline structure of cellulose makes it highly resistant to attack. This arrangement makes lignocellulosic biomass recalcitrant and a complex substrate to convert into valuable products.
Figure 1.4. Schematic representation of the structure of lignocellulosic biomass. The picture shows that lignocellulosic biomass (on top in the left) if composed many of cellulose. This cellulose is embedded in lignin and hemicellulose (top in the right). Above, the structure of cellulose is drawn. Glucose monomers form long polymer chains which form the microfibrils.

Cellulose chains, with a degree of polymerization between 10,000 and 15,000, are linked by strong hydrogen bonds which form cellulose chains into microfibrils, making it crystalline in nature.
and very recalcitrant to degradation (Fig 1.4). Three types of cellulases are involved in the complete hydrolysis of cellulose into sugar, and all of them cleave β-1,4-glycosidic bonds (Fig 1.5): endo-β-glucanases (EG) randomly break cellulose fibers; exo-β-glucanases (or cellobiohydrolases, CBH) cleave cellulose chains and release cellobiose; and cellobiose is further hydrolyzed to glucose by β-glucosidase.

Figure 1.5. Synergistic action of cellulases in cellulose degradation. Endo-1,4-β-glucanases break down randomly the internal β-1,4-glucosidic bonds of the cellulose chains, whereas exo-1,4-β-glucanases cleave off cellobiose units from the end of the chains. These cellobiose units are broken down by β-glucosidases into glucose monomers.

Most endoglucanases and exoglucanases have a two-domain structure that contains a catalytic domain (CD) and a carbohydrate binding module (CBM) [46-48] which are connected by a peptide.
linker that maintains the separation between the CD and the CBM (Fig 1.6). The CD contains the enzyme active site, responsible for cellulose hydrolysis. The CBM is a contiguous amino acid sequence that anchors the CD onto the surface of cellulose through hydrogen bonding and van der Waals interactions [49, 50]. According to sequence similarities within their CDs and CBMs, cellulases can be grouped into different families [51].

Figure 1.6. Schematic structure of the different domains of a typical endoglucanase. It contains a catalytic domain (CD) and a carbohydrate binding module (CBM)). These two domains are connected by a peptide linker, which is known to maintain the separation between the CD and the CBM.

In order to produce cellulases, fungi and bacteria have been heavily exploited, but the use of fungi has been more important because of their capability to produce high amounts of cellulases
and hemicellulases which are secreted to the medium for easy extraction and purification. In addition, fungal enzymes are often less complex than bacterial cellulases and can be more rapidly cloned and produced via recombination in a rapidly growing bacterial host, for example E.coli. However, for several reasons the use of bacterial cellulases is becoming widely exploited. On the one hand, bacteria often have a higher growth rate than fungi, which allows a higher recombinant production of enzymes. On the other hand, bacterial cellulases are more complex, therefore they are often expressed in multi-enzyme complexes. In this way, their function and synergy increase. Although those two reasons are important, there is a more important one; bacteria inhabit a wide variety of environmental and industrial niches, it is for that reason that the cellulolytic strains produced are extremely resistant to environmental stress. These include strains that are: thermophilic, psychrophilic, alkaliphilic, acidiophilic or halophilic. Thus, besides surviving the harsh conditions found in the bioconversion process, they often produce enzymes that are stable under extreme conditions present in bioconversion process. Those enzymes should be suitable as they may increase rates of enzymatic hydrolysis, fermentation and product recovery [12, 52]. Thus, using cellulases from bacteria has a greater potential for improvement than those from fungi. Fungal cellulases are produced in large quantities, which is the basis for their use in industrial processes so far. However, some bacterial cellulases present much higher specific activities and therefore have the potential to be excellent components for a more efficient saccharification processes [52].

Lignocellulosic substrates present a high degree of chemical and structural diversity (comprising mainly cellulose, hemicellulose, and pectin, as well as the non-carbohydrate lignin polymer, all in varying proportions), which makes it extremely recalcitrant such
that only some microorganisms are able to degrade it. But due to the different components that form lignocellulosic biomass, multiple enzymes that cooperate synergistically are used in the hydrolysis step. The core enzymes for cellulose hydrolysis are cellulases and the reaction can be favored by other enzymes that break down the structure of lignin and hemicellulose increasing cellulose accessibility (Table 1.2).

<table>
<thead>
<tr>
<th>Component</th>
<th>Enzyme</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cellulose</td>
<td>Endo-1,4-β-glucanase, exo-1,4-β-glucanase, β-glucosidase</td>
</tr>
<tr>
<td>Hemicellulose</td>
<td>Endo-xylanase, β-xylosidase, acetyl xylan esterase, endo-mannanase, β-mannosidase, α-glucuronidase, ferulic acid esterase, α-galactosidase, p-coumaric acid esterase</td>
</tr>
<tr>
<td>Lignin</td>
<td>Laccase, Lignin peroxidase, Manganese peroxidase</td>
</tr>
<tr>
<td>Pectin</td>
<td>Pectin methyl esterase, pectate lyase, polygalacturonase, rhamnogalacturonan lyase</td>
</tr>
</tbody>
</table>

*Table 1.2. Some of the main enzymes required to degrade lignocellulose to monomers*[53]. The column in the left shows the different components of the lignocellulosic material. The column in the right shows the enzymes that are able to hydrolase the components of the lignocellulosic material.

Some parts of the cellulose structure may be amorphous in nature, which are easier to degrade. In regard to hemicelluloses, they consist of short highly branched chains of various sugars: mainly
xylose, and further arabinose, galactose, glucose and mannose [54]. They are classified according to the main sugar in the backbone. Due to the branches, hemicelluloses are amorphous in structure and relatively easier than cellulose to degrade. Concerning lignin, it is a complex three-dimensional network formed by the polymerization of phenyl propane units and forms a protective seal around the other two components. It is the most abundant natural phenolic polymer (Fig 1.7). It is formed mainly by three monomers: p-coumaryl alcohol, coniferyl alcohol and sinapyl alcohol. These monomers are linked together by alkyl-aryl, alkyl-alkyl and aryl-aryl ether bonds (Fig 1.7). It makes the cell wall impermeable and resistant against microbial and oxidative attack[55].

![Diagram of Laccase enzymatic activity.](image)

**Figure 1.7. Diagram of Laccase enzymatic activity.** Oxidation of phenolic subunits of lignin by laccase, Laccase functions via the catalyzation of one-electron substrate oxidations with a concurrent four-electron reduction of molecular oxygen to water.
In order to hydrolyze lignocellulosic material, we can use free enzymes. However, one efficient approach for degradation of lignocellulosic material in nature is the integration of cellulases and other associated enzymes into multi-enzymatic complex named the cellulosome [56]. Some anaerobic bacteria have evolved this specialized structure, a self-assembled nanomachine, the so-called cellulosome, that is highly efficient in this process. This molecular machinery is extremely complex and varied and seems to reflect the adaptation of the bacteria to the complexity of the plant cell wall. The cellulosome is composed of a scaffoldin (non-catalytic) subunit, two dockerins (recognition modules) and a cohesin, that are able to integrate various enzymes into the complex [57, 58]. *Clostridium thermocellum* is the most studied cellulosome producer. Its scaffoldin subunit comprises a series of 9 repeating cohesin modules, a single carbohydrate-binding module (CBM) and an X-dockerin that interacts with an attaching scaffoldin at the cell surface[59]. Cellulosomal enzymes contain a catalytic module and a specific dockerin module, which binds to the cohesins of the primary scaffoldin (Fig1.8).
Figure 1.8. Schematic representation of a cellulosome. In the legend the different components of the cellulosome are described. Each cohesin is selective for each dockerin, so each catalytic module has its own place in the designer cellulosome.

We are able to mimic this natural machinery creating a designer cellulosome for promoting synergistic action among enzyme components [60, 61]. This artificial complex uses recombinant chimaeric scaffoldins composed of cohesin modules originating from different bacterial species. Each cohesin binds specifically to the matching dockerin of the same species, containing the different enzymes. These nanomachines allow controlling the composition and architecture of the cellulosome assembly and have demonstrated to be efficient in the degradation of lignocellulosic materials [62-64].

The improvement of cellulases may be of great interest for different industries such as textile, food industry, but especially for the bioenergy industry because these enzymes are necessary in the bioconversion of biomass to bioethanol. The enzymes that we selected for improvement in this thesis are have been targeted thinking of their possible implementation in the production of bioethanol.
The degradation of lignocellulosic materials can produce large amounts of value-added products that can be obtained by different thermo-chemical and biochemical processes. Lignocellulosic materials, consisting of approximately 75% polysaccharide sugars, can be converted to bioethanol and fine chemicals (Fig 1.9). In the industrial process, a thermochemical pretreatment is required to liberate cellulose from hemicellulose and lignin [65-67].

Figure 1.9. Thermochemical and biochemical processing of lignocellulosic biomass [68].

Global climate change due to excessive carbon emissions, as well as the uncertainty and price instability of petroleum resources, have encouraged the development of new sources of energy that are sustainable and environmentally friendly, and can reduce the dependence on fossil fuels. In the past decade, numerous efforts have been made to implement bioethanol as a semi-renewable fuel, as it is capable of alleviating some of the issues associated with fossil fuels, especially those related to the environment [69].
New mandates have compelled the rapid implementation of bioethanol [44, 70], and it is currently used in fuel at a concentration of between 5 and 25%, depending on the country.

Plant cell wall cellulosic biomass is the most abundant source of energy-rich carbon in the biosphere [71]. This source of energy, gives several advantages compared with oil. Bioethanol can be produced virtually in any country, thus reducing oil dependence. Also, it reduces CO$_2$ emission palliating part of the environmental problem related to fossil fuels. In addition, the use of bioethanol is less toxic to humans as it does not contain sulfur and produces a lower level of particulates and toxic emissions such as sulfur dioxide. Besides, its volatility is lower and hence the smog formation is reduced [72]. Natural resources are nowadays used for the production of bioethanol because of their low cost and abundant supply. These sources include city and agricultural waste, giving the opportunity to generate value for waste product. Bioethanol generated from lignocellulosic material s nowadays called second-generation bioethanol as opposed to first-generation ethanol that uses food crops. However, generating bioethanol from cellulose is more complex and expensive than in the case of first-generation ethanol. The conversion of lignocellulosic bio-wastes into products of interest requires several steps (Fig. 1.10), which include the pre-treatment of the substrate, its saccharification to obtain fermentable sugars, and finally fermentation using microbes to produce the final chemical products of interest.

Once that lignocellulose is pretreated, enzymatic hydrolysis also known as enzymatic saccharification, is carried out at 50 ºC. The biological degradation of cellulose into glucose and pentose monomers is achieved using multiple enzymes in defined ratios that cooperate synergistically. Then, sugars can be fermented at 50 ºC to obtain ethanol and other chemicals by different
organisms. These two steps can be done separately (separate hydrolysis and fermentation, SHF) or simultaneously (simultaneous saccharification and fermentation, SSF) [73].

For successful cellulose degradation to sugar, enzymes must withstand the conditions of the industrial bioconversion process, such as high temperature, generally above 50 °C, and low or high pH [11, 67]. The lower efficiency of the enzymes under these conditions, make the saccharification process a critical bottleneck in the bioconversion of cellulose.

Figure 1.10. **Enzymatic lignocellulose conversion process into valuable products such as bioethanol.**

In this thesis we have improved a set of enzymes that can produce sugar from lignocellulosic material in a wide range of conditions. The activity of the reconstructed cellulase enzymes was compared to that of modern enzymes. In particular, we used a bacterial endoglucanase from *Thermotoga maritima*, *T.reesei* enzyme preparation and Ctec2 enzymes cocktail. These enzymes are commonly used for the bioconversion of cellulose. The ancestral cellulases showed considerably higher specific activities than those of modern ones under a broad range of temperatures and pH.
values with various substrates. We observed that an efficient bioconversion can be achieved by reconstructing a set of three enzymes as compared to other methodologies were hundreds of variants need to be tested. The reconstructed endoglucanase enzyme also displayed higher efficiency when integrated in a bacterial cellulosome, a macromolecular machine for cellulose degradation [59, 74], that has been also proposed for industrial implementation [12, 75]. The intention is to integrate the other two enzymes in the same cellulosome.

Our ancestral enzymes also showed very good synergy with other lignocellulosic enzymes such as laccase and xylanase, as well as incorporated into a bacterial cellulosome. We anticipate that the incorporation of additional enzymes with complementary activities towards cellulosic biomass degradation may result in even higher synergies and overall activities. Our resurrected enzyme targets a critical step of the process and is expected to result in an important reduction of the enzyme cost in industrial biomass degradation. Finally, the ancestral enzymes could be combined in the future with other reconstructed lignocellulosic enzymes to generate highly efficient cocktails providing the long-awaited improvement of the saccharification of cellulosic substrates.
Chapter 2: Methods for phylogenetic analysis

This chapter of this thesis, explains the methodology used in phylogeny in order to create phylogenetic trees and reconstruct ancestral sequences.

2.1. Introduction

The common name used for the mathematical and statistical methods used to infer ancient information (in the form of strings of characters) from current data is ancestral reconstruction. The main use of this technique, although it has some non-biological applications such as the phylogenies of the phonemes and vocabulary of ancient languages [76], oral traditions of extinct cultures [77] or ancestral marriage practices [78], is in the field of phylogenetics. That is why, the most common used characters are either protein or nucleic acid sequences, and the current data comes from the extant species that have been sequenced.
Phylogenetics is the study and correlation of the evolutionary relationships between extant individuals, species and populations and their corresponding ancestors [79]. Nowadays, by means of the so-called ancestral sequence reconstruction, it is possible to reconstruct ancestral biological macromolecules; polynucleotide sequences of DNA and distinct types of RNA, or amino acid sequences of proteins.

This technique is based on a sufficiently realistic statistical model of evolution to accurately recover ancestral states. In order to determine the route of the evolution [80] the genetic information already obtained through methods such as phylogenetics is used.

One of the precursors of the modern phyligenetics is cladistics, in fact, the idea of reconstructing ancestors from measurable biological characteristics, comes from cladistics. In cladistics, the organisms are classified based on the common characteristics that they share. Cladistics appeared as early as 1901 and infer the evolutionary relationships of species on the basis of the distribution of shared characteristics, of which some are inferred to be descended from common ancestors. The first person who is known to have carried out a cladistics analysis for birds is Peter Chalmers Mitchell [81, 82], followed by the works of Robert John Tillyard for insects (1921) [83] and Walter Max Zimmermann for plants (1943).

Emile Zuckerkandl and Linus Pauling in 1963 were the first persons that carried out works in ancestral sequence reconstruction. In 1955, Frederick Sangerce started developing techniques for sequencing the primary structure of proteins. This helped, Zuckerkandl and Pauling to propose that, based on the amino acid sequence of extant proteins, it is possible to infer the phylogeny of that protein and the sequences of all the common ancestors, including the earliest point of the tree [84, 85]. Beside
those pioneers works, it was in 1971 when Walter M. Fitch developed the first algorithm for ancestral sequence reconstruction using the principles of maximum parsimony [86].

2.2. Theory

Every effort on reconstructing ancestors starts with a phylogeny, a hypothetical tree that includes the order in how species are correlated between each other by descent from common ancestors, starting with the last universal common ancestor (LUCA). In a phylogenetic tree, terminal nodes correspond to the extant species. These nodes are successively connected to their common ancestors by branches. The common ancestors are the inner nodes. At the end, all the species and, therefore, all the evolutionary lines converge in the LUCA (Fig. 2.1).
Figure 2.1. Schematic representation of a phylogenetic tree. The three kingdoms are represented, Bacteria in green, Archea in red and Eukaryotes in blue. The last node (root) represents LUCA: Last Universal Common Ancestral. LUCA is the common ancestor of the three kingdoms.

2.2.1. Methods

The maximum parsimony method is a non-parametric statistical method. In order to infer the tree, it uses a set of extant sequences minimizing the amount of mutations that are necessary to match the available data. Some years later, in 1975, David Sankoff optimized this algorithm adding a cost to the mutations [87]. Because of the development of this work David.L.Swofford developed the first phylogenetics program [88] in 1989, called PAUP. It soon became very popular in the phylogenetics community.
At the same time, the exponential increase of the computing power made the implementation of much more complex algorithms possible: maximum likelihood approaches [89-91] or Bayesian methods [92-96].

The maximum likelihood algorithm searches for the most probable tree when the phylogenetics model and the extant sequences are estimated. In the Bayesian approach, the computer program searches for the highest posterior probability, which is determined on the one hand by the likelihood of the data under a certain evolutionary model and on the other hand by a set of prior probabilities set for the trees. Nowadays most of the procedures for ancestral sequence reconstruction are based in maximum parsimony, maximum likelihood and Bayesian inference.

In this thesis, Bayesian inference has been used for computing the phylogenetic tree, whereas maximum likelihood has been chosen to infer the extant sequences.

2.2.1.1 Parsimony

Parsimony refers to the principle of selecting the simplest of competing hypotheses. In the context of ancestral reconstruction, parsimony endeavours to find the distribution of ancestral states within a given tree which minimizes the total number of character state changes that would be necessary to explain the states observed at the tips of the tree.

One of the earliest examples of maximum parsimony implementation is Fitch's method [86], which assigns ancestral character states by parsimony via two traversals of a rooted binary tree. In spite of being really used, it has some evident limitations, Fitch’s approach overestimates the amount of rare changes [97].
Maximum parsimony is very useful due to its low computational costs and high efficiency for huge datasets and when ab initio phylogenies are needed [98] to optimize more complex algorithms. They are still used in some cases to seed maximum likelihood optimization algorithms with an initial phylogeny. However, the underlying assumption that evolution attained a certain end result as fast as possible is inaccurate. Parsimony methods impose five general assumptions that are not valid most of the times:

1. *Variation in rates of evolution.* Fitch's method assumes that changes between all character states are equally likely to occur; thus, any change incurs the same cost for a given tree. This assumption is often unrealistic and can limit the accuracy of such methods [99].

2. *Rapid evolution.* It assumes that mutations are rare. This assumption is not correct in cases of rapid evolution, such as some retroviruses [100-102].

3. *Changes in time among lineages.* Those methods accept that the same amount of evolutionary time has passed along every branch of the tree without taking into account the variation in branch lengths in the tree. They are often used to quantify the passage of evolutionary or chronological time. This limitation makes the technique responsible to infer that one change occurred on a very short branch rather than multiple changes occurring on a very long branch, for example [103]. In addition, it is possible that some branches of the tree could be experiencing higher selection and change rates than others, some periods of time may represent more rapid evolution than others, when this happens parsimony becomes inaccurate [104].
4. *Statistical justification.* Without a statistical model underlying the method, its estimates do not have well-defined uncertainties [101, 103, 105].

5. *Convergent evolution.* When considering a single character state, parsimony will automatically assume that two organisms that share that characteristic will be more closely related than those who don’t.

### 2.2.1.2 Maximum likelihood

This method assumes that the ancestral states are those which are statistically most likely, based on the observed phenotypes. The first works developed using an approach of this method was developed in the context of genetic sequence evolution [89, 90, 106]; similar models were also developed for the analogous case of discrete character evolution.

Using a model of evolution needs to take into account the fact that not all events are equally likely to occur. But it does not mean that they need to happen just because they are more likely to take place. Sometimes, the one with the less probability occurs, and in those cases, maximum parsimony may actually be more accurate because it is more willing to make large, unlikely leaps than maximum likelihood. Maximum likelihood is really reliable in reconstructing character states. However it not so good in giving accurate estimations of the stability of proteins as overestimates, since it assumes that the proteins that were made and used were the most stable and optimal [107].

In maximum likelihood Markov process models the evolution of the sequence, assuming that all the mutations are independent [108]. The likelihood of the phylogeny is calculated from a sum of intermediate probabilities of the nodes for the proposed tree.
This basic model is frequently extended to allow different rates on each branch of the tree. In reality, mutation rates may also vary over time (due, for example, to environmental changes). This can be modelled by allowing the rate parameters to evolve along the tree, at the expense of having an increased number of parameters. A model defines transition probabilities from states $i$ to $j$ along a branch of length $t$ (in units of evolutionary time). At each node, the likelihood of its descendants is summed over all possible ancestral character states at that node:

$$L_x = \sum_{Sx \in \Omega} P(Sx) \left( \sum_{Sy \in \Omega} P(Sy|Sx,txy)Ly \sum_{Sz \in \Omega} P(Sz|Sx,txz)Lz \right),$$

where the node $x$ is the ancestor of $y$ and $z$. $Sx$ represents the sequence of the $i$-th node, $tij$ refers to the branch length from $i$ to $j$.

$\Omega$ is the set of all the possible combinations (the four nucleotides or the 20 basic amino acids).

Thus, the objective of ancestral reconstruction is to find the assignment for all $x$ internal nodes that maximizes the likelihood of the observed data for a given tree.

**Marginal and joint likelihood**

The problem for ancestral reconstruction is to find the combination of character states at each ancestral node with the highest marginal maximum likelihood. In order to find the most probable evolutionary lineage to the common ancestor, two different conventions have been proposed. First, one can consider the probabilities of all the descendants for a certain ancestor and
calculate the joint combination with the maximum likelihood. This approach is called joint reconstruction. And second, instead of calculating the global likelihood, one can successively select the most likely ancestor for every node. This procedure is referred to marginal reconstruction. Joint reconstruction is more computationally complex than marginal reconstruction. Nevertheless, efficient algorithms for joint reconstruction have been developed with a time complexity that is generally linear with the number of observed taxa or sequences [91].

### 2.2.1.3 Bayesian inference

Bayesian inference employs both the likelihood of the experimental data, described before, and a prior knowledge about the possible solutions. Thus, the aim in ancestral sequence reconstruction is to obtain the posterior probabilities for every internal node of a known tree. Furthermore, the posterior probabilities can be combined with the posterior distributions over the parameters for a given evolutionary model and the structure of all possible trees. This results in the following applications of Bayes’ theorem:

\[
P(S \mid D, \theta) = \frac{P(D \mid S, \theta) P(S \mid \theta)}{P(D \mid \theta)}
\]

\[
\alpha P(D \mid S, \theta) P(S \mid \theta) P(\theta),
\]

where \( D \) is the experimental data, \( S \) corresponds to the ancestral states and \( \theta \) represents the phylogenetic tree and the evolutionary
model. Equation (2.2), \((D \mid S, )\) represents the likelihood of the experimental data that could be computed, \((S \mid \theta)\) refers to the prior probability of an ancestral node for a known tree and model and \((D \mid \theta)\) corresponds to the probability of the data for a known tree and model, integrated for all possible ancestral states.

Note that two different formulations have been given (2.2) and (2.3), one for each of the applications of Bayesian inference, the empirical and the hierarchical Bayes. Empirical Bayes approach estimates the probabilities of several ancestral nodes for a given tree and model of evolution. On the other hand, hierarchical Bayes approach calculates these probabilities over all possible trees and model of evolution, comparing how likely they are, with a given experimental data [109].

2.3. Methodology

In order to construct the ancestral trees of cellulases, bioinformatics tools have been used. The use of different software allows the whole process of constructing those trees. These different programs are described below.

First of all, a query was selected and making a blast of it in the protein databank, the amino acid sequences were obtained. It is a comprehensive resource for protein sequence and annotation data that includes different databanks. After downloading these sequences, they have to be aligned, cleaned and finally the tree has to be constructed, the following programs were the ones used. The scheme of the used methodology and software is shown in Figure 2.1.
Methods for Phylogenetic Analysis

**Figure 2.1. Methodology used for the reconstruction of ancestral sequences consisting in four steps:** (1) Selection of extant sequences, (2) creation of a multiple alignment, (3) construction of a phylogeny, and (4) reconstruction of ancestral sequence. Close to each step the icon of the program that has been used for each step is shown.

### 2.3.1. Selection of extant sequences: Uniprot

In order to reconstruct a phylogenetic tree first, homologous sequences of the protein of interest need to be found. This means that, sequences of the different species chosen descend from the same common ancestor. In this way, those residues that are exactly the same at the same position are identical by character state in the particular sequences. Homologue sequences in this work, were retrieved from UniProtKB [110] online database using BLAST (Basic Local Alignment Search Tool) [111]. By using
this tool, regions of local similarity were searched between sequences, which gave later the possibility to infer evolutionary relationships. UniProtKB (Universal Protein Resource Knowledge Base) [39] is a catalog of information on proteins. To find the protein of interest (for now on, query), one can either use the search tool (Fig. 2.2) or directly enter the protein sequence, or its UniProt identifier.

![Image](image.png)

**Figure 2.2. Search tool of the Uniprot Database.** The search of the sequences was made by using a query. This query is the sequence of the specie used for the search of the homologous sequences.

The first step in this database was the selection of the query of the desired protein. This is the sequence from which the database will select the rest of the homologous proteins. Once the selection of the query was done, the BLAST tool was used to find the homologous sequences. Some parameters were set (those parameters are described below table 2.1):
Methods for Phylogenetic Analysis

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target Database</td>
<td>The search is performed against this database. Different phyla can be chosen for search.</td>
</tr>
<tr>
<td>E-Threshold</td>
<td>Statistical measure of the number of expected matches in a database. The bigger this value is,</td>
</tr>
<tr>
<td></td>
<td>the more unlikely to be significant a match.</td>
</tr>
<tr>
<td>Matrix</td>
<td>It gives a probability score for the position of each aminoacid in an alignment. For this</td>
</tr>
<tr>
<td></td>
<td>probability, the BLOSUM [112] matrix is based on the frequency with which that substitution is</td>
</tr>
<tr>
<td></td>
<td>known to occur among consensus blocks within related proteins.</td>
</tr>
<tr>
<td>Filtering</td>
<td>The filtering can be made by masking the lookup table or by lower complexity regions.</td>
</tr>
<tr>
<td>Gapped</td>
<td>Gaps can be introduced after having selected the sequences.</td>
</tr>
<tr>
<td>Hits</td>
<td>The number of hits in a search can be chosen.</td>
</tr>
</tbody>
</table>

*Table 2.1. Description of the parameters that can be changed in Uniprot Search tool. The search of the homologous sequences was done by changing these parameters.*

The values used for each parameter in this work were the ones shown in Figure 2.3.

*Figure 2.3. Uniprot search tool. According to the description of the parameters done in Table 2.1, these were the selected parameters in the search of our sequences.*
After selecting all the parameters as shown in Figure 2.3 “Run BLAST” was pressed in order to obtain the homologous sequences of the previously selected query. The process can take several minutes depending on different reasons, such as complexity of the query sequence, length of the sequence or the applied parameters. Once the process was over, homologous sequences were shown.

The next step consisted on selecting the sequences of interest. To do this, the identity between the regions of the different sequences with the query was taken into account. Finally, the selected sequences were download in the most appropriate format for the following steps: FASTA format, which is a text file that is commonly used for the aminoacid sequences. This format is shown in Figure 2.4.

Figure 2.4. Fasta format file. Example of sequences in fasta format, this type of files can be opened in text format.
2.3.2. Creation of a multiple alignment: MUSCLE

Following the methodology to construct ancestral trees, the next step was the multiple alignment of the sequences selected in the previous step. MEGA is a very versatile tool for phylogenetic and molecular evolution analysis. This program is a package that is useful for: aligning sequences by ClustalW and MUSCLE, estimating phylogenetic trees by a variety of methods (Neighbor Joining, Maximum Parsimony and Maximum Likelihood), estimating rates of molecular evolution, inferring ancestral sequences and drawing those trees in different ways. Besides it’s multiple applications, in this work, it was used for two main purpose, to do the alignment and to select the best model [113]. Regarding to the alignment, over the last years, many algorithms were developed for this this purpose, with Clustal [114] and MUSCLE [115, 116] being the most popular in the phylogenetic community. In this work, the MUSCLE algorithm was used for all the multiple sequence alignments. In order to make the sequence alignment of the selected sequences, the FASTA file downloaded from the UniProtKB was loaded in MEGA [117-119]. The sequences appeared unaligned as it is shown in Figure 2.5.

**Figure 2.5.** Sequence alignment before aligning. This figure shows the sequences before aligning and editing the alignment.
Once this appeared, MUSCLE was run by pressing “align with MUSCLE”. Another window appeared the showing the different parameters that can be changed before running the algorithm (Fig. 2.6). The parameters are shown in Table 2.2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gap Opening Penalty</td>
<td>Increasing or decreasing this value gaps become more or less frequent in the alignment</td>
</tr>
<tr>
<td>Gap Extension Penalty</td>
<td>The bigger this value is, the shorter the gaps are in the alignment. Terminal gaps do not penalize</td>
</tr>
<tr>
<td>Max Memory in MB</td>
<td>The upper limit of the memory used by the algorithm in the computer. By living it default, the use of all the computer resources is avoid</td>
</tr>
<tr>
<td>Max Iterations</td>
<td>Maximum number of allowed iterations</td>
</tr>
<tr>
<td>Clustering Method (1,2 iterations)</td>
<td>The clustering method used for the first two iterations</td>
</tr>
<tr>
<td>Cluster Method</td>
<td>The clustering method use for the rest of iterations</td>
</tr>
<tr>
<td>Max Diagonal Length</td>
<td>Maximum length of the diagonal of the matrix made by aligning the sequences</td>
</tr>
</tbody>
</table>

Table 2.2. Description of parameters that can be changed in MEGA software in order to align the protein sequences. By changing these parameters the alignment was done.

In this work, the following parameters were used for the aligning (Fig 2.6).
Methods for Phylogenetic Analysis

Figure 2.6. Screenshot of the window with the parameters that can be changed for the alignment of the sequences. According to the description of the parameters done in Table 2.1, these were the selected parameters in the search of our sequences.

In order to run the algorithm “Compute” was selected. Depending on how many iterations have been selected the process will last longer or shorter. Once all the iterations were finished, the alignment of all the sequences was obtained, as shown in the Figure 2.7.

Figure 2.7. Image of the protein sequences once the alignment was done and it was edited. The edition of the alignment was done manually and after it an almost no-gap matrix was obtained.
Analyzing the alignment some asterisk can be seen in the top of the alignment, which means that the residue below the asterisk is conserved in all the species. The colors, they are related to the biochemical properties of the aminoacids. So as to remove ambiguously aligned regions, GBLOCKS [120] can be used or there is a possibility of doing it manually. The other tool of MEGA used in this work was the choosing of the best model (Fig 2.8).

![Table: Maximum Likelihood fits of all different amino acid models](image)

**Figure 2.8. Screenshot of the parameters available to modify in order to select the best model for the construction of the tree.** a) Best model selection in MEGA, b) Parameters used for the tree construction available in MEGA.

### 2.3.3. Computing a phylogenetic tree

The way the phylogenetic trees were made in this work was by using BEAST (Bayesian Evolutionary Analysis Sampling Trees) [121, 122]. This a package of programs for Bayesian analysis of molecular sequences using MCMC (Markov chain Monte Carlo), a class of algorithm for sampling the probability distribution based on constructing a Markov chain. BEAST can be used for reconstructing phylogenies using MCMC to average over tree space. In this way, each tree is weighted proportional to its
posterior probability. It is a cross-platform program for Bayesian MCMC analysis of molecular sequences. It is entirely orientated towards rooted, time-measured phylogenies inferred using strict or relaxed molecular clock models. It can be used as a method of reconstructing phylogenies but is also a framework for testing evolutionary hypotheses without conditioning on a single tree topology. MCMC is used to average over tree space, so that each tree is weighted proportional to its posterior probability. It includes a graphical user-interface for setting up standard analyses and a suit of programs for analyzing the results. It is a software package that allows a phylogeny analysis [122].

2.3.3.1. Beauti

By using Bayesian Evolutionary Analysis Utility (BEAUti) a input file (.xml) was created in order to run BEAST. It is a graphical user interface that allows to set the evolutionary model and options for the MCMC Once the aligned sequences are imported (through a NEXUS file), the interface with several tabs permits modifying many parameters as it is shown in Figure 2.9 and explained in Table 2.3.

Figure 2.9. Graphical user-interface (GUI) application of BEAUti. The figure shows the parameters that can be set to generate the file to construct the tree. The description of each parameter is shown in Table 2.3.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partitions</td>
<td>Makes possible to load sequences that were not included initially and making partitions</td>
</tr>
<tr>
<td>Taxa</td>
<td>Gives the possibility of making subgroups with the taxa. Also, this subgroups can be force to be monophyletic</td>
</tr>
<tr>
<td>Tips</td>
<td>Allows data selection of individual taxa</td>
</tr>
<tr>
<td>Traits</td>
<td>Phenotypic trait analysis can be set</td>
</tr>
<tr>
<td>Sites</td>
<td>Selection of the substitution model is allowed and the site heterogeneity model</td>
</tr>
<tr>
<td>Clocks</td>
<td>Gives the chance to choose the clock model. The mutation rate of biomolecules is used in those models to estimate when they diverged</td>
</tr>
<tr>
<td>Trees</td>
<td>The tree prior is set</td>
</tr>
<tr>
<td>States</td>
<td>Permits to reconstruct the states of all the ancestors or only choosing some subgroups</td>
</tr>
<tr>
<td>Priors</td>
<td>Sets the prior distribution for the subgroups</td>
</tr>
<tr>
<td>Operators</td>
<td>Allows to use or not some of the parameters in other tags</td>
</tr>
<tr>
<td>MCMC</td>
<td>MCMC value for the computing of the tree can be choose</td>
</tr>
</tbody>
</table>

**Table 2.3. Description of the parameters available to adjust in order to make the tree.** By changing these parameters the correct file is generated in order to run the tree (explained in section 2.3.3.2).

As it was explained before, this programs, has a tool that gives the opportunity of forcing phylogenetetic groups. In this work this tool was used and it can be seen in the **Figure 2.10**.
Methods for Phylogenetic Analysis

Figure 2.10. Description of the way the interface allows to make different groups in a set of sequences.

“Generate BEAST File…” was the next step to obtain the XML file for a further phylogeny computing using BEAST.

2.3.3.2. Beast

In order to run BEAST the previously generated .xml file was used, The output was a .log file. The log file records a sample of the states that the Markov chain found. In order to compute the phylogeny, BEAST graphical interface was opened first (Fig. 2.11). Here, the XML file was opened. The application also has the option to activate the BEAGLE library [123]. BEAGLE is a high-performance library that takes advantage of the parallel
processors available in most of current PCs. Using this option is highly advisable to improve the performance of the program.

Figure 2.11. Graphical user-interface (GUI) application of BEAST. The file generated by Beauti (xml file) can be run to construct the tree.

Once the XML file of interest was selected and BEAGLE library option chosen, the program was run and the phylogeny started computing (Fig. 2.12).

Once the process was over, the program generated the log files that contain in the information of the process. An examination of this output was needed to determine whether the Markov chain was run for long enough to obtain accurate estimates of the parameters. Another application, Tracer, was used for this analysis (Fig 2.13).
2.3.3.3. TreeAnotator

By using TreeAnotator the sample of trees generated by BEAST was summarize in a single consensus tree. The obtained tree contains information about the posterior probabilities of the nodes in the consensus tree, the posterior estimates and the rates. In the following figure (Fig 2.12) the different parameters are described.

![Figure 2.12. TreeAnotator graphical interface and its options. This tool allows to select the most probable tree from the set of trees obtained by running the xml file in Beast.](image)

2.3.3.4. Tracer

Tracer is a graphical interface (Fig. 2.13) that makes possible the monitorization and analysis of the MCMC output carried out in BEAST. In order to perform the analysis, the log file obtained previously (POINT) that corresponds to the analysis of the phylogenetic computation was opened, many parameters related to MCMC analysis appeared on the left side of the interface, which were pondered by their Effective Sample Sizes (ESSs).
Figure 2.13. Tracer graphical interface. Normal distribution of the probabilities is shown in this image.

A low ESS means that the trace contained a lot of correlated samples and may not represent the posterior distribution well. It is advisable to run BEAST again until ESS reaches a value higher than 100.

2.3.3.5. Figtree

Finally, there are plenty of programs to draw the tree obtained by the previous steps. Figtree was used in this work, which is a program included in BEAST package.

This program is able to read tree files in both Nexus and Newick format. In the case of this format, Figtree has an extended version that includes parameters as fonts. It permits re-rooting, rotating
Methods for Phylogenetic Analysis

clades, showing or hiding labels etc. This is shown in Figure 2.14.

Figure 2.14. FigTree graphical interface and its options. A consensus tree can be seen in the figure.

2.3.3.6. TreeGraph

TreeGraph 2 is a graphical editor for phylogenetic trees which allows you to apply lots of graphical formats to the elements of your tree. Moreover, it supports several (visible or invisible) annotations (e.g. support values) for every branch or node. These annotations can be imported from Nexus tree files or text files containing data in a table (e.g. exported from a spreadsheet program).
2.4. Reconstruction of ancestral sequences

Following the whole procedure described previously, a consensus tree was obtained and once being confident with the statistics, the ancestral sequences of interest were inferred. There are many algorithms and programs to obtain these sequences, but PAML (Phylogenetic Analysis Using Maximum Likelihood) [124, 125] was used to infer all the ancestral sequences in this work. PAML is based on the maximum likelihood algorithm mentioned in previous sections.
2.4.1. PAML

PAML is a package of programs for maximum likelihood analysis of protein and DNA sequences. In this thesis, Codeml was used in this thesis for the reconstruction of ancestral codons and proteins. In order to run this, it is needed to include in a new file a sequence data file, a tree file (in Newick format), a matrix file (Jones matrix in this case) and the control file (Fig. 2.16).

![PAML program running the calculations for the ancestral reconstruction of proteins](image)

**Figure 2.16. PAML program running the calculations for the ancestral reconstruction of proteins**

Once everything was ready, the executable file was run and a rst file was created, containing all the information concerning the process the posterior probabilities and the joint and marginal protein reconstructions. **Figure 2.17** shows the information obtained in this type of files.
Figure 2.17. Rst file obtained in PAML. All the information of the tree is contained in this file, including joint and marginal probabilities of each node.
2.5. Applications

The applications for ancestral reconstruction have increased exponentially in the last two decades. In the field of molecular evolution the more outstanding advances have been in the optimization of the fluorescence performance of opsins [126] and GFP proteins [127], novel anticancer drug’s mechanism and design [32], the uric acid and evolution in mammals [128], the amino acid persistence in proteins [57] or mammalian diving capacity evolution [129].

Regarding to other applications, the fields of usage have been such as, calculating spatial migration traits in order to infer the location of the ancestors [130], inferring ancestral ranges of species from phylogenetic trees in order to obtain historical biogeographic ranges [131] and genome rearrangements [132].
3.1. Molecular biology techniques

The subsequent points explain the procedure of the molecular biology techniques used in this thesis in order to obtain both the ancestral and the extant cellulases in the laboratory. First, the DNA sequences that encode the protein of interest must be purchased. After that, the genes need to be inserted into a bacterial expression vector, this is called ligation. Later, the plasmids are transformed into host bacteria in order to induce the protein expression. Once the protein of interest is expressed, the cells should be lysed to liberate the proteins. Finally, the resulting protein must be purified. This process is shown in Figure 3.1, step by step, since the cloning to the purify protein.
Experimental methods

Figure 3.1. Schematic representation of the molecular biology techniques used to produce proteins. First, the gene of interest is inserted in the expression plasmid and transformed in the bacteria. Then, the bacteria is grown and after the protein expression test a big culture is induced. Finally, the protein is obtained by cell lysis.

3.1.1. Cloning of commercial plasmid

Cellulases encoding genes were codon optimized for E.coli and purchased in a commercial plasmid (Life Technologies). This plasmid contains an antibiotic resistance gene for the proper selection. This antibiotic was carbenicillin for all the genes but for the exoglucanase one that was kanamycin. The antibiotic ensures the proper selection of bacteria, being the only E.Coli colony grown in the plate. 1 µL of the commercial plasmid (50 ng/µL) was transformed into E.coli-XL1Blue competent cells (Agilent Technologies) following the manufacturer’s protocol [133]. Once
transformation was performed, competent cells were grown in 400 µL of SOC medium (Invitrogen) for one hour and spread in LB-agar-antibiotic (the selected one in each case) plates and incubated overnight at 37 ºC.

Single colonies were isolated and grown in 10 mL of LB media + 1% 100 mg/mL kanamycin for 16 h at 37 ºC gently stirring. The harvesting of cells was made by centrifugation (14000 rpm, 10 min, 4 ºC, Eppendorf Centrifuge 5810R) and plasmids were extracted using a so called miniprep kit, DNA-plasmid extraction kit (Thermo Scientific) following the company’s protocol [134]. Purified plasmids were eluted in 50 µl of nuclease-free water and their concentration was measured in the Nanodrop 2000L system.

3.1.2. Digestion of commercial plasmid

The enzymatic digestion of the commercial plasmid containing the cellulase gene was carried out after amplification. In order to perform the enzymatic digestion, a double digestion strategy with BamHI – KpnI cutting was used. BamHI and KpnI restriction sites are flanking the borders of the cellulase gene.

The enzymes used for the digestions were purchased from Thermo Scientific and the protocol used was the manufacturer’s Fast Digest protocol. The final digestion volume is adjusted to 50 µL and incubated at 37 ºC for one hour. The screening of the digestion products was made in a DNA-agarose gel (1%) in TAE buffer. The running of the DNA-agarose gel was carried out using the BioRad agarose electrophoresis equipment for approximately 90 min. After this time, the band corresponding to the cellulase was extracted from the gel and the gene was purified with a
Experimental methods

DNA-extraction kit from Thermo Scientific following the usual protocol [135]. Concentrations were also measured using the Nanodrop 2000L.

The host used for the insert was pQE80 plasmid, it can be seen in Figure 3.2.

![Figure 3.2. Scheme of the PQE80 plasmid use for hosting the cellulases.](image)

In the section point (3.1.3) the protocol for the construction is explained.
3.1.3. pQE80-cellulase construct ligation

Once the digestion is made and the gene purified, the genes encoding the cellulase must be ligated onto a high-efficiency bacterial expression vector with compatible cohesive ends. The previously digested BamHI-pQE80-KpnI open plasmid was used (Fig. 3.3). This plasmid was a kind gift from Professor Julio Fernandez’s lab at Columbia University. It also contains an ampicillin resistance gene. For the ligation of the gene encoding cellulase and the pQE80 plasmid Invitrogen’s T4-DNA ligase protocol [136] was used. The mol ratio between the amount of plasmid vector and the cellulase gene insert is 3:1. With the following formula, the calculations for the needed amount of plasmid and DNA inserts were done. Ligations were incubated overnight at room temperature. Thereafter, to stop the process, ligations were diluted 5 times with deionized water.
**Figure 3.3.** Schematic representation of the product of the ligation between the plasmid and the desired cellulase: PQE80-beta-glucosidase ligation in this case.

### 3.1.4. Cloning of pQE80-cellulase plasmid

5 µL of the recombinant plasmid (depending on the concentration) were transformed into E. coli-XL1Blue competent cells following the same protocol described in 3.1.1 section. Competent cells are later spread out onto LB-agar-ampicillin plates and incubated overnight at 37 °C. In the same way previously described, single colonies were taken out and grown in 10 mL of LB media + 0.1% 100 mg/mL ampicillin for 16 h at 37 °C. Finally, cells were harvested by centrifugation (14000 rpm, 10
min, 4 °C) and plasmids were extracted using the same so called miniprep kit, DNA plasmid extraction kit. The purified plasmids were eluted in 50 µl nuclease-free water and their concentration was measured in the Nanodrop 2000L. The plasmids are screened and verified in a DNA-agarose (1%) gel and concentration is calculated using the Nanodrop 2000L system.

3.1.5. Screening

Once the pQE80-cellulase constructions were made and amplified, an amount between 1-10 µL of the plasmid was transformed onto E.coli-BL21 competent cells following the seller protocol [137]. After transformation, cells were grown as previously was made with EColi-XI1blue in 400 µL of SOC medium for 1 hour at 37 °C and spread out in LB-agar plates with the corresponding antibiotic. Plates were incubated overnight at 37 °C to grow the colonies. Some single colonies were isolated and grown in 10 mL LB medium + antibiotic for 8 hours or until the optical density (OD) of the medium reached 0.6. ODs were measured with the Nanodrop 2000L.

In order to induce the overexpression of cellulases by T7 promoter activation, 5 µL of IPTG (isopropyl-β-D-thiogalactopyranosid, Sigma Aldrich) 100 mg/mL was added to the half of the volume of medium and the solution was incubated overnight at 37 °C. 1 mL of each colony was taken then, to screen the overexpression. Bacteria were harvested by centrifugation (14000 rpm, 10 min, 20 °C). Supernatant was discarded and bacteria were resuspended in 20 µL of extraction buffer. 20 µL of 2xSDS page Sample Buffer solution was added to each sample for the denaturation and charging of the protein in acrylamide electrophoresis gel separation. The samples were again centrifuged (14000 rpm, 30 min 20 °C) and boiled at 95 °C for 3 min.
Experimental methods

The screening was carried out by running 20 µL of each of the solutions are run in an 8-12% acrylamide gel for approximately 1 hour in a BioRad acrylamide electrophoresis system. 12% gel has been used in the case of endoglucanase as it size is 33kDa. However in the case of exoglucanase 70kDa and beta-glucosidase 82kDa, 8% acrylamide gel has been used. After the run, gels were cleaned in deionized water for 30 min. Proteins in the gel were stained with Bradford solution (Thermo Scientific) for 20 min and cleaned with deionized water again. Negative controls without IPTG are also added to the gel to visualize the overexpression better (Fig. 3.4).

![Figure 3.4. 8% Acrylamide electrophoresis gel for beta-glucosidase screening (~82000 KDa). Protein ladder (Thermofisher) can be visualized in the left side of the picture. C1-C4 nomenclature refers to the 4 colonies isolated. The ones in which iptg is written are the induced ones.](image)

3.1.6. Protein production

The best overexpressed colony was selected and 1 mL of LB media with the desired bacteria was added to 1 L more LB media + 0.1% 100 mg/mL of the corresponding antibiotic + 0.1% 50
mg/mL chloramphenicol (it was added to maintain the ability of overexpression of the bacterial pLys system). The culture was incubated for about 8h until OD > 0.6 at 37 °C shaking (250 rpm). Once the desired OD was reached, 0.1% 100 mg/mL IPTG was added to induce the overexpression of the protein. The culture was again incubated overnight (16 h more or less) at 37 °C while shaking.

After doing this, bacteria were separated from the media by centrifugation (4000 rpm, 4 °C, 20 min) and the supernatant discarded. The pellet was then resuspended in 16 mL of extraction buffer and 160 µL of protease inhibitor (Merck Millipore) was added and incubated rocking (5 rpm) for 30 min at 4°C with 160 µL of 100 mg/mL lysozyme (Thermo Scientific) solution for the enzymatic destabilization of the bacterial membrane. Once this is done, a series of reagents are added: 1.6 mL of 10% Triton X-100 (Sigma Aldrich) for the chemical destabilization of the bacterial membrane; 80 µL of 11 mg/mL DNase I (Invitrogen) for the enzymatic degradation of DNA; 80 µL of 1 mg/mL RNase A (Ambion) for the enzymatic degradation of RNA; 160 µL of 1M MgCl2 (Sigma Aldrich) as a catalyst to increase the enzymatic activity of DNase and RNase. The suspension is incubated again for 10 min at 4 °C with rocking prior to the cell lysis. Cell lysis was carried out by French press (G. Heinemann HTU DIGI-F Press). Cells were introduced in the press chamber and lysed at 18000 psi during 30 min. The lysis product obtained was then centrifuged in a high-speed centrifugation system (33000 rpm, 4 °C, 90 min; Beckman Coulter Avanti J-26 XPI).
3.1.7. Protein purification

Regarding to the purification process, four different enzymes have been purified during this thesis and different purification processes have been carried out for them. The details have been described below.

3.1.7.1. Ancestral endoglucanase

The purification of ancestral endoglucanase was carried out first by temperature and then using a HisTrap column. After the centrifugation described in section 3.1.6 (in this case with 30 min is enough) the supernatant was transferred to a 50 ml tube and it was incubated in a water bath at 50 ºC for 20 min. Then, the sample was cooled in ice for 5 min and centrifuged to eliminate debris at 4000xg for 10 min.

After the temperature step, the second step was carried out with the HisTrap cobalt affinity resin (Thermo Scientific). All the cellulase constructs contain a HisTag composed of 6 consecutive histidines in the N terminus of the construct which poses the ability to specifically bind to the cobalt affinity column. This binding was later eluted by adding imidazol in the buffer. A 150mM imidazole buffer was used for the elution.

3.1.7.2. Ancestral exoglucanase

For exoglucanase, the first purification process was carried out by means of a HisTrap nickel affinity resin (Thermo Scientific). In this case, niquel one was used, as the exoglucanase is harder to
purify. The niquel resin has a stronger affinity but it is not as specific as the cobalt one is. This binding can was later eluted by adding imidazol in the buffer. A 150mM imidazole buffer was used for the elution.

The second purification process used was by means of size exclusion and it was carried out with an ÄKTA pure fast protein liquid chromatography (FPLC) system (GE Healthcare) with a Superdex 200 column of 30 cm (GE Healthcare). Fractions of interest were collected from the chromatogram and stored in Acetate buffer 50mM (pH 5.5).

3.1.7.3. Ancestral beta-glucosidase

In the case of beta-glucosidase, the same process was used with some changes. In the first purification process instead of using niquel resin, cobalt resin was used.

Regarding to the second purification process, the buffer used for the elution in the size exclusion process was PBS (pH7).

3.1.7.4. Extant T.maritima

The extant T.maritima was purified in the same way of the ancestral endoglucanase, both for the first purification step and for the second purification step.

3.1.8. T.reesei cocktail protein determination

The determination of the protein content of the cocktail was first made by the dry weight method [138] for protein content determination. For that porpoise size exclusion chromatography was used, using a Superdex 200HR column, eluted in water. Then the sample was freeze dried and it was weighted. Second, absorbance at 280 was measured of a purified fraction and used densitometry and mass spectrometry for determining
concentration of endoglucanase. Moreover, the protein concentration was determined by the BCA assay (Pierce) [139] using a BSA standard supplied with the kit and a standard of our ancestral endoglucanase LFCA.

3.2. Biochemical assays

3.2.1 CMC

This assay is specific for measuring the endoglucanases activity. Endo-b-1,4-D-glucanase (EC 3.2.1.4) randomly cleaves accessible intermolecular b-1,4-glucosidic bonds on the surface of cellulose. Water-soluble derivatives of cellulose such as carboxymethylcellulose (CMC) and hydroxyethylcellulose (HEC) are commonly used for endoglucanase activity assays because insoluble cellulose has very low accessible fractionation of b-glucosidase bonds to cellulase [140-142]. The reaction of hydrolysis can be determined in different ways: by measuring the changes in reducing sugars, viscosity or color but the assay recommended for the endoglucanase (CMCase) assay is a fixed conversion method. This method, requires 0.5 mg of absolute glucose released under the reaction condition [143]. The reducing sugars concentration is finally measured by the DNS method [144].

Cellulolytic activity of ancestral endoglucanase (LFCA) was tested at 50 mM and pH 4.8 citrate buffer with 2 % CMC (Sigma), 30 min at various incubation temperatures. Cellulases from T.maritima and T.reesei (1,4-(1,3:1,4)-β-D-Glucan 4-glucano-hydrolase (EC 3.2.1.4), C2730 Sigma Aldrich) were used as controls. In addition two blanks were also prepared; the substrate blank (0.5 ml of CMC solution + 0.5 ml of citrate buffer) and the enzyme blank (0.5 ml of CMC solution + 0.5 ml
of dilute enzyme solution). Both the substrate and enzyme blanks were treated identically as the experimental tubes. Enzymatic reactions were terminated by placing the tubes into an ice-water bath. Enzymatic activity was determined quantitatively by measuring soluble reducing sugars released from the cellulosic substrate by the dinitrosalicylic acid (DNS) method. A volume of 3 ml of the DNS solution was added to each sample and the reaction mixtures were boiled for 5 min. After boiling, tubes were cooled and after adding 20 ml of distilled water, absorbance was measured at 540 nm.

![Figure 3.5. Color change between blanks and reactions with DNS method. The yellow one is the blank, the darker the color is the bigger the reaction. It means that more sugar has been produced in the reaction.](image)

A glucose standard curve was used to determine the concentration of the released reducing sugars. For this porpoise, the following standards were prepared: GS1 – 0.125 ml of 2 mg/ml glucose + 0.875 ml of buffer. GS2 – 0.250 ml of 2 mg/ml glucose + 0.750 ml of buffer. GS3 – 0.330 ml of 2 mg/ml glucose + 0.670 ml of buffer. GS4 – 0.500 ml of 2 mg/ml glucose + 0.500 ml of buffer. The glucose released by the enzyme solutions was calculated with
deduction of the enzyme blank absorbance based on the glucose standard curve.

The determination of the pH dependence was done as following: purified enzymes were diluted in 50mM buffer at different pH values between 4 and 12. Activities were measured with 2% CMC at 70ºC for 30 min. All assays were performed in triplicate and the average value with standard deviation was determined.

3.2.1.1. Residual and long-term activity measurements

On the one hand the determination of the residual activity was carried out, to determine when the enzyme loses half of its activity. The enzymes diluted in citrate buffer 50 mM at their optimum pH, were incubated at different temperatures (60-90ºC). The residual activity was measured on 2% CMC for 30 min at 60ºC. The amount of reducing sugars was measured and quantified by the DNS method. The parameter $T_{50}$ is defined as the temperature at which an enzyme loses 50% of its optimal activity after a 30 min heat treatment [145].

On the other hand, a study of the activity of the enzymes in different times was done, the long-term activity. In this case, all measurements were conducted in 50 mM citrate buffer, pH 4.8 on 2% CMC at 60ºC for a period of 10 to 240 minutes. After hydrolysis, the reducing sugar concentration was measured by the DNS method.

3.2.1.2. Inactivation constant (Kin) determination

The objective of this assay was the determination of the inactivation constant, for this purpose, enzymes were incubated at
80°C during different time intervals diluted in their optimum pH. The amount of reducing sugar was measured and quantified by the DNS method. The inactivation constant ($K_{in}$) was calculated using the equation $\log \left( \% \text{ residual activity} \right) = 2.303 \times K_{in} \times t$, where $t$ is time [11]. The half-lives of the enzymes were calculated from the plot.

### 3.2.2 CellG3

Endoglucanase activity was also measured using another different method, the CellG3 method of an endoglucanase assay kit (K-CellG3, Megazyme International, Ireland)[146]. As controls, cellulases from *T.maritima* and *T.reesei* (C2730, Sigma Aldrich) were used. Enzyme samples were diluted in acetate buffer (100 mM, pH 4.5) and after the addition of CellG3 substrate enzyme solutions were incubated at different pH’s and temperatures. The incubation was carried out for 10 min. Cellulase cleaved a bond within BCI(PNP)β-G3, the non-blocked reaction product containing the 2-chloro-4-nitrophenyl substituent was instantly cleaved to D-glucose and free 2-Cl-4-nitrophenol (CI(3)NP). Finally, the hydrolysis reaction was stopped by addition of Trizma base solution (pH 9) and the CI-phenolate color was developed and measured at 400 nm (NanoDrop 2000C). CellG3 Unit was defined as the amount of enzyme required to release one micromole of 2-chloro-4-nitrophenol from CellG3 in one minute under the defined assay conditions, the enzyme activity was calculated multiplying the measured absorbance at 400 nm by 9.64 and by the dilution factor [146].
Figure 3.6. Colorometric change reaction of CellG3 kit [15]. Cellulase cleaved a bond within BCI PNPβ-G3, the non-blocked reaction product containing the 2-chloro-4-nitrophenyl substituent was instantly cleaved to D-glucose and free 2-Cl-4-nitrophenol (ClPNP). The color changed to yellow at 400nm absorbance in presence of beta-glucanase.

3.2.3 Filter-paper

In this case, this assay was used for the determination of a total cellulase system made of three cellulases: endoglucanases, exoglucanases, and β-glucosidases. Total cellulase activities were measured using insoluble substrates, including pure cellulosic substrates such as Whatman No. 1 filter paper or any other lignocellulosic substrate [147]. Filter-paper assay FPA is the most common total cellulase activity assay recommended by IUPAC.
The assay based on a fixed conversion degree, measures the hydrolysis of both, crystalline and amorphous cellulose of the filter paper. In this case, the activity of the total cellulase is described in terms of filter-paper units (FPU).

The filter paper activity (FPA) of cellulase enzymes was carried out in a mixture containing 0.5 mL diluted enzyme by 50 mM citrate buffer (pH 4.8) and 50 mg of Whatman No. 1 filter paper and incubated at various temperatures for 1 h. CellicCTec2 (Novozymes) enzyme cocktail was used as a control. Apart from the reactions, three blanks were also prepared: Reagent blank (1.5 ml of 50 mM citrate buffer) enzyme blank (1.0 ml of 50 mM citrate buffer + 0.5 ml enzyme dilution) and substrate blank (1.5 ml of 50 mM citrate buffer + filter paper strip). All the blanks were treated identically as the experimental tubes. The reaction was finished placing the tubes on ice. The reducing sugars released were determined using the DNS method. 3 ml of DNS was added to all the tubes and after boiling for 5 min they were placed on ice again to stop the reaction. 0.5 ml of the colored solutions were withdrawn into 1.5-ml microcentrifuge tubes and centrifuged at 10000 g for 3 min. Finally, 2.5 ml of distilled water was added to 0.2 ml of the supernatant and the absorbance was measured at 540 nm, where the absorbance of reagent blank was used as the blank.

In order to determine the released reducing sugars a standard curve was made by means of preparing the following standards: GS1: 1.0 ml of glucose standard + 4.0 ml buffer = 2 mg/ml (1.0 mg/0.5 ml). GS2: 1.0 ml of glucose standard + 2.0 ml buffer = 3.3 mg/ml (1.65 mg/0.5 ml). GS3: 1.0 ml of glucose standard + 1.0 ml buffer = 5 mg/ml (2.5 mg/0.5 ml). GS4: 1.0 ml of glucose standard + 0.5 ml buffer = 6.7 mg/ml (3.35 mg/0.5 ml). Add 0.5
ml of GS1–4 solutions to 13 × 100 mm test tubes, and add 1.0 ml of 0.050 M citrate buffer.

Filter paper unit (FPU) is defined as 0.37 divided by the amount of enzyme that produces 2.0 mg glucose equivalents in 1 h from 50 mg of filter paper. All experiments were carried out in triplicates.

3.2.2.1. Lignocellulosic substrates hydrolysis

The protocol used for this assay was the same that the one for filter paper, the only difference is the substrate. But not only this, we have also added more cellulolytic enzymes such as laccase and xylanase. 50 mg of different lignocellulosic substrates in 50 mM citrate buffer at pH 4.8 were used. Enzyme hydrolysis was performed for 1 hour at 50°C. Endoglucanase alone or in combination with Laccase and Xylanase were used for hydrolysis of the lignocellulosic material. Three different enzyme combinations were used differing in the endoglucanase used: ancestral, *T. maritima* or *T. reesei*. Cellulose degradation was determined by determining percentage of hydrolysis as described elsewhere[148].

3.2.3 Avicel

In this case, a crystalline substrate was used for the cellulolytic activity with mixtures of the free enzymes (0.5 μM each) at 0.5 μM buffer acetate (50 mM final concentration) with 1 % Avicel (FMC, Delaware USA) at various temperatures and pH’s for 24 hours. 0.4ml of the enzymes solutions was placed together with 1.6ml of Avicel solution. Also two blank were done: a substrate blank (1.6ml of Avicel solution + 0.4ml of acetate buffer) and an enzyme blank (1.6ml of acetate buffer + 0.4ml of
Enzymatic reactions were stopped by placing the tubes into an ice-water bath, and the tubes were then be centrifuged for 2 min at 14,000 rpm at room temperature. Enzymatic activity was determined quantitatively by measuring soluble reducing sugars released from the cellulosic substrate by the dinitrosalicylic acid (DNS) method. A volume of 150 μL of the DNS solution was added to 100 μL of sample (supernatant fluids), and after boiling the reaction mixture for 10 min, absorbance at 540 nm was measured. Released sugar concentrations were determined using a glucose standard curve. Glucose concentration was determined using a glucose assay kit [150](GOD; Sigma-Aldrich) according to the manufacturer’s instructions. All assays were performed at least twice in triplicate.

3.2.4. Thermal stability of the ancestral endoglucanase: Circular Dichroism

The thermal stability of the ancestral endoglucanase was determined by Circular dichroism (CD); measurements were made with a JASCO J-815 CD spectrophotometer. For each construct, spectra were generated by averaging five wavelength scans. Thermal unfolding transitions were monitored at 222 nm, with a 0.5°C step size, within the range of 55 to 110°C, in a thermal-resistant 10-mm quartz cuvette. Thermal denaturations at pH 4.8 were carried out in 50 mM citrate buffer both with 0.5M Glycerol and without glycerol [151].

3.2.5. Ancestral endoglucanases Kinetic parameters determination

In order to determine the kinetics parameters of the ancestral endoglucanase, Km and Vmax, ten different substrate concentrations were used in the range of 2 to 20 mg/ml CMC for endoglucanase. The Km and Vmax were determined
directly from the hyperbolic curve fitting of Michaelis-Menten equation generated using Phyton inhouse script. $K_{cat}$ was determined by the formula $V_{\text{max}}/E_t$, where $E_t$ is the total enzyme concentration in $\mu$mol/ml [152].

### 3.3. Cellulosome

#### 3.3.1. Minicellulosome

Two mini-scaffoldins were designed in this study consisting of components from *C. thermocellum* CipA scaffoldin.

##### 3.3.1.1. Minicellulosome constructs

The X-module and type II dockerin dyad and the CBM were amplified from pET28-XDock and pET28-CBM, respectively (a kind gift by Prof. Ed Bayer, Weizmann Institute, Israel). Cohesin 7 was amplified from pAFM-c7A[153]. First, XDock was amplified with primers incorporating NdeI, NheI, KpnI and SpeI sites at the 5’ end and 2 STOP codons and a XhoI site at the 3’ end. The resulting fragment was cloned into pET28 vector using NdeI and XhoI sites. Then the CBM was amplified and cloned into the previous vector using NdeI and NheI sites. Next, cohesin 7 sequence was cloned using KpnI and SpeI sites to generate pET28-Scaf1. A second copy of cohesin 7 was then cloned into this vector in SpeI site to generate pETScaf2, containing 2 tandem cohesins. Both miniscaffoldins carried a hexa-histidine tag at the N-terminus.

Integration of the LFCA endoglucanase into the minicellulosome was accomplished by cloning the LFCA endoglucanase sequence into a pET28a vector between the NcoI and EcoRI sites. Then, the
sequence of *C. thermocellum* Cel8A dockerin (and N-terminal linker) was PCR amplified and cloned at the C-terminus of the LFCA sequence between EcoRI and XhoI sites thus generating pET28-LFCA_Dockerin that carries a C-terminal hexahistidine tag. LFCA_CB is generated by replacing the Cel8A dockerin with a sequence containing the linker between Cel8A catalytic domain and dockerin, followed by the CipA CBM. Both mini-scaffoldins and LFCA endoglucanase fusion proteins were expressed in *E. coli* BL21 star. Expression of miniscaffoldins was carried out at 16°C with 0.1 mM IPTG overnight, while LFCA fusions and Cel8A were expressed at 37°C for 3 hours in 1 mM IPTG. Cultures were lysed by enzymatic means in 1 mg/ml lysozyme, 1% Triton X-100, 5 µg/ml DNAseI and 5 µg/ml RNAse A and centrifuged to remove cell debris. Clarified samples were incubated at 55°C for 20 min, cooled in ice and centrifuged to eliminate aggregated proteins. Affinity purification was then carried out using HisTrap columns in an ÅKTA Purifier FPLC (GE healthcare). Sample purity was evaluated by SDS-PAGE and proteins were concentrated in Tris 50 mM, NaCl 300 mM, CaCl$_2$ 1 mM pH 7, quantified by absorbance at 280 nm with a NanoDrop (ThermoScientific) and stored in 50% glycerol. Mini-cellulosome assembly assays were performed by native-PAGE. Different relations of proteins were incubated in 50 mM Tris, 300 mM NaCl, 1mM CaCl$_2$ pH 7 at 37°C for 1h before running the gel. SdbA cohesin was also added to block XD in the scaffoldin. The true enzyme-scaffoldin ratio was determined from this analysis according to that ratio were no free protein was found in excess. This ratio was used in the following experiments.

Microcrystalline cellulose binding was assayed as described previously [21]. Briefly, 10 µg of protein was incubated with 10 mg of Avicel (SigmaAldrich) at 4°C for 1h with gentle agitation. Samples were centrifuged and the supernatant was stored as the
unbound fraction. The pellet was washed three times and used as the bound fraction. Both samples were then analyzed by SDS-PAGE and BSA was used as a control.

3.3.1.2. Minicellulosome Activity assays

Proteins were incubated in Acetate buffer pH 5.5 containing 100 mM NaCl, 12 mM CaCl$_2$ and 2 mM EDTA for 1 h at 37°C to allow complex formation. Enzymes were used at 0.5 µM (for Avicel and PASC analyses) and at 0.35 µM for CMC assays. Scaffoldins were added at equimolar concentration according to native-PAGE analysis. BSA was added in all samples to minimize unspecific enzyme-substrate interactions. Avicel assays were conducted for 24 h in an orbital shaker in 2-ml tubes containing a wing magnet to improve stirring so that this insoluble substrate did not precipitate. PASC was prepared as described elsewhere [154]. Assays in this substrate were conducted in similar tubes but in a heating block for 30 min. After incubation time, samples were centrifuged and the soluble sugars present in solution in the supernatant were determined by the DNS assay. Absorbance was measured in a 96-well plate using a FLUOstar fluorimeter (BMG Labtech, Germany) in the absorbance mode. CMC assays were conducted in a heating block using azo-CMC (Megazyme) as a substrate. The activity was determined according to the manufacturer’s indications.

3.3.2. Cellulosome

3.3.2.1. Cellulosome design
Plasmid and Primers design for cloning the Ancestral cellulases in the cellulosome system.

3.3.2.2. Cellulosome construction

3.3.2.2.1. Cloning of the new designs in the expression plasmid by PCR

PCRs was performed with Phusion High Fidelity DNA polymerase F530-S (New England Biolabs, Inc), and PCR products and plasmids were restricted with Fastdigest enzymes (Thermo scientific, USA). T4 DNA ligase used for ligation (Fermentas UAB, Vilnius, Lithuania). PCR products were purified using a HiYield™ Gel/PCR Fragments Extraction Kit (Real Biotech Corporation, RBC, Taiwan), and plasmids were extracted using Qiagen miniprep kit (Valencia, CA) (Fig 3.7). Competent *Escherichia coli* XL1 cells were used for plasmid transformation.

**Figure 3.7. Phusion PCR.** Schematic representation of the procedure need for phusion PCR.
3.3.2.2.2. Recombinant protein expression

We produced the recombinant proteins in *E. coli* BL21 (DE3) grown in 2 L LB (Luria Broth) and 2 mM CaCl$_2$ with the appropriate antibiotic at 37 °C until A$_{600}$≈0.8–1 and induced by adding 0.1 mM (final concentration) isopropyl-1-thio-β-D-galactoside (IPTG) (Fermentas UAB Vilnius, Lithuania). Cell growth were left at 16 °C overnight. Cells were harvested by centrifugation at 5000 rpm for 5 min. Pelleted cells were resuspended in 30 mL TBS containing 5 mM imidazole (Tris-buffered saline, 137 mM NaCl, 2.7 mM KCl, 25 mM Tris–HCl, pH 7.4). The His-tagged enzymes will be purified on a Ni–NTA column (Qiagen). Acrylamide gels SDS-PAGE (10 %) were used to assess the purity of the recombinant proteins and absorbance at 280 nm indicated their concentrations. We stored the proteins in 50 % (v/v) glycerol at −20 °C.

3.3.2.2.3. Gel electrophoresis

Each enzymatic component was first calibrated for its optimal ratio for full complex formation with the chimaeric scaffoldin. The three enzymes were then be mixed at their optimized ratio with the scaffoldin to ensure full complex formation. Protein mixtures supplemented with 12 mM CaCl$_2$ and 0.05 % Tween 20 and incubated for 2 h at 37 °C. The electrophoretic mobility of the proteins were then analyzed by PAGE under non-denaturing conditions with gels comprising a 4.3 % stacking gel and a 9 % separation gel. Migration were carried out at 100 V. The gels were stained using InstantBlue Coomassie-based staining (Expedeon, USA).
3.3.2.2.4. Affinity-based ELISA

The specificities of the cohesins for the chimeric dockerin-bearing enzymes were examined semiquantitatively by a sensitive enzyme-linked affinity assay in microtiter plates [155]. MaxiSorp ELISA plates (Nunc A/S, Roskilde, Denmark) were coated overnight at 4°C with predetermined concentrations (designated below) of the desired CBM-Coh(100 ml/well) in 0.1M sodium carbonate (pH 9). The following steps were performed at room temperature with all reagents at a volume of 100 ml/well. The coating solution was discarded and blocking buffer (TBS, 10 mM CaCl2, 0.05% Tween 20, 2% BSA) was added (1 h incubation). The blocking buffer was discarded, and incremental concentrations of the desired EndoDockT or ExoDocG constructs, diluted in blocking buffer, were added. After a 1 h incubation period, the plates were washed three times with wash buffer (blocking buffer without BSA), and the primary antibody preparation was added. Following another 1 h incubation period, the plates were washed three times with wash buffer and the secondary antibody preparation was added. After another 1 h incubation, the plates were again washed (four times) with wash buffer and 100 ml/well TMB β Substrate-Chromogen were added. Color formation was terminated upon addition of 1M H2SO4 (50 ml/well), and the absorbance was measured at 450 nm using a tunable microplate reader.
Chapter 4: Phylogenetic results

In this chapter I will describe the computational procedures utilized for the reconstruction of ancestral forms of cellulase enzymes. I have worked with three different cellulases enzymes, i.e., endoglucanase, exoglucanase and β-glucosidase. I will describe how the sequences for these enzymes are retrieved from internet databases and how they are handled in order to construct a phylogenetic chronogram. This chronogram is the base of the reconstruction process since it provides the overall framework of molecular relationships of all the sequence used in its construction. I will describe the different computational methodologies for tree building based on their statistical basis. These methods include parsimony, maximum likelihood and Bayesian inference. Although I have worked with all of them, I will mainly report the results on Bayesian inference [92-96]. In each phylogenetic tree we can find a series of internal nodes connecting different branches. Each one of these nodes represents the most probable sequence of the common ancestor of the connecting groups of species. I will select one node of each tree
based on different considerations such as age, position in the tree and importance, given the selected group of species. These genes encoding the sequences of the nodes will be synthesize, and reconstructed in the laboratory for testing.

4.1. Reconstruction of an ancestral bacterial endoglucanase

The first cellulase enzyme that we have reconstructed is a bacterial endoglucanase. This enzyme hydrolyzes amorphous and crystalline cellulose by randomly cutting the cellulose fibers giving rise to nanofibers that vary in length all the way from a single glucose molecule to a polymer composed by hundreds of glucose units. The first step in the reconstruction is collecting a number of sequences of endoglucanases from different species. This number varies depending on availability, but generally is recommended to be high enough as to provide a good representation of different phyla. In the case of our endoglucanase we searched in the Uniprot [110] database using BLAST [111] tool as described in the methods section (2.3.1). We searched the endoglucanases from family Cel5A with CAZy [156] identification code 3.2.1.4. This family is important from and industrial point of view, since there are examples of enzymes from this family with outstanding properties used in industrial applications.

I selected as query the Cel5A endoglucanase from Bacillus Subtilis using the default parameters in the Uniprot BLAST website. From this search I retrieved 32 homologous sequences that were available at the time of the search in august of 2014. All the sequences are available in Appendix I including also the Uniprot ID. Most sequences belonged to three different bacterial phyla, i.e., Firmicutes, Actinobacteria and Proteobacteria. The sequences were used to construct a multiple alignment using MUSCLE [115, 116] software. A fragment of this alignment can
Experimental Results

be found in the material and methods section (2.3.2). Upon close inspection of the alignment, we can identify a well aligned block section that corresponds to the catalytic domain of the endoglucanases. This block shows no major gap or unstructured regions. This contrasts with the carbohydrate binding module (CBM) that the sequences show either in the C or N termini. The CBM is poorly aligned demonstrating a molecular diversity that might reflect different origins for this module. For this reason, I decided to focus the analysis on the catalytic domain of the homologous sequences.

Taking the aligned portion corresponding to the catalytic domain, the alignment was manually edited to remove long gaps or portions that are poorly represented in all the sequences. Once the alignment is edited, I used Bayesian inference, to infer the phylogenetic relation between the endoglucanase enzymes. I used BEAST [121, 157] software as described in the methods section. In the final tree, we can see the sequences belonging to the three phyla used in the construction of the chronogram (Actinobacteria, Firmicutes and Proteobacteria) correctly separated. We used fossil and genetic data from the Time Tree Of Life (TTOL) [31, 158] to calibrate the tree. This can be done either as a variable for several nodes in the estimation of the tree or by directly retrieving that age of nodes from the TTOL. In the case of the endoglucanase, I used age data for some nodes from the TTOL as calibration point and estimated the age for all other nodes (see methods section for details). As a result of a calibrated tree, we obtain a chronogram. In our case the mutation rate is not fixed and is uncorrelated for each brand. Thus, we obtain an uncorrelated relaxed clock chronogram that follows a lognormal distribution for node age. The estimated age for the oldest node is more than 3 By (~3000
million years). From the tree, we have identified the common ancestor corresponding to each phylum, i.e, LPCA (Last Proteobacteria Common Ancestor), LACA (Last Actinobacteria Common Ancestor) and LFCA (Last Firmicutes Common Ancestor) (Fig 4.1).

Finally, I used PAML [124, 125] software for the reconstruction of the most probable aminoacid sequence of each node. PAML assigns to each position of the inferred sequence the residue with the highest posterior probability. This sequence is not unique and depends on the sequence used for the tree. However, the phenotypes displayed by the protein or gene of this sequence must be robust and independent of the sequences used. In general, the true value of a reconstructed sequence lies on the phenotype rather than on the genotype. Finally, I selected the sequence of the node corresponding to the LFCA, for synthesis and laboratory resurrection. I chose this particular node because old enough and likely displays phenotypes that can provide information about life and environmental condition of our planet about 3 Bya, such as the high temperature of oceans.
Figure 4.1. Uncorrelated relaxed clock chronogram for bacterial endoglucanases Cel5A. A total of 32 sequences were used from three different phyla, Proteobacteria, Actinobacteria, and Firmicutes. The different species are indicated by UniProt identification codes. Divergence times were estimated using Bayesian inference and information from the TTOL. Geological scale and times are indicated. The internal node corresponding to the Last Firmicutes Common Ancestor (LFCA) was selected for reconstruction.
4.1.1. Ancestral endoglucanase sequence analysis

From the tree in figure 4.1 we selected the node that represents the last common ancestor of Firmicutes (LFCA) that lived ~2.8 Bya for the reconstruction. We speculate that this may have been one of the earliest cellulase enzymes. The ancestral reconstruction used a maximum likelihood \([89-91, 159, 160]\) assignment at each site for the residue with the highest posterior probability. The posterior probabilities of all the aminoacids of the reconstructed sequence are shown in (Fig 4.4). The average posterior probability value is 0.95, which makes reliable the reconstruction.

![Posterior probability distribution for each inferred residue of the ancestral endoglucanase LFCA](image)

**Figure 4.2. Posterior probability distribution for each inferred residue of the ancestral endoglucanase LFCA.** Each position corresponds to the residue with the highest posterior probability. The average posterior probability value is 0.95.

Overall, the ancestral sequence from LFCA displays 71% identity with respect to their modern descendants. This implies that approximately 84 mutations occurred from the ancestral LFCA EG protein to the modern descendant that was used as a query.
4.2.1. Reconstruction of an ancestral bacterial exoglucanase

In the case of exoglucanase, I used the same procedure used in the case of the endoglucanase for the sequence selection; we took the sequences from the Uniprot database. The exoglucanase family I selected in this case was GH48 with CAZy identification code 3.2.1.176; I chose this family, because of its importance in industry. The selected query for the BLAST of the exoglucanase extant sequences was exoglucanase from *Thermobifida fusca*. Using this query, I selected 33 sequences for the alignment (See Appendix II). The sequences we chose, belonged to Actinobacteria and Firmicutes phyla. I generated a sequence alignment using the selected sequences and observed the same phenomena we previously saw for the endoglucanase; the catalytic domains of all of the sequences aligned correctly, forming a similar block of alignment than in the previous case. Nevertheless, it was not the case for the CBM, some sequences had the CBM at the C-termini and others at the N-termini, and there were numerous gaps. Thus, similarly to what we did for the endoglucanase, we made the alignment of the catalytic domain, without the CBM.

We used the sequences of the aligned catalytic domain to construct the phylogenetic chronogram of the exoglucanase CBHI. I inferred the tree using Bayesian inference (Fig 4.2). The root of the chronogram is dated 3 Bya and we could identify the common ancestors of the two phyla present in the tree, LACA (Last Actinobacteria Common Ancestor) and LFCA (Last Firmicutes Common Ancestor). I used PAML for the reconstruction of the most likely aminoacid sequence for each node and I selected the LACA node for the synthesis and laboratory experiments. I selected this exact node because of its
age and position in the tree. The age of the nodes was obtained using information from the TTOL.

Figure 4.3. Uncorrelated relaxed clock chronogram for bacterial exoglucanase GH48. A total of 33 sequences were used from Actinobacteria, and Firmicutes phyla. Divergence times were estimated using Bayesian inference and information from the TTOL[31]. Geological scale and times are indicated. The internal node corresponding to the Last Actinobacteria Common Ancestor (LACA) was selected for reconstruction.
4.2.2. Ancestral exoglucanase sequence analysis

I made the reconstruction of the oldest exoglucanase ancestor that belonged to Actinobacteria phyla (LACA) of about 1300 Bya. We used maximum likelihood for the ancestral reconstruction, obtaining the residue with the highest posterior probability in each position. These posterior probabilities of all the aminoacids of the reconstructed sequence are shown in (Fig 4.5). The obtained average posterior probability value was 0.92, which makes reliable the reconstruction.

![Poster probability distribution for each inferred residue of the ancestral exoglucanase LACA. Each position corresponds to the residue with the highest posterior probability. The average posterior probability value is 0.92.](image)

4.3.1. Reconstruction of an ancestral bacterial β-glucosidase

Regarding to the β-glucosidase chronogram, we used the same procedure described for the two previous enzymes. I retrieved the
sequences for the reconstruction from the Uniprot database. In this case, I selected bgIII family of beta-glucosidases with CAZy identification code 3.2.1.21 as it is common to use enzymes from this family. I selected *Clostridium Thermocellum* as a query. I made the selection of the sequences in the same way of the previous cases, making a BLAST in Uniprot. 34 sequences of Actinobacteria and Firmicutes phyla were chosen for the reconstruction. As well as in the previous cases, we took the sequences for an alignment. The alignment was well-resolved without significant gaps or unstructured portions. The β-glucosidase cellulases do not have binding domains; they do not need this CBM as they are responsible of breaking the small cellobiose units and not crystalline cellulose regions. Due to the lack of this module, the alignment was easily resolved using the whole sequence of each specie for the alignment (All the sequences in appendix III).

I obtained a chronogram that diverged 3 Bya and we could identify the common ancestors of the two phyla present in the tree, LPCA (Last Proteobacteria Common Ancestor) and LFCA (Last Firmicutes Common Ancestor). I made the reconstruction of the most likely aminoacid sequence for each node using PAML and I selected the LACA node for the synthesis and laboratory experiments. The reason why I selected this node is its age and position in the tree. As well as in the other two cases, I determined the age of each nodes doing the datation with the TTOL.
Figure 4.3. Uncorrelated relaxed clock chronogram for bacterial beta-glucosidase BglII. A total of 34 sequences were used from Actinobacteria, and Firmicutes phyla. Divergence times were estimated using Bayesian inference and information from the TTOL. Geological scale and times are indicated. The internal node corresponding to the Last Firmicutes Common Ancestor (LFCA) was selected for reconstruction.
4.3.2. Ancestral beta-glucosidase sequence analysis

Regarding to the beta-glucosidase, we selected the oldest ancestor of the Firmicutes phyla (LFCA) as well as in the endoglucanase case. We used maximum likelihood in this reconstruction too. The posterior probabilities for each aminoacid are represented in the (Fig 4.6). The value of the average of the posterior probabilities is 0.88.

![Figure 4.6. Posterior probability distribution for each inferred residue of the ancestral beta-glucosidase LFCA. Each position corresponds to the residue with the highest posterior probability. The average posterior probability value is 0.88.](image-url)
Chapter 5: Experimental Results

In this chapter, I will describe the experimental results that show the performance and chemical properties of the reconstructed enzymes (endoglucanase, exoglucanase and β-glucosidase), obtained from the chronograms analyzed in Chapter 4. I will first describe the results of experimental synthesis and production of the ancestral enzymes using the standard molecular biology procedures. I will then report on the chemical performance of the resurrected enzymes under different conditions of temperature, pH and substrate. The ancestral enzymes will be compared individually and in a cocktail with extant endoglucanases used in biotechnological applications. Moreover, we tested the activity of the ancestral enzymes in a cellulosome, a bacterial molecular complex that incorporates several enzymes into scaffolding that enhances the performance of individual enzymes.
5.1 Ancestral cellulases production

In order to bring back to life the ancestral cellulases, we asked a company (Invitrogen) to synthesize the reconstructed sequences and we cloned them into an expression vector and expressed in the *E. coli* strain BL21. I carried out the procedure of cloning as I have explained in materials and methods section (3.1.6.). I show the high level of expression of the ancestral cellulases are in **Figure 5.1** SDS/PAGE acrylamide gel. When IPTG was added to induce the protein expression, an overexpressed protein band appeared at a determined molecular weight (endoglucanase ~33 kDa, exoglucanase ~69 kDa and β-glucosidase ~83 kDa). In all the cases this weight coincided with the molecular weight of the ancestral cellulase, which means that the transformation was correct and the bacteria was expressing the desired protein. Furthermore, it can be observed that in addition to the overexpressed band, there were other protein bands that were produced naturally by these bacteria.
Figure 5.1. SDS acrylamide gel electrophoresis analysis of ancestral cellulases expression in *E. coli* bacteria. (a) Endoglucanase expression analysis in 12% acrylamide gel. Lane 1, protein marker (M); lane 2 to 9, colonies 1 to 4 where the protein has been induced with IPTG in the second lane of each colony; lane 10, protein marker (M). (b) Exoglucanase expression analysis in 8% acrylamide gel. Lane 1, protein marker (M); lane 2 to 9, colonies 1 to 4 where the protein has been induced with IPTG in the second lane of each colony. (c) β-glucosidase expression analysis in 8% acrylamide gel. Lane 1 to 2, colony 1; lane 3, protein marker (M); lane 4 to 9, colony 2 to 4. The protein has been induced with IPTG in the second lane of each colony.

The reason behind why the ancestral proteins expressed in such a big amount comparing with the extant ones is unknown, but it seems to be common in ancestral proteins [161]. In Figure 5.2 we can see that the production of the ancestral cellulases is much
higher than that of the extant ones. In addition to running acrylamide gel of the reconstructed ancestral enzymes I run an acrylamide gel comparing the expression of the *T. maritima* endoglucanase and the ancestral endoglucanase LFCA (Fig. 5.2). In this comparison, from a similar bacterial culture amount grown under the same conditions, the intensity of the band corresponding to LFCA endoglucanases is more than twice that of *T. maritima* endoglucanase.

![Image](image)

**Figure 5.2. SDS-PAGE acrylamide 12% gel for the purified enzymes.** On the top panel, LFCA (33 kDa) and on the bottom panel, Endo-β-glucanase from *T. maritima* (37 kDa), with and without IPTG induction. After induction, the expressions level of LFCA EG is more than twice that of *T. maritima*, using the exact same protocol, amounts and expression system.

Furthermore, I run an acrylamide gel of the commercial *T. reesei* mixtures. The results can be seen in Fig 5.3a. In this gel, there is no a single band, as expected from a commercial enzyme preparation, although a major band appears between 54 and 60 kDa [162, 163].
5.2. Mass spectroscopy and protein concentration determination

To further analyze the content of the enzyme preparation from *T. reesei*, we used mass spectrometry. Mass spectrometry (MS) is an analytical technique that ionizes molecules and sorts the ions based on their mass to charge ratio (m/z). For the application of the qualitative analysis of the enzymatic profiles, the mass analyzer that has the most impact is the time of flight (TOF). It is the most suitable analyzer in this type of studies due to its sensitivity, speed in full scan mode, high resolution and ability to identify unknown compounds. It is possible because TOF presents the possibility of measuring the exact mass of the detected ions using isotope distribution (True Isotopic Pattern, TIP). The time-of-flight analyzer discriminates according to the speed difference acquired by the ions inside a flight tube of known length as a function of its m/z ratio. TOF analyzer is based on the fact that all the ions generated in the ionization source have the same kinetic energy, their speed being inversely proportional to the square of their mass. A voltage (pulse) is applied to them to accelerate the ions, throwing them to a tube under high vacuum with a constant kinetic energy. Ions that have the same kinetic energy but different values of m/z will have different velocity. The higher m/z ions will travel at a slower speed than the lower m/z ions arriving later at the detector.

For the determination of the concentration of *T. reesei* enzyme, we have used two methodologies described in details in the material and methods section. The first method used was Pierce BCA Protein Assay Kit [139] using a BSA standard supplied with the kit and a standard of our ancestral endoglucanase and the second methods was the Difference Dry Weight method [138]. These values have been also contrasted directly with other
researchers that use the same preparation [164]. We obtain a similar concentration that previously reported, around 125 mg/mL.

Figure 5.3. Endo-β-glucanase content determination for T.reesei preparation. (a) present, the major component endo-β-glucanase represents about 70% of the total mix, as determined by gel densitometry. (b) Mass spectrum of T.reesei preparation in which endo-β-glucanase at 54 kDa represents also about 70 %, in close agreement with gel in (a). SDS-PAGE page acrylamide 8 % gel of T.reesei preparation. The main band represents the endoglucanase of Trichoderma reesei (~54 kDa). Although several enzymes are present, the major component endo-β-glucanase represents about 70% of the total mix, as determined by gel densitometry.
5.3. Endoglucanase Assays

We carried out several assays in order to test the activity of the ancestral endoglucanase, first of all, a study of the specific activity was done. We measured the activity of the ancestral endoglucanase against two extant endoglucanases in a range of temperatures and pH values as it is shown in Figure 5.4.

![Graphs showing activity assays for endoglucanase enzymes.](image)

**Figure 5.4. Activity assays for endoglucanase enzymes.** (a) Specific activity as a function of temperature for LFCA, *T. maritima*, and *T. reesei* cellulases at pH 4.8. (b) Specific activity as a function of pH (4-10) for LFCA, *T. maritima*, and *T. reesei* cellulases at 70°C.

The first assay that we performed was the determination of the specific activity in a range of temperatures, from 30 to 90 °C. This range is broader than the range generally used in industrial applications that does not go above 60 °C. It is also broader than the typical range presented in the literature for improved endoglucanase testing, which is about 40-80 °C [165, 166].
In the assay, I used a common soluble substrate such as carboxymethyl cellulose (CMC) [167]. As mentioned, I performed the assay at different temperatures and pH 4.8, incubating the substrate enzyme mix for 30 min and measured the reducing sugar concentration with the DNS method [168]. We determined the activity of the enzyme spectrophotometrically measuring the absorbance at 540 nm, as the amount of reducing sugar in mg released per min and per mg of enzyme used (see Materials and Methods). We show elevated activity at up to 90°C in the presence of 5% glycerol, which is generally used as stabilizer (Fig 1b). This working temperature is very high for endoglucanase, which generally operate in the range 40-60 °C.

The same measurements were performed in a broad pH range at 70°C. Figure 5.4 shows that the ancestral endoglucanase has higher specific activity with soluble CMC and 5% glycerol than bacterial and fungal cellulases at all temperatures tested. It also outperforms the assayed extant enzymes in all the range of pH. In Figure 5.5 we show the same assays repeated at other temperatures and pH values.
**Experimental Results**

Figure 5.5 Endoglucanase activity measurements using CMC/DNS method. (a) Specific activity assay at 30 °C as a function of pH (4-12) for LFCA, T. maritima and T. reesei cellulases. We determined the reducing sugar mg equivalent released per minute and per mg of enzyme. (b) Specific activity assay at 50 °C as a function of pH (4-12) for LFCA, T. maritima and T. reesei cellulases. (c) Specific activity assay in a range of temperatures (30-90 °C) at pH 10 for LFCA, T. maritima and T. reesei cellulases. All assays were triplicated. Values are reported as average ±S.D.

In figure 5.5 three assays are shown, in the first one (Fig 5.5a) I show the activity assay I carried out at 30°C; I run the same assay also at 50°C (Fig 5.5b). In both cases, the ancestral endoglucanases outperforms the two commercial endoglucanases. The other assay that is shown is the one corresponding to pH 10 (Fig 5.5c) that I run in a range the temperatures 30-90°C. As well as in the previous assays, at pH of 10 the endoglucanase shows a higher activity than the commercial endoglucanases from *T.maritima* and *T.reesei* in all the tested temperatures. The methodology used in those assays was the CMC method, the same described for Figure 5.4. In those cases, we see the exactly the same than in Figure 5.4, the ancestral endoglucanase shows a higher specific activity in all the tested conditions.
Another method available for the measurement of the specific activity of an endoglucanase is the use of a specific kit for endoglucanase activity. In this case (Fig 5.6) a CellG3 kit (Megazyme) [169] was used for the determination of the endoglucanase specific activity. I measured the specific activity of the endoglucanases at pH 5 and a range of temperatures (30-90ºC). In this case, I measured the activity of the ancestral endoglucanase against the endoglucanase of Thermotoga Maritima. The preparation T.reesei was not used in this case, as it is a mixture of enzymes and not a purified single enzyme. In all the temperatures I tested, the activity of the ancestral endoglucanase was much higher than the T.maritima, specially from 50ºC to 70ºC, where the difference is even higher.

![Figure 5.6. Endoglucanase activity measurements using CellG3 method](image)

*Figure 5.6. Endoglucanase activity measurements using CellG3 method. Activity in CellG3 U/mg of two different endoglucanases: LFCA; Thermotoga maritima, measured in a range of temperatures (30-90ºC) at pH 5 with Megazyme endo-cellulase assay kit [4]. Both enzymes show similar profile but LFCA displays higher specific activity. All assays were triplicated. Values are reported as average ±S.D.*
The next property we evaluated for those endoglucanases was the long term activity. We define the long term activity as the reducing sugar production per minute. The hydrolysis rate was determined by measuring the activity at 60 °C and reaction times ranging from 10 to 250 min as shown in Figure 5.7, the ancestral enzyme reached over 90% reducing sugar production within 10 min.

![Figure 5.7: Long-time activity measurements. Reducing sugars production was measured at different reaction times at 60 °C. Experimental details are provided in the Materials and Methods. In each case, three replicates were collected. The average ± S. D. values are shown for each measurement.](image)

We used exponential fits to determine a rate of 0.24 µg of reducing sugar/min for ancestral endoglucanase, 0.032 for T. maritima and 0.069 for T. reesei.

Another thing we determined when characterizing the enzyme was the stability of the ancestral endoglucanase to temperature
incubation compared to *T. maritima*, and *T. reesei* cellulases as Figure 5.8 shows. By means of this graphic representation, we determined the temperature at which they lost half of their activity after 30 min of incubation and then ascertaining their residual activity at 60°C. We estimated a value of 85°C for *T. maritima*, 79°C for LFCA EG, and 72°C for *T. reesei*. In the case of the T.reesei preparation, it shows a biphasic-like decay, which we hypothesize that could be a consequence of the action of other enzymes in the preparation (Fig 5.8).

**Figure 5.8. Pre-incubation experiments at different temperatures for the different endoglucanases conducted for 30 min.** Residual activity was determined on 2% CMC for 30 min at 60°C using DNS. Relative activity is determined for each individual enzyme. Each enzyme was pre-incubated at its best performing pH value.

Another relevant characteristic that we determined were the kinetics for the thermal inactivation of the enzymes at 80°C. In
order to study this, the residual activity (the remaining activity, after the loss of activity with the incubations) was plotted against the time, showing that it follows a clear first order kinetics for *T. maritima* and LFCA EG (Fig 5.9). In the case of the preparation of *T. reesei* we observe a biphasic behavior, the reason for this behavior was not clear, although we suspect that perhaps it could be an effect of the different enzymes present in the preparation.

**Figure 5.9. Endoglucanase inactivation at 80°C.** (a) Residual activities were measure at different incubation times. (b) Specific activity was determined after incubation. The activity of non-incubated enzyme was used as a reference for 100% residual activity. Each assay was repeated five times. The values are presented as average ±S.D.

From the plot (Fig 5.9) I determined, the inactivation constant ($K_{in}$) and half-life ($t_{1/2}$) [2], we used a linear fit for the determination. The values obtained in each case are shown in Table 5.1.
Table 5.1. Half-life and inactivation rate constant (K_{in}) of LFCA, T. maritima and T. reesei. In the case of T.reesei the biphasic behavior observed could reflect the action of more than one enzyme in the preparation T.reesei.

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Half-life (min)</th>
<th>Rate constant K_{in} (min^{-1})</th>
</tr>
</thead>
<tbody>
<tr>
<td>LFCA</td>
<td>56</td>
<td>0.29</td>
</tr>
<tr>
<td>T. maritima</td>
<td>178</td>
<td>0.10</td>
</tr>
<tr>
<td>T. reesei</td>
<td>9</td>
<td>2.26</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.30</td>
</tr>
</tbody>
</table>

The values I obtained for the three enzymes are in consequence with what it was expected. The half-life for T. maritima and LFCA cellulases at 80°C was 178 and 56 min, respectively (Table 5.1). The t_{1/2} value for T. maritima was high, as expected for an extreme thermophile, although it still displays lower specific activity than the ancestral LFCA endoglucanase. However, for T. reesei this value is smaller, 9 min. That means that it lost more than 50% of its activity within the first 10 min of incubation, with a slower decay after that, which again likely reflects the action of more than one enzyme.

Regading to the inactivation constant value (K_{in}) we calculated a K_{in} of 0.29 min^{-1} for LFCA EG, 0.10 min^{-1} for T. maritima, and 2.26 min^{-1} and 0.30 min^{-1} for the fast and slow phases of T. reesei, respectively.

The experiments described since now, have been performed using CMC, a soluble laboratory substrate useful for determining endoglucanase activity. However, cellulases must be able to
hydrolyze cellulose in lignocellulosic materials, such as agricultural, industrial, or city waste. To test this, we used cardboard as a source of cellulose. In cardboard, cellulose, lignin, and hemicellulose are present at approximately 60%, 15%, and 15%, respectively. We performed activity assays using isolated LFCA endoglucanase and in combination with laccase and xylanase, enzymes that can degrade lignin and hemicellulose, respectively. We determined the percentage of cellulose hydrolyzed in a 50 mg sample of cardboard within 1 h at 50°C and pH 4.8. As shown in Figure 5.10, the endoglucanase enzymes degraded very small amounts of cellulose on their own, no more than ~5%, with the commercial T. reesei being slightly more efficient, probably due to the action of other enzymes in the preparation.
5.3.1. Lignocellulosic substrates hydrolysis

**Figure 5.10. Hydrolysis of cardboard lignocellulosic material.** We used 50 mg of milled lignocellulosic material (cardboard, newspaper and pine softwood) in 50 mM citrate buffer at pH 4.8. Enzyme hydrolysis was performed for 1 hour at 50°C. Endoglucanase alone or in combination with laccase and xylanase were used for hydrolysis of the lignocellulosic material. Three different enzyme combinations were used differing in the endoglucanase used: ancestral, T. maritima or T. reesei. Released sugars are quantified with the DNS method. Cellulose hydrolysis yield was determined as described elsewhere [148, 170].
Experimental Results

The experiments described above were carried out using CMC, a soluble laboratory substrate useful for determining endoglucanase activity. However, for industrial applications, cellulases must be able to hydrolyze cellulose in lignocellulosic materials, such as agricultural, industrial, or city wastes in synergy with other enzymes such as lacasse and xylanase. This is important, for instance, for the pretreatment of lignocellusic biomass using enzymes. To test this aspect, we used cardboard, newspaper and softwood from pine tree as a source of cellulose. These three materials have different content of cellulose, kinging and hemicellulose. While cardboard contains around 60% cellulose and around 15% of lignin and hemicellulose, newspaper and pine softwood contain less cellulose, less than 50% [171-173] and more lignin, ~22 and ~30, respectively; and hemicellulose, ~18 and ~25, respectively. We performed activity assays using isolated LFCA EG and in combination with laccase from *Trametes pubescens* and xylanase from *Thricoderma viride*, enzymes that can break down lignin and hemicellulose, respectively. We determined the percentage of cellulose hydrolyzed in a 50 mg sample of lignocellulosic material [171-173], within 1 h at 50°C and pH 4.8. In the case of cardboard, the three EG enzymes degraded very small amounts of cellulose on their own, no more than ~5%, with the commercial *T. reesei* endoglucanase being slightly more efficient, probably due to the action of other enzymes in the preparation (Fig 5.10a). Conversely, LFCA EG worked best when used synergistically with laccase and xylanase hydrolyzing approximately half the cellulose present in the sample, as compared to *T. reesei* that degraded ~16% and *T. maritima* that degraded less than 10%. In the case of newspaper and softwood, the lower amount of cellulose and higher content of lignin is reflected in the lower efficiency of cellulose degradation, although still LFCA EG outperforms the other endoglucanases (Fig 5.10a, and 5.10b).
These results highlight not only the potential of LFCA EG to work with lignocellulosic substrates, but also the advantage of using multi-enzyme cocktails containing cellulases, laccases, xylanases and other enzymes for efficient pretreatment of raw materials and subsequent hydrolysis of cellulose.

As the graph in Fig 5.10 shows, LFCA endoglucanase worked best used synergistically with laccase and xylanase and could hydrolyze approximately half the cellulose present in the sample, as compared to T. reesei that degraded ~16% and T. maritima that degraded less than 10%. These results show, the potential of the ancestral endoglucanase to work with lignocellulosic substrates and the advantage of using multi-enzyme cocktails containing cellulases, laccases, xylanases and other enzymes for efficient pretreatment of raw materials and hydrolysis of cellulose on the other. This synergy is really interesting under an industrial point of view, as using enzymes able to hydrolyze different substrates we can hydrolyze complex lignocellulosic materials.

5.3.2. Thermal stability of the ancestral endoglucanase:
Circular Dichroism

In order to determine the thermal stability of the ancestral endoglucanase, we performed a circular dichroism experiment. (Fig 5.11). We examined the thermal denaturation of the ancestral endoglucanase versus the ancestral endoglucanase with 0.5% glycerol shifting the melting point. Figure 5.11 shows elipticity indicating thermal denaturation transitions of endoglucanase (in blue) and the endoglucanase with 0.5% of glycerol (in green) at pH 4.8 from 55°C to 110°C. The circular dichroism at 222 nm should report transition midpoint temperatures (Tm) are reported
in Table 5.2. The signal was fitted to a two stated model using baselines for the unfolded and folded states and with the thermodynamics states adjusted by using the modified Gibbs-Helmholtz equation [174] from where we determined the thermodynamic values of enthalpy, entropy and the heat capacity.

![Figure 5.11](image)

**Figure 5.11. Thermal denaturation of ancestral reconstructed endoglucanase.** In blue the thermal denaturation of endoglucanase at pH 4.8 and 50mM citrate buffer from 55 to 110°C. In green, the thermal denaturation of endoglucanase at pH 4.8 and 50mM citrate buffer from 55 to 110°C in presence of 0.5% glycerol.

In Table 5.2 the values for the denaturing temperature and the thermodynamic constants are represented.
<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Endoglucanase</th>
<th>Endoglucanase+ %0.5 Glycerol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tm (°C)</td>
<td>80.3</td>
<td>83.3</td>
</tr>
<tr>
<td>ΔH (kJ/mol)</td>
<td>592</td>
<td>428</td>
</tr>
<tr>
<td>ΔS (kJ/mol/K)</td>
<td>1.68</td>
<td>1.20</td>
</tr>
<tr>
<td>Cp (kJ/mol/K)</td>
<td>9.8*10^{-6}</td>
<td>5.5*10^{-5}</td>
</tr>
</tbody>
</table>

Table 5.2. Thermodynamic constants of reconstructed ancestral endoglucanase

The Tm values (Table 5.2) show that the glycerol has a stabilizer effect as it was expected. By adding glycerol the Tm of the protein increased in 3°C. Besides, it is a high value of Tm for this type of proteins that usually have values around 60-75°C [175] obtaining similar enthalpy and entropy values [176]. We observe a marginal stabilization when using glycerol in the free energy and in the melting temperature with an increasing of 3°C.

5.3.3. Ancestral endoglucanases Kinetic parameters determination

We further studied the reaction processes and catalytic events for the ancestral endoglucanase by determining $K_M$, a measure of substrate affinity, and $K_{cat}/K_M$, a measure of catalytic efficiency. In order to determine these kinetic parameters, Michaelis-Menten equation [177].

$$v = \frac{d[P]}{dT} = \frac{V_{max} [S]}{K_M + [S]}$$

Where Vmax represents the maximum rate achieved at saturating substrate concentration and $K_M$ (Michaelis constant) is the substrate concentration at which the reaction rate is half of the Vmax.
Michaelis–Menten saturation curve for an enzyme reaction shows as it is shown in (Figure 5.12a) the relation between the substrate concentration and reaction rate [178].

**Figure 5.12.** Hanes–Woolf plot and Lineweaver–Burk plot for reconstructed ancestral endoglucanase.

Results obtained for endoglucanase activity are shown in (Table 5.3).
Table 5.3. Kinetic constants of ancestral endoglucanase

<table>
<thead>
<tr>
<th>Kinetic constants</th>
<th>Endoglucanase</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_M (mM)$</td>
<td>0.007</td>
</tr>
<tr>
<td>$V_{max} (mM/s)$</td>
<td>1.96</td>
</tr>
<tr>
<td>$K_{cat} (s^{-1})$</td>
<td>217</td>
</tr>
<tr>
<td>$K_{cat}/K_M (mM/s)$</td>
<td>2.4*10^4</td>
</tr>
</tbody>
</table>

From the kinetic values determined from the plots, we observe that the enzyme has a high substrate affinity ($K_M$) and the measured catalytic efficiency is also high ($K_{cat}/K_M$) in comparison with the constants measured in the literature for endoglucanases [179] [180, 181].

5.4. Enzyme Cocktail Assays

After studying the activity of the ancestral endoglucanase against some extant ones, I developed, a study of the ancestral cocktail against a commercial one, as it is explained in the next points of this work. The use of cellulase cocktails is really important in order to completely hydrolyze the cellulose into sugar monomers. Moreover, there is a huge industrial interest in the degradation of lignocellulosic materials that is why; we have included a laccase in our studies.

5.4.1. Ancestral Enzymes Cocktail

The first assays I carried out for the ancestral cocktail are shown in Figure 5.13. In this assays, I measured, the specific activity of the ancestral cocktail against a commercial one Ctec2 cocktail (Fig 5.13). We run the assay in three different temperatures 50°C (Fig 5.13a), 60°C (Fig 5.13b) and 70°C (Fig 5.13c) and in three different pH values (5,7 and 11) in each case, As the figure
Experimental Results

shows, the ancestral cocktail outperforms the commercial cocktail Ctec2 in all the cases (Fig 5.13). This significant difference in the activity is even bigger in the case of the neutral and the basic pH (Fig 5.13b and 5.13c), where the activity of the commercial cocktail is really small.

![Figure 5.13](image)

*Figure 5.13. Specific activity as a function of pH for ancestral enzyme cocktail and commercial enzyme cocktail Ctec2. a) 50°C, b) 60°C, and c) 70°C. Hydrolysis was carried out for 1 h using filter-paper as a substrate. All assays were triplicated. Values are reported as average ±S.D.*

Once I measured, the good performance of the ancestral cocktail in comparison with the commercial one Ctec2, I carried out a study of the stability. The study was developed in the same way we did in the case of the endoglucanases in section 5.3. For this purpose I performed several assays.

I determined the long term activity of both cocktails (ancestral cocktail and Ctec2 commercial cocktail) measuring the activity at 60 °C and reaction times ranging from 10 to 250 min as shown in Figure 5.14. In this figure, we see that the commercial cocktail Ctec2 reached almost the %100 of the reducing sugar production in a short time comparing with the ancestral cocktail. The plot shows (Fig 5.14) that the commercial cocktail reaches the %100
of its reducing sugar production in 50 minutes, in contrast, the ancestral one needs 250 minutes to reach it. However, the production of reducing sugars of the ancestral cocktail is higher than the commercial one from the very beginning.

Figure 5.14. Long-time activity measurements for ancestral and commercial enzyme cocktail Ctec2. Reducing sugars production was measured at different reaction times at 60 °C. Experimental details are provided in the Materials and Methods. In each case, three replicates were collected. The average ± S. D. values are shown for each measurement.

From this plot (Fig 5.14) we calculated the hydrolysis rate of both cocktails. We obtained a value of 0.14µg of sugar per minute in the case of the ancestral cocktail and a rate of 0.067µg of sugar per minute for the commercial one Ctec2.

In addition, I also evaluated the stability for temperature incubation. This study is shown in Figure 5.15 making the
comparison of the stability of the ancestral cocktail versus the commercial one Ctec2. By means of this graphic representation, we determined the temperature at which they lost half of their activity after 30 min of incubation, making after the measurement of the activity at 60°C. The estimated values of temperature were the commercial Ctec2 and the ancestral cocktail lose half of their activity were 73°C and 76°C for each case (Fig 5.15).

**Figure 5.15. Pre-incubation experiments for ancestral and commercial enzyme cocktail Ctec2 at different temperatures conducted for 30 min.** Residual activity was determined on 2% CMC for 30 min at 60°C using DNS. Relative activity is determined for each individual enzyme. Each enzyme was pre-incubated at its best performing pH value. All assays were triplicated. Values are reported as average ±S.D.

Continuing with the thermal stability of the enzyme cocktails, I determined the kinetics for the thermal inactivation of the enzymes at 80°C. The residual activity plotted against the time
followed a clear first order kinetics for both ancestral and commercial cocktail Ctec2, it can be seen in Figure 5.16.

Figure 5.16. Ancestral and commercial enzyme cocktail Ctec2 inactivation at 80°C. (a) Residual activities were measured at different incubation times. (b) Specific activity was determined after incubation. The activity of non-incubated enzyme was used as a reference for 100% residual activity. Each assay was repeated five times. The values are presented as an average ±S.D. All assays were triplicated. Values are reported as average ±S.D.

I calculated the inactivation constant ($K_{in}$) and half-life ($t_{1/2}$) from the plot, for what I used the same procedure used in point 5.2. The half-life for ancestral and commercial cocktails Ctec2 at 80°C was 55 and 13 min, respectively (Table 5.2).
Experimental Results

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Half-life (min)</th>
<th>Rate constant $K_{in}$ (min$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ancestral cocktail</td>
<td>55.6</td>
<td>0.39</td>
</tr>
<tr>
<td>Ctec2</td>
<td>12.7</td>
<td>1.72</td>
</tr>
</tbody>
</table>

*Table 5.2. Half-life and inactivation rate constant ($K_{in}$) of ancestral and commercial enzyme cocktails Ctec2.*

The values obtained in this case are similar to the previous ones, what makes sense. In the case of the commercial cocktail, we obtained a slightly higher value than for *T. reesei* but similar in magnitude. In the case of the ancestral cocktail is almost the same we measured for the ancestral endoglucanase. Regarding to $K_{in}$, the values also are similar that we obtained previously, 0.39 in the case of ancestral cocktail and 1.72 in the case of the commercial one.

We carried out all the experiments described until now in 5.3 point, using filter-paper, a common substrate very used for the determination of the activity of cellulase cocktails. Nevertheless, there is a huge interest in the hydrolyzation of crystalline cellulose in industry [182]. That is why, we decided to test the activity of the ancestral cocktail using a completely crystalline substrate as it is the case of Avicel. For this porpoise, we carried out the assay as shown in *Figure 5.17* at different temperatures (40-70ºC) in three different pH values (5.7 and 10). The incubation for this assay was performed for 24 hours in agitation. The figure clearly shows (*Fig 5.17*) that the specific activity of the ancestral cocktail is also higher in this case. This difference becomes even bigger when the temperature and the pH values are higher. So, this assay shows that the ancestral cocktail is not only
able to degrade amorphous cellulose in a better way, but also its activity is higher with crystalline cellulose. Although this are really interesting results, both filter-paper and avicel are %100 cellulose containing substrates and there is a need of hydrolyzing not only cellulosic substrates but also lignocellulosic materials as it was mentioned before. In this direction, there are several assays we performed in point 5.3.2 for those lignocellulosic materials used also before for the endoglucanase.

Figure 5.17. Ancestral and commercial enzyme cocktail Ctec2 activity measurements using avicel. Specific activity assay at pH (5, 7, 10) for ancestral and commercial cocktail at different temperatures. a) 40°C, b) 50°C, c) 60°C and d) 70°C. We determined the reducing sugar mg equivalent released per minute and per mg of enzyme. All assays were triplicated. Values are reported as average ±S.D.
5.4.2. Lignocellulosic substrates hydrolysis

We carried out activity assays with three different substrates (cardboard, newspaper and wrapping paper), values we obtained at 50 °C and pH 4.8 can be seen in Figure 5.18. These values show that the ancestral cocktail has higher activity in all the substrates when we added laccase to the cocktail. As laccase degrades lignin, cellulose was more accessible for the cellulases present in the cocktail. In these conditions, where cellulose was more accessible, we measured the higher specific activity of the ancestral cocktail, almost doubling the activity of the commercial cocktail Ctec2. We expect that the commercial cocktail Ctec2 contains other enzymes such as xylanases and laccase in addition to the cellulases that favor the degradation. This can be the reason for the observed lower activities of the ancestral cocktail in absence of laccase.

![Figure 5.18. Activities of ancestral enzyme cocktail (CKA), commercial enzyme cocktail (CTec2), ancestral enzyme cocktail in presence of T. pubescens laccase (CKA + L) and commercial enzyme cocktail in presence of T. pubescens laccase (Ctec2 + L) at 50 °C and pH 4.8. Assays were carried out in three different substrates: cardboard, newspaper and wrapping paper.](image)
It can be also observed that the activity is substrate dependent. Substrates have different cellulose, hemicellulose and lignin content [183]. In addition, they went through different mechanical and chemical processes in their production. This may define the arrangement of the lignocellulosic fibers and therefore, as well as being different in composition they are different in structure. This diversity of substrate’s characteristics may affect the ability of enzymes to reach their specific substrate and degrade it.

Comparing these results with the ones obtained in filter-paper ideal substrate (Fig 5.13.), lower values are obtained in lignocellulosic substrates. Filter-paper is pure cellulose that is synthetized in the laboratory. In contrast, lignocellulosic biomass is a complex substrate due to its structure and composition. Lignin and hemicellulose form a protective shell around cellulose, which obstructs enzymatic attack and thus, lower amount of glucose is released.

We repeated those assays in different temperatures; we plotted in (Fig 5.19) the activity values obtained at a temperature range of 50-70 ºC. In all the tested substrates, the highest activity was obtained at 50 ºC with the ancestral cellulases cocktail together with the laccase. As we saw before in Figure 5.18, the laccase influence was really positive and the activity of ancestral cellulases was higher than using the commercial ones Ctec2. Regarding temperature, its increment resulted in lower activities. At 70 ºC similar values were observed when the assay was carried out with or without laccase. This may happen due to the fact that laccase is not active at that temperature. If laccase losses its activity the accessibility of cellulose is reduced and thus, less cellulose is degraded to glucose.
Experimental Results

Figure 5.19. Activities of ancestral enzyme cocktail (CKA), commercial enzyme cocktail (CTec2), ancestral cocktail in presence of T. pubescens laccase (CKA + L) and commercial enzyme cocktail in presence of T. pubescens laccase (CTec2 + L) at 50-70 ºC temperature range and pH 4.8. Assays were carried out in three different substrates: cardboard (a), newspaper (b) and wrapping paper (c).

As mentioned before, the commercial cellulases cocktail Ctec2 contains other enzymes that favor the hydrolysis. Thereby, in some cases lower activities of the ancestral cocktail were observed when laccase was not added.
5.5. Minicellulosome

In this part of the thesis, we constructed different complexes with the ancestral endoglucanase. This work was carried out in collaboration with Mariano Carrión Vázquez group in the Cajal Institute (CNIC).

Apart from testing the activity of the ancestral endoglucanase as free enzyme, another interesting aspect was testing the performance of LFCA endoglucanase incorporated into a cellulosome, which is a macromolecular complex containing several lignocellulose-degrading enzymes attached to scaffoldin via dockerin protein domains. There are some organisms of cellulolytic bacteria such as Clostridium thermocellum that use the cellulosome to degrade cellulose in nature. The use of it has been suggested for industrial applications, due to the increased activity [184]. For this reason, we made different constructs fusing endoglucanase enzymes to domains present in the cellulosome. As it can be seen in Figure 5.20 we bonded dockerin at the C-terminus of the ancestral endoglucanase (LFCA-Dock) to allow its incorporation into a mini-scaffoldin containing a single (Scaf1) or two tandem (Scaf2) cohesins. In this Figure 5.20 there is a scheme of the different domains separately and also a schematic representation of the created complexes. Two different controls were used LFCA endoglucanase fused to a cellulose binding module (LFCA-CBM) and C. thermocellum Cel8A endoglucanase (CtCel8A), a major endoglucanase in its cellulosome [185].
5.5.1. Minicellulosome construction

![Diagram of cellulosome construction](image)

**Figure 5.20. Cellulosome construction.** Schematic representation of the incorporation of ancestral endoglucanase (LFCA) into a mini-cellulosome. The different molecular elements assembled are represented.

LFCA-Dock incorporation into two mini-scaffoldins occurred at molar ratios of 1.1:1 (LFCA-Dock:Scaf1) and 2:1 (LFCA-Dock:Scaf2), which was close to the expected ratio since cohesin-dockerin binding occurs in a 1:1 ratio[184], indicating precise complex formation (Fig 5.22a). Furthermore, LFCA-Dock incorporated into the cellulosome and LFCA-CBM was capable
of binding microcrystalline cellulose (Fig 5.21), while the other proteins were not. This indicates that, as expected, only when a CBM was present, specific microcrystalline cellulose binding could occur.

**Figure 5.21. Controls of cellulose interaction of LFCA and variants.** Only LFCA-CBM and LFCA-Dock+Scaf1 bind microcrystalline cellulose (Avicel) and are found in the precipitated fraction (P) while other variants and the control protein BSA are found in the soluble (S) fraction. BSA was added to all samples to minimize nonspecific interactions.

5.5.2. Minicellulosome activity assays

We studied the effect of the incorporation of the ancestral endoglucanase into the cellulosome carrying out activity assays with Avicel. Taking into account that this is a microcrystalline cellulose substrate this is targeted by the CBM used (Fig 5.22b). These assays were run at 70ºC. We chose this temperature taking into account that no major loss of activity was expected to happen during the long incubation time needed. Moreover, this was the temperature for which the highest activity was observed in the
CellG3 assay. Free LFCA endoglucanase showed a higher activity with this substrate than native CtCel8A but dockerin incorporation into LFCA resulted in a lower activity than that of the original LFCA endoglucanase, but it was still slightly higher than that of CtCel8A (Fig 5.23c). Importantly, when LFCA-Dock was incorporated into Scaf1, the resulting activity was extraordinarily improved.

**Figure 5.22. Cellulosome activity assays.** (a) Native-PAGE shows that a new band appears upon incubation of LFCA-Dockerin and a mini-scaffoldin, indicating complex formation. Activity of the free and mini-cellulosome bound LFCA on Avicel (b), PASC (d), and CMC (e). Each experiment was carried out in triplicate and the average ± S. D. values are shown.
Concerning to the complex LFCA-CBM, we measured a similar activity than that of LFCA-Dock. This made us to think that the enhancement was due to a substrate targeting effect. Incorporation into Scaf2, whereby two tandem identical cohesins allow for the formation of a cellulosome with two enzymes, did not provide further activity improvement in either case, for LFCA endoglucanase and for CtCel8A. However, this result does not mean that further synergy could if we use different enzymes together with LFCA EG. Similar results were observed at all of the tested pH values (Fig 5.23) and at lower temperatures (Fig 5.24). At higher temperatures above 80 °C, though, the situation was reversed and CtCel8A showed higher activities (Fig 5.24), perhaps due to the long reaction times.
Experimental Results

Figure 5.23. Activity of LFCA as a function of BSA concentration and pH in Avicel. (a) Dependence of LFCA activity on the BSA concentration. (b) The activity of LFCA-variants and (c) CtCel8A at different pH values. Both assays were performed at 50°C using Avicel as substrate in triplicate. Values show average ± SD.
Figure 5.24. Temperature dependence of LFCA and CtCel8A enzymes free and integrated into a mini-scaffold in different substrates. Avicel (a,b), PASC (c,d). Assays were done in triplicate. Values are shown as average ± SD.

5.6. Cellulosome

I developed this work during a short stay in Pf Edward Bayer’s laboratory in the Weizmann Institute of Science as part of a collaboration with his group. The aim was to learn the designer cellulosome techniques and the implantation of our ancestral enzymes in their designer cellulosome. The intention was to redesign and create a new cellulosome composed of ancestral
Experimental Results

cellulases (endoglucanase, exoglucanase and beta-glucosidase) as it is shown in Figure 5.25.

Figure 5.25. Schematic representation of the designed ancestral cellulosomes. In the legend can be seen which was the meaning of each component. One of the designs was done to assemble the ancestral endoglucanase and the ancestral exoglucanase enzymes (the one on the top). The other one was designed in order to include the three ancestral enzymes (endoglucanase, exoglucanase and beta-glucosidase). The CBM chosen for those cellulosomes was the one from Clostridium Thermocellum.
5.6.1. Recombinant protein expression

In order to produce designer cellulosomes of the ancestral enzymes, first of all, I carried out the production of both the scaffolds and the previously designed enzyme-dockerin complexes. The production was made as explained in material and methods. Once the production was done, I run acrylamide gels for each recombinant protein as shown in Figure 5.26.

\[ \text{Figure 5.26. Screening of the recombinant proteins.} \]

In those gels, the needed proteins and scaffolds can be seen in their correct sizes. After having all the proteins and scaffolds produced, I did the complexation for the designer cellulosome. For this porpoise, I used the previously described (material and methods) protocol.

5.6.2. Cellulosome construction

I tested the complexation running native gels of each enzyme with its corresponding dockerin and scaffolding. The one for beta-glucosidase was the first one and as Figure 5.27 shows, I obtained a good complexation.
We need to test the enzyme/scaffold ratios used in Figure 5.27, in order to choose the most appropriate one. In this case, ratio 1 would be the best one as few enzyme and scaffold is left and a big quantity of the complex is formed, as it is shown in the gel (Fig 5.27).

In the case of the endoglucanase, Figure 5.28 shows that we needed to try more ratios to see if the complexations took place or not. I made the same for both of the scaffolds.
In spite of the fact that, I tried more ratios of complexation, we did not manage to ensure that the complex was taking place. That is way, I carry out an ELISA [155] affinity assay. All the cohesins in each scaffoldin bound their respective dockerin in a specific way and failed to bind (or bound very poorly) other nonmatching dockerin-bearing molecules. The scaffoldin-borne cohesins bound their matching dockerins as efficiently as the individual monovalent scaffoldins did, indicating that the binding capabilities of the scaffoldins were reliable and selective. All specific cohesin-dockerin interactions, for each scaffoldin, were of similar intensity as judged by the affinity enzyme-linked immunosorbent assay ELISA procedure, thus indicating that similar amounts of protein were bound in each well, suggesting a molar equivalent of the 1:1 scaffoldin (cohesin)-to-dockerin ratio.
Experimental Results

By means of this assay (Fig 5.30) we conclude that we succeeded with the complexation.

In the case of exoglucanase, the same native gels run for endoglucanase were done as Figure 5.29 shows.

![Native gel of the complex form between Exo-DockG and the cellulosome scaffold. The number below represents the scaffold/enzyme ratio used in each case.](image)

As well as it happened for the endoglucanase, we were not able to see the complexation in the native gels. That is way; also in this case I performed an affinity ELISA. Nevertheless, in this case, the ELISA did not show positive results and that means that the complexation did not succeeded. We hypothesize that this could be because of steric effects and that there is a need on trying by the N-terminal of the ancestral exoglucanase. We designed new primers to a future approach.
In Figure 5.30 the ELISA experiments results are shown, the endoglucanase-dockerinT complex (in red) was formed as the affinity is really similar to the control (in black). In contrast, the exoglucanase-dockerinG complex was not formed, as the line (in green) has nothing to do with the control.

**Figure 5.30. Affinity-based ELISA.** The absorbance is plot against the cohesion concentration. Black line represents the control, red one represents the Endo-dockT and green one Exo-DockG.
Chapter 6: Discussion

In this thesis, I demonstrate that reconstructed ancestral enzymes have improved physicochemical properties that make them suitable for industrial applications. Overall, I demonstrate that the ancestral sequence reconstruction technique can be employed as a general technique for enzyme design. I show the high performance over a broad range of temperatures and pH values, of three reconstructed ancestral cellulases (endo-glucanase, exo-glucanase and beta-glucosidase). The activity of the endoglucanase was notably improved, both as a free enzyme and as a part of a cellulosome complex. The activity of this endoglucanase was compared with two commercial endoglucanases, _T. maritima_ and _T. reesei_, which are commonly used in industry. It is important to note that the comparison between ancestral endoglucanase and _T. maritima_ endoglucanase is very significant taking into account that _T. maritima_ is perhaps the most extremophile bacterium known. We also made a comparison between the ancestral enzymes cocktail and a commercial (Ctec2 cocktail) used in industry, in which our ancestral enzyme cocktail outperformed the commercial one. The specific activity of the ancestral cocktail was higher in all the tested conditions, not only using laboratory substrates as filter-paper but also using other lignocellulosic compounds such as cardboard.
The use of our technology, Ancestral Sequence Reconstruction, enables a bigger improvement of enzymes activity. Using this technique, phylogenetic relationship is studied by means of sequences of enzymes from modern organisms. This is the first time that such a work has been carried out explicitly focusing on industrial enzymes. Until now, ancestral proteins were reconstructed in order to study their evolution [21, 32]. We now extend the reach of the ancestral enzymes bringing real applications to industry such as the bioethanol production [44].

The ancestors of this modern proteins and enzymes often show exceptional properties related to its thermal stability, chemical or kinetic. High activities are expected because of the promiscuity, lower selectivity and therefore more effective with different types of substrates [2-8]. Ancestral enzymes are not necessarily more stable than modern extremophiles [22]. However, those enzymes living in the Hadean and Archean eons, with an estimated temperature of 60-70ºC in the oceans were thermophiles [20, 21, 186, 187]. This thermophilic phenotype is captured by ancestral reconstruction and exhibited by our ancestral enzymes. Nevertheless, ancestral enzymes display other properties too, such as broad pH, higher expression yields or chemical promiscuity [20, 22, 24, 27, 161]. These properties make them more efficient than the contemporary enzymes, including extremophiles and make them a good alternative in industry.

Ancestral enzymes are known as generalists, which mean that they have a higher applicability comparing with the extant ones. Those extant enzymes are known as specialists, they have evolved to be more effective in a certain organisms and conditions [188]. For example, in the case of thermophile enzymes they are more effective than others under high temperatures, but they may show elevated substrate specificity. Therefore, ancestral reconstruction
is a good methodology for protein engineering having a high applicability in biotechnology [28, 32, 34].

The first enzyme we reconstructed was the endoglucanase, we achieved high specific activities under different conditions and substrates, only testing one enzyme, which makes ASR more efficient than any other protein engineering technique. Other currently available techniques like directed evolution need many variants through long cycles of evolution and testing. Moreover, improving the thermal stability of an enzyme maintaining its catalytic activity unchanged is a challenge in protein engineering.

In this thesis, I also reconstructed the catalytic domain of two other cellulases, i.e., exoglucanase and β-glucosidase. It was not possible to reconstruct the CBM of endoglucanase and exoglucanase, as the sequences of the selected species showed a large variability in these domains, including their position either in the C or N termini. The CBM is poorly aligned demonstrating a molecular diversity that might reflect different origins for this module. For this reason, I decided to focus the analysis on the catalytic domain of the homologous sequences. This procedure was used for both endoglucanase and exoglucanase. In the case of beta-glucosidase, we were able to align all the sequence together as they only had one domain. Taking into account that the CBM is the responsible of hydrolyzing crystalline cellulose, we hypothesize that a change in cellulose structure might be after the origin of the addition of this domain. The fact that beta-glucosidase does not have this domain can supports this idea, as the linkages that it breaks are not crystalline.

The initial results of this thesis were those for endoglucanase. In comparison with the enzymes commercially available and that are used in industry of T. maritima and fungal endoglucanase from T. reesei, our ancestral endoglucanase showed higher activity
especially at high temperatures and high pH values. Although it is true that the thermophilic endoglucanase from *T. maritima* is slightly more stable than the ancestral endoglucanase, the fact that the specific activity of the ancestral endoglucanase is higher in all conditions tested greatly compensates, and makes it suitable to be used even in the hardest conditions. Apart from testing the ancestral endoglucanase as a free enzyme, it was also tested as a part of a minicellulosome complex. In both cases, the ancestral endoglucanase outperformed the extant ones. It is remarkable, that in the case of the minicellulosome, not only amorphous cellulose was tested, but also crystalline one, which is of a big interest for industry [182].

Moreover, the ancestral endoglucanase also showed very good synergy with other lignocellulosic enzymes such as laccase and xylanase. This synergy was proved using cardboard as substrate.

We obtained similar results when using the ancestral cocktail of enzymes. Ancestral cellulases showed again, that they are much more active than commercial cellulases, exhibiting almost twice their specific activity. This result supports the theory that ancestral proteins can be adapted to harsh conditions obtaining better performances. These ancestral enzymes could obtain higher bioconversion yields and improved the efficiency of the process. Therefore, they represent a promising implementation for industrial usage to overcome the limitations of the current process.

Furthermore, it was observed that the addition of other lignocellulosic enzymes, such as laccase had a positive influence on the hydrolysis, leading to higher activity of the cellulases. This positive effect is caused because of the ability those lignocellulosic enzymes ability to break down the complex network of lignin and thereby, increase cellulose accessibility to
enzymatic attack. In the experiments presented here, it has been possible to observe the effect in different lignocellulosic materials, such as paper and cardboard. Even though different lignocellulosic substrates are composed of the same components, they differ in the percentage of the components and the structure. As a result, different performances are obtained in each substrate. Moreover, its recalcitrant structure and different components makes it difficult the hydrolysis, and lower activity values are obtained compared to the ones obtained in pure cellulose substrates such as filter paper. We expect that other lignocellulosic enzymes, such as laccases and xylanases, including fungal cellulases can benefit from ancestral reconstruction, which can help to generate highly efficient cocktails providing the improvement of the saccharification of cellulosic substrates for numerous industrial applications.

In the final part of my thesis, I show the addition of these improved enzymes to the designer cellulosome. Although it was not fully completed, we expect an optimized degradation of all types of lignocellulosic materials. The synergy of techniques, the ancestral sequence reconstruction and the designer cellulosome will suppose a biologic advance.

Lots of efforts have been made in the last years in order to improve the performance and efficiency of enzymes in biotechnology. In fact, it has been one of the paradigms of modern molecular biology. There is still a need for further improvements, especially for industrial applications. Besides the improvement in the efficiency of enzymes, there is a need to upscale the production, in order to supply industrial demand. For that purpose, we can use different organisms that have been genetically modified to produce large amounts of enzymes e.g. *Trichoderma reesei* RUT C30 [189] or *Bacillus Subtilis* [190].
Furthermore, nowadays the possibility of designing an organism able to produce the desired improved enzymes is opened using CRISPR cas9 technique [191]. One can expect that in the future, the combination of existing techniques, such as directed evolution or rational design with ancestral sequence reconstruction could lead to novel enzymes with multiple improved properties and even new tailored functions.


46. W. Tang, X.C., H. Zhang, F. Chen and X. Li, Limitation of the Development on Cellulose Hydrolysis by Cellulase


57. Bayer, E.A., et al., The cellulosomes: multienzyme machines for degradation of plant cell wall


68. Menon, V., Rao, M., Trends in bioconversion of lignocellulose: Biofuels, platform chemicals &


159


118. Kumar, S., G. Stecher, and K. Tamura, MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0


141. Hong, J., X. Ye, and Y.H. Zhang, *Quantitative determination of cellulose accessibility to cellulase based on adsorption of a nonhydrolytic fusion protein*


162. Nidetzky, B., W. Steiner, and M. Claeyssens, *Cellulose hydrolysis by the cellulases from Trichoderma reesei: adsorptions of two cellobiohydrolases, two endocellulases and their core proteins on filter paper*
172. Kinnarinen, T. and A. Hakkinen, Influence of enzyme loading on enzymatic hydrolysis of cardboard waste


182. Ling, Z., et al., Unraveling variations of crystalline cellulose induced by ionic liquid and their effects on


Appendix I

List of endoglucanase proteins from the species used in the construction of the phylogenetic tree

**Q59665 Cellvibrio japonicus**

ANGQPASFSGMSSFNTEWGGGYNNAAQVVSWLKSDWNAKLVRAMGVEDGGYLTDTPANKDRVTQVVDAAAIANDMYVIDWHSHNAHQYQSQAIAFFQEMARKYGANNHVIYEYIYNEPLQVSVSWNTIKPYAQAIVAIAIRAIIDPNLIIIVGTPTWSQVDVAANDPITYQNIAYTLHFYAGTHGQYLRDKAQTLNRGIALFVTIEWGSVNAANGDGAVANSETNAWVSFMKTNHISANAWALNDKVEGASALVPGASANGGWVNSQLTASGALAKSIIS

**Q8VUT3 Pseudomonas sp**

APISTNGNQLLFGERAVAGDESVAGPSLFNNNGWGGEFYNAGAVASAQQDWNAEI
RAAMGVDEGGYLEDASALNRVRAVVDAAIANDMYVIDWHSHHAE
SYTQAAVSVSFQQMASEYGQHNDVIYEIYNEPLSVWSNTIKPYAEQVIGAIRAVDPNLIVVGPTWSQVDVAANDPITYNNIAYTLHFYAGTH
TQYLRDKAQQYALDMGIFIPLFVTIEWGTNVANNDGDCVAYNETNTWMD
FKANNISHANWALNDKAEAGSSALVGTNPSGNWADNQYTASGTFVRDIVR

**C9DDS3 Pectobacterium carotovorum**

TPVETHQQLSIENGRLVDEQGKRVLRGVSSHGLQWFQGDYVKNDSMKWLRRDDWGINVFRAMYTADGYISNPXNKLKAVAAACGLGVYIIIDWHILSDNPNKIAAQAKTFFAEMAGLYGSSPNIYFENEPNNGVTTWNGQIPYALEVTDTIRSKDPDLIIIVGTGTWSQDIHDADNQLPDNPSTMYALHYAGTHQQLFRLRDIDYADSQSRGAAIFVSEWGTSDASNGGGPFLPESQTWIDFLNNRGRVSVWNWSLTDKSEASAALAPGASKSGGWTEQNLSTSGKVFREQIR

**Q59394 Pectobacterium sp**

169
Appendix I

TPVETHGQLSIENGRLVDEQGKRVQLRGISSNLQWVGDYVNKDSM
KWLWDDGWGINFRAMYATAENGYIANPSLANKVEAVAAAQGLGV
YIIIDWHTLSNDPNTYKQAKIAKFAEMAGLYGNSPNIYEIAENPNG
SVTWNGQIRPYALEVTDTSRKDPDNLIIIVGTWSQDIHDAADNQLP
DPNTLYALHYAGTHGQFLRDRIDYAQSRGAAIFVSEWGTSDASNG
GPFLEGQTWIDFLNNRGISWVNWLSLDKSETSAALVAGASKSGGW
EQNLISTGKFVREQIR

Q59395 Pectobacterium atrosepticum

TPVETHGQLSIENGRLVDEQGKRVQLRGISSNLQWVGDYVNKDSM
KWLWDDGWGINFRAMYATAADGYISNPSLANKVEAVAAAQGLGV
YIIIDWHTLSNDPNTYKQAKIAKFAEMAGLYGNSPNIYEIAENPNG
SVTWNGQIRPYALEVTDTSRKDPDNLIIIVGTWSQDIHDAADNQLP
DPNTLYALHYAGTHGQFLRDRIDYAQSRGAAIFVSEWGTSDASNG
GPFLEGQTWIDFLNNRGISWVNWLSLDKSETSAALVAGASKSGGW
EQNLISTGKFVREQIR

D0KFU8 Pectobacterium wasabiae

TPVETHGQLSIENGRLVDEQGKRVQLRGISSNLQWVGDYVNKDSM
KWLWDDGWGINFRAMYATAENGYIANPSLANKVEAVAAAQGLGV
YIIIDWHTLSNDPNTYKQAKIAKFAEMAGLYGNSPNIYEIAENPNG
SVTWNGQIRPYALEVTDTSRKDPDNLIIIVGTWSQDIHDAADNQLP
DPNTLYALHYAGTHGQFLRDRIDYAQSRGAAIFVSEWGTSDASNG
GPFLEGQTWIDFLNNRGISWVNWLSLDKSETSAALVAGASKSGGW
EQNLISTGKFVREQIR

R9F9F1 Thermobifida fusca

TPVERYGKVQVCQHTQLCDEHGPNPVQLRGMSTHGIQWFDHCLTDSSL
DALAYDWWKADIIRLSMYIQEDGYTENPRGTDRMHQLIDMATARGLY
VIVDWHILTPGDYPYNLDRAKTFFAEIAQRHASHTNVLYEIAENPNG
VSWASIKSYAEEVIPVRQRPDSVIIVGTRGWSGPAAIAANPVNASN
IMYAHFHYAASHRDNLYNLRASELFPVFVTFGETETYTGADANDNF
QMADRYIDLMERIKGWTWNYSDDFRSGAVFQPGTCASGGWSGS
SLKASQGWVRSKLQ

R4T6Y4 Amycolatopsis orientalis

TPVSINGKLVHCQVCLNQYQGKPIQLRGMSTHGIQWYQVCTASLD
ALANDWKADILRVAMYIQDDGYESNPRKTDMMHNYYIEATKRGMY
VLVDWHQLDPGDPNVNVDLAKTFFTEIAQRHKDKVNIIYDVANE
GVSWADVKRYAEEVIPVIRAQDPDSVVFLGTHGWSTDETDILNNVPN
ATNIMYTFHFYAASHQDEHYDALARTADKLPVFVTETFQTITYTDG
GNDFTYSQKYLDDLAAKKGWTNWNFSDDFRSGAVFKTGTCAGNSF
TGSTLKPAGVWVRDRIR

F4FAV2 Verrucosispora maris

TPVQINGQLVRVCVNLNCQYQRPIQLRGMSTHGQWFGNCYNNASL
DALATDWRADLRFRIAMYVQEQQGYETDPAGTNRVNVLVEEATRRGM
YAMIDFHHLTPGDPMFNLERAKTFFAASVARHASKNNVIYEIANEPN
GVSVSTIKNYADQVIPVIRANDPDAVVIVGTRGWSSNHEIVNNPVPN
ASNVMYAFHFYAAHRSDDNYRAEVERAAARPLFVTEFGTVTDYTDG
GVDELASSTQWLDLRLKIGYANWTFSDKAEGSAALRPGETNCNSNY
TGSTLTPSGVFMRRIR

D9TBA5 Micromonospora aurantiaca

TPVAINGQLQVCGVNLNCQYQRPIQLRGMSTHGQWFAFCYTDASL
DVLANEWRSDLLRISMYVQEQQGYETNPAGTNQVNTLVDAEARGVGM
YALIDFHTLTPGDPMYNLDRAKTFFANVSARNAAKKNVIYEINENPN
GVSVSTIRNYAEQVIPVIRANDPDAVVIVGTRGWSSNDEVINNPVAR
QNIMYTFHFYAAHSDKNYESVRAAASRLPLFVTEFGTVTDYTDG
VTDASSNAWLDLRLKISYANWTLSDAPEGSAALRPGETCSGSFGG
TSLTEGAFMRIR

E8SBH4 Micromonospora sp

TPVAINGQLQVCGVNLNCQYQRPIQLRGMSTHGQWFAFCYTDASL
DVLANEWRSDLLRISMYVQEQQGYETNPAGTNQVNTLVDAEARGVGM
YALIDFHTLTPGDPMYNLDRAKTFFANVSARNAAKKNVIYEINENPN
GVSVSTIRNYAEQVIPVIRANDPDAVVIVGTRGWSSNDEVINNPVAR
QNIMYTFHFYAAHSDKNYESVRAAASRLPLFVTEFGTVTDYTDG
VTDASSNAWLDLRLKISYANWTLSDAPEGSAALRPGETCSGSFGG
TSLTEGAFMRIR

G8S2I4 Actinoplanes sp

TPLAANGQLKVCGAGLCNQNGKIQRLRGVSSHGIIHWFGPACYTAAM
DALATDWNADLFRIAMYVQEQQYESDPTGTKVNSLVDMAEAHGGM
YALIDFHLNPGDPMNLARAKEFFAKVAAARRNAAKKNVIYEINENPN
GVSVWAGIKSYAEQVIPVIRANDPDGIVIGTRGWSSSAEIDNPVNAT

171
Appendix I

T1V343 Amycolatopsis mediterranei

TPLAANGQLHVCGVHLCEANRAIQLRGMSTHGLQWFDSYNDASL DALANDWHADLLRIAMYVQEKGYETNPAWTDRVNSLVGEAEERGM YAIYFHTLTPGDPYNLDRATKTFFAAVAAARNAARKNIYIEIANE PN GVSWSAGIKSYAEQVIPVIRAADPDAVVIVGTRGWSSNETEIVNNPV AGNIMYTZHFYAAASHKDNYRATVSRAATRLPLVETFTGVTATGGG ALDQASTTAWLDLLDLQISYANWTYSDADESSAALQPGTCAGGDY GTGLRTASGALVRNIR

D6KDP0 Streptomyces sp

TPVGVNGQLHVCGVHLCNQNYNHIQQLRGMSTHGIQWFSCYNAASL DALATDWDKADLRIAMYVQEDGYETDPAGTSRVMGLVDMAAERG MYALIDFHTLTPGDPYNLDRATKFFASVAAARNAAKNNIYIEIANEP NGVSWSAIAKNYAEQVIPVIRAADPDAVVIVGTRGWSSNEASEVNNPV NAANIMYTZHFYAASHKDNYRSTVSRAASQPLVETFVTSATGGG AVDQASSTATAWLDLLDLQISYANWTYSDAPEGSAALKEPTCGGSY GGSALTEGALVKRS

S1SNP7 Streptomyces lividans

TPAAVNGQLHVCGVHLCNQNYRPIQRLRGMSTHGIQWFPGTGYGASL DALAQWDKSDLRLVAMYVQEDGYETDPAGTSRVMGLVDMAEDRG MYALIDFHTLTPGDPYNLDRARTFSSVAAARDKNIYIEIANEP NGVSWSATAVKSYAEQVIPVIRAADPDAVVIVGTRGWSSNEASEVNNPV VNAINTIMYAFHYAASHKDDYRAASVRAATRALPLVETFVTSATGGG GGAARDSSSAWLDLLDLQISYANWTYSDAPEGSAAFRPACETCGT DYSNGLTEGALLKRS

A5A6G0 Paenibacillus sp

GQLKVQGNQLVQGQAVQLVQMSHQLQWYGNFVNHSLQWMR DNWGINVFRAAMYATADGYTDPSVKNVKVEAQQASIDLGLVIID WHLSDGINPNTYKASKQFFQEMATLYGNTPNVIEIANEPVGGVSW ADVKSYSYEEVITAIRAIDPGVVIVGSPQTSQDIHLAADNPVSHSNVM YALHYSFTGQFLRDRITYAMNKGAAIFVTEWGDSAGSNGGPY
PQSKEWIDFLNARKISWVNWSLADKVETSAALMPGASPTGGWTDAQ
LSESGBKVRDQIR

**I0BQW9** *Paenibacillus mucilaginosus*

AAA岁月GKLKVQGADLGGEQGQVQLRGMSSHIIHWGYDLNPGS
LKWLDWDWNSLFRVAMYTAEKGYITDPSVKEKVKVKEAVQAAILDLGL
YVIIIDWHILTDGDPNTYKTQAKAFFQEMAALYGQYPNIVELYCNPN
GNWTVAGQIKPYAQELTCAIRADPNIIIVGTNPWQSDVQADASPL
PYGNIMYAAHFLYAGTHGQWLRDKNIDYARSKGAAALVFTEWGSASDAS
GDGGPFLREAEQEWIEFMSRGIISWANWSLADKEETSAALLPGANPS
GGWPASQLSSASQFVRSKLR

**I7L2V5** *Paenibacillus polymyxa*

TPVERYGQLSVKNGKLVDKNGQPVQLKGISSHGVQWFDLVPNQDT
MKWLRDDGWISLFRVAYLTEENGYIANPSLKNKVEAIAEAAQKLGL
YVIIIDWHILSDGDPNTHKNEAKAFFNEFSTKYGHLNIVIYLEANEPNG
NVNWNQIRPYASEVSQVIRAKDPDNIIIIVGTGTSQVDVHDAADHP
LDKNTMYTVHFYAGTHQQSLDRDYALNKGVIGATEWGSASDAS
NGGPFLNESKLVWTDMSRISWANWSLSDKNETSAAALLPGARAKG
GWPSQLTASGKFVKQAIL

**G7VSG4** *Paenibacillus terrae*

TPVERYGQLSVKNGKLVDKNGQPVQLKGISSHGVQWFDLVPNEDS
MKWLRDDGWISLFRVAYLTEEDGYITNPSLKNKVEAIAEAAQKLGL
YVIIIDWHILSDGDPNIIHKNEAKAFFNEFATQYGNLNPVYILEANEPNG
NVNWNQIRPYASEVSQVIRAKDPDNIIIIVGTMWQSDVHDAADNH
LPDKNTMYTVHFYAGTHQQYLRDRDYALNKGVIGATEWGSASDAS
SGNPPFLNESKLVWTDMSRISWANWSLSDKNETSAAALLPGARAKG
KGGWPDSQLSSGKFVKQAIL

**D91A39** *Bacillus megaterium*

TPAAAKNGQLSIKGTQVLNRDGKAVQLKGISSHGVRWYGDVFVNKDSDL
KWLDDGWGIVFRAAMYTDGGYIDNPVKNKVEAIAAEKELG
YVIIIDWIIHDGYPQNEHKEKEAKEFKEKEMSSLGNTPNIYIEANEPNG
DVWKRDKIPYAEVEISVRKDNPDIIIIVGTGTWSQVDVНАADDQIL
KDNVMYALHFYAGTHQQSLRDKKYALSGPAPIFTEWGSASDAS
GNGGFVLDQSREWLNYLDSKNSWVNWNLSDKQETSSALKPGASKT
GWPLTDLTASGTFVRENIL

173
Appendix I

**P23549 Bacillus subtilis**

TPVAKNGQLSGKIQNLAVKQGKIVQGISSHGLQWYGQFYVKNDCL
KWLDDGTVFAAMYTADOGGNPSVKNKMKEAAAEKELGIY
VIIDWHILNDGPNPQMKEKAEFFKEMSSLYGNTPMYVENIYCEPNG
VENKRDQKPYAEVIRKNDPDDIIIVGTWQSDVNDADDQLK
DANVMDALHFGYAGTHGFQLRDKANYALSNGAPIFVTEWGTSDASG
NKGVFLGDQREWLOYLSKTSWVZNWNSDQIESSASLPGASKTG
GWRGLDSLASGTVFVRENIL

**D8WN01 Paenibacillus campinasensis**

TPVAKNGQLSGKIQNLAVKQGKIVQGISSHGLQWYGQFYVKNDCL
KWLDDGTVFAAMYTADOGGNPSVKNKMKEAAAEKELGIY
VIIDWHILNDGPNPQMKEKAEFFKEMSSLYGNTPMYVENIYCEPNG
VENKRDQKPYAEVIRKNDPDDIIIVGTWQSDVNDADDQLK
DANVMYALHFGYAGTHGFQLRDKANYALSNGAPIFVTEWGTSDASG
NKGVFLGDQREWLOYLSKTSWVZNWNSDQIESSASLPGASKTG
GWRGLDSLASGTVFVRENIL

**P06565 Bacillus cellulosilyticus**

SVVEEHGQLSISNLVGVRGEQVQLKGMSSHGLQWYGQFYVYESM
KWLDDGTVFAAMYTADOGGNPSVKEKVEAAEADLDIYY
VIIDWHILSDNPPNIEYKEEAKDFDEMSLEYGDYPVIYRIVIYEANPNGD
TVDNQKIPYAEVIRKNDPDDIIIVGTWQSDVHDANQLED
NVMYAFHYAGTHGFQLRDKANYALSNGAPIFVTEWGTSEATGDG
GVFLDEAQVWIDFMDERNLSWANWLSLTHKDESSAAALMPGASPPTG
GTWEAEALPSGTVFVRENIL

**O85465 Bacillus agaradhaerens**

SVVEEHGQLSISNLVGVRGEQVQLKGMSSHGLQWYGQFYVYESM
KWLDDGTVFAAMYTADOGGNPSVKEKVEAAEADLDIYY
VIIDWHILSDNPPNIEYKEEAKDFDEMSLEYGDYPVIYRIVIYEANPNGD
TVDNQKIPYAEVIRKNDPDDIIIVGTWQSDVHDANQLED
NVMYAFHYAGTHGFQLRDKANYALSNGAPIFVTEWGTSEATGDG
GVFLDEAQVWIDFMDERNLSWANWLSLTHKDESSAAALMPGASPPTG
GTWEAEALPSGTVFVRENIL

**U5MQR0 Clostridium saccharobutylicum**

174
ATTSFGQLKVSQLQLCDNSNGKPIQLKGMSSHLQWYGQFVNYDSM KFLRDKWGVNVIRAAMYTNEGGYISNPSSEKEKKIKIVQDAIDLNYV IIDWHILSDNNPNTYKEQAKSFQEMAEYYGKYSVNIYEICENPNGT NWANDIKPYANYIIPAIDAIPNIIIVGTSTWQDVAADNPRLYSNIMY TMYCFYAGHTQSLRDKINAMYAKGIAIFVTEWGTSDASGNGGYYL DESQKVVDVMASKISWNWALCDKSEASAALKSGSSTTGGWTDS DLTTSGLFVKKSIG

**R6S0D3 Eubacterium siraeum**

TPVSQHGQLSVKNGQLVDKSGKGYQLRMSTHGLTWFPFEPVNESAF KTLRDDWNTNVRLAMYVDEGCYMGKNSGLEELGKVDCIKLDM YVIIDWHVNLNPDSKYTNEAKSFFETVSKRYAKPNVIYEICNENPG GASWYGKPYAEKIIPIVIRKNAPSNNIVGPTWQSEIDKPLDPLNY KNVMYAHFYAATHAGLRSVENCRAQLPVVFSEFGTCDASGGG ANDFNETQKWLFDQGISYCIMESICNKDETCSVLRPGTSANGNW SESNLTENGKWMRNWF

**R5K1B8 Clostridium sp**

TPLENHGALQVKTLVDKNGAPYQLKVSTHGLAWFPFYVKNDA FQTLRDDWGANVRLAMYTEDEGYCNREQLQVLSDGVEYATELG MYVIIDWHILQPDVYQGEAEDFFAEMSAKYAKYDNVIYIEICNED NNGATWDGSVPKYEIIPAIIPKNDKDAIIIVGPTWQDVAADNPI TQTENMYAIHYAATHTDGIRSKVSYALDKGPLPVVFSEFSICDASGGG SNDYDQAACKILDIDQHLYCIMESICNKDETCSVLRPGTSANGNW EDSDLGTGKWIRIQL

**R5JFB6 Coprococcus sp**

TPVENHGKLGVKTLVDKNGDKYQLKGLSTHGMWFQYVSEETF KTLRDDWGANLIRLAMYTDGNYCDEKAEIKKLLDDGVYASDLGM YVIIDWHILLNNPNNHIEDAKDFFNTSKEYEYDNLYEICNPNG GTTWTDVKSQAETIPVIRANDKDAIIIVGPTWQDVAASENPVYID NIMYAAHFYAATHQKELRDKISAIADGPLPVFSEFSICDASGNGAID YNEADAWFEDNKLYNLASYASWLSNKAETASLFFSSSTTVSDFSESDIS DTGKYIRDKIL

**R5W9F0 Coprococcus eutactus**

TPFDNHGQLSVKTDIVDESQSKYQLKGVTSTHGWFPDQYVNYDKAQ SIRDDWDANLVRAMYTDGCDKDSIRGLVDAGYTAATELGMY
Appendix I

VIIDWHILNDNPNHSHIDDAKEFFDDDVSAYSSNHNVIYEICNEPNGGTSWSDKYAEIIIVPKVDKNAIIIVGTPNWSQDVDDPITYDNIIMAYAHFYAATTHKDLDNNKVTASNGLPVPVFSEFSLCDASNGGIDYDSDWDVLINDNNLASYASWSLCNKNESALIKPDSTATSTITIDDSDTGKYVRDKI

I5AVZ7 Eubacterium cellulosolvens

TPFENGHKLSVKGTDLVDESQKQQLYQLRSTHGLAWFPQYVNVNADAFTRTLNDWGANLVRLAMYTDGGYCNQGELKDLIDRGVNYASDLGMYVIIDWHILYNDPQHKEEAKAFFDEMKTYYADRGNVLYEICNEPNATTWADVKKYAEEVIPVIRKNAPIAAICGTPWSQDVQVAADPIKGGNMLMYTLHYAATTHKDELDRKMQTAIAASGTVPVFISEFSICDASNGTLDYSSEAEKWNLIKENVLSFAGWSLSNKDEASAVRSGCDRLSDWaTEGELSDGWNWLKLVS

R6QN32 Butyrivibrio sp

TPVGHQGQLSVKQVGLDSDLNKQKQQLYQLRSTHGLWFPQYVNVKDAFKTLRDNWGANVRLAMYTDGGYCSKADLEAKIDEGVKAASELGMYVIIDWHILSDNPNPTYKDEAVKFFNKMSKKSQNNVNYEICNEPNNGGVDWNTIKTYADTIPIKNSTINPNAILVGTPTWSQDVAADNAANPVAKKNVMTLHYAAGTHKDNRNLTAARNAGTVPFISEFSICDASNGGGIDSTSANAWKHLNDNNVSYVGWLSLCNKAETSALIVSSCSKLSGWTDSLETSGKWLRFNIA

C4Z6Y9 Eubacterium eligens

VTTVPTGRLQVSGTKLTDSENSINQLRGVSTHGISWFPDYVNYDAAFATLRDDWGANVRAMAYPEENGYLDKAALQIIDNNGVYTELGMYVIDWHVLNYAPRHTQEACDFAEAKSYSGHDNVYNIEICNEPVGADWNSDIKPAYETVIGTIRFDHALILVGTNTWSQDVSDSVGNTLDDGNYMVVAHWFAGTHKERNKISTALNAGVPVFISECSICDASNGGGIDYASANETLDFQSNQLSFIAWSLSNKAETSALISSGSASKGWSDDGLSETGWFKSAIS

Resurrected endoglucanase sequence

TPVETHGHQLSVKQGQLVDENGKPVQLRGMSSHGLQWGFDFVNDSMKWLDDGWAINVFVAMYTAEGGYITNPVSNKVKVEAEEAAIDLGYVIIDWHILSDNDPNTYKQEKAFFQEMAAKGYGNPNIYIECNEPNNGGVTWSNQIKPYAEEVIPAIRANDPDNIIIIVGTPWTSQD
VHDAADNPLPYSNIMYALHYAGTHGQSLRDKIDYALSKGVAIFV
TEWGTDASGNENPFLNESQKWIDFMNRSNISWANWLSLDKSET
SAALMPGASPTGGWTDSNLSASGKFVREQIR
Appendix II

List of exoglucanase proteins from the species used in the construction of the phylogenetic tree

**O65986 Clostridium cellulovorans**

VVPNNEYVQHFKDMYAKIHNNANGYFSDEGIYHAVETLMVEAPDY
GHETTSAFSYYMWEAMNAKLTDGFSGFFKAWDVEKYIIPGETD
QPSASMNNDPNKATAYAAEHPSMPSMQLQFGAAVGSPLYNEL
KSTYGTSQVYMHWLLDNYGFWGFQATSTPVYINTFQRGQESC
WETVQPCKDEMKYGRGRDFLDLFTGDSQYATQFKYTNAPDADAR
AVQATTYAQLAASKWGVDISSYVAKSTKMDFLRFSDKHYFRAK
GNSTQAGTYDSQYLLNWYYAWGGGISSNWSWHRGSHHNGFYQ
NPMAAWILNTSDFILPSMQNAADVNDWNSLKRQIEFYQWQLQSAEGGIA
GGASNSNNGSYAQAPGRTRTFYGMGYTPHPVYEDPQSNWFQGQA
WSMQRVAEYYSSKVAPAKSSDWGKAKWACANQFDAAAKKFKIP
AKLVVTGQPDFTWGTGSYTNARNLQKVEAYGEDLGAGQSLSNALSY
YAKALESSTDADKVAYNTAKESTRKILDYLYASYQDDKGIAVTET
RNDFKRFNQSVYIPSGWTGKMPANDGIVQSGAFLSIRSKYQDPSWP
NVEAALA

**O82831 Clostridium josui**

PVNKVQYQRFESMNKIKDPNSGYFSEEGIPYHSVETLMVEAPYGH
VTTSEMSSYMYWLEAMNYGRFTGDSGFNKSWTTEKYLIPTEKDPQ
NSSMRSYDANKPATYAPEFQDPSKYPSPLDTSQPVGDPNQLTSAY
GTSMLYGMHWLDDFVDNWFVGVRADGTPSYINTFQRGEQESTW
ETIPQPCWEHKGGQGGFLDLFTKDGTGAPKQFYTNAPDADAR
VQATYWANENWANKEQKGSVTEVGAKSMGDLRYSSFDKYFRKIG
QPTQAGTYDSAHLLSWYYAWGGGVDSTWISIGNCSNHGFCQYQNP
PFAAWVLSTDSSFPSSGNTWAKLQLRQLEFYWQLQSEGAIAAG
GATNSWNGYRESISPTSTFYGMGYVENPYYADPGSNWTFGMQVW
SMQROAELYYKTDTRAKNLDDKWAKWNVSEIKFNAODTFQIPGTL
DWEQGDWPDTQGTYGTPMNHLKVVNYTDLGCASSLANTLYY
AAKSGDTSKENGAKKLDAMWNNSDGMISTIEQRGDFYHRFLQDE
**P50900** *Clostridium stercorarium*

SSDDPYKQRFLELWEELHDPSNGYFSSHGIPYHAVETLIVEAPDYGH
TTSEAMSYYLWLEALYGFKTGDFSYGFMKAWEIEKYMIPTEQDPN
RSMAGYNPAKPATAYAPEWEEPSMYSQLDFSAPGIDPIYNELVSTY
GNTIYGHMWLLDDVNYGWGFRADRISPSAYINTFQRGSQESVVE
TIPQPCWDDLTIGGRNGFLDFVGDSQYSAQFKYTNAPDADARAIQA
TYWANQWAKEHVGNSQYVKKASRMGEDLRYAMFDKYFRKIGDS
KQAGTGYDAHHLLSWYYAWGGITADWAREIGSCHVHAYQNP
MTAWILANDPEFKESPNGANDWAKSLERQLEYFQWLQSAEGAIAG
GATNSKYKRYETLPAIGSFTYGMBAYEHPVYLDPGNTWFGFQAWT
MQRVAEYYYLTDRAEQLDDKKEWDCVSVRLDNSDGFTEIPGNLE
WSQPDWTGTGTYNPNLHVSYGTDLGAAGSLANALLYYAKT
SGDDEARMNAKELLDRMNLYRDGSAPETREDYVRFFEQEVYV
PQGWSGTMPNGDRIECPVFSLDIRSKYLDNDPYPKLLQAYNEGKAPV
FNYHRFQAQCDAIAANGLYSILFG

**A9KT91** *Clostridium phytofermentans*

TRGTYEQRFMDLWSDIKNPKNGYFSPQQIPYHSIETMIVEAPDYGHVT
TSEAMSYYMWLEAMYGKTGDFSYGATWNAVAEKYMPIPTADQPP
TSMKSYTPSKPATAYPEQDPSQYPAKLDSSAPGSDPISWQLVAAY
NTIYGMHWWLDVDNWYFGSRGDTSKPSYINTFQRGEQESTWETIP
QPCWDTMKYGGTNGFLDFLTGDSYAQFQKTDAADARAIQAA
YWASEWAKDYGVNVNDTYSKATMMGDYRYSMDKFYRKIGNST
VAGTGYASHYLLSWYYAWGGITADWAVGSSHIHPQMN
AAWVLSQNSKFKPKTTNQADWATSLLRQLEYQWLQSESSEGIGAG
ASNSKNGRYETWPGATATFYGMGYEAPIVYKDPGNTWFGFQAWS
MQRVAYYKTNDVKAKQILDKWAVKSSVTVKADGTDTPSTLD
WSQPDWTGTGYSYTSNKLHVTVDGSDTLDGTVSLANLLYSSA
ANDVAAKNSATELLDDRWVKLYRDDKGAAPEARADYKRFFEQTVY
VPSTFNGKMPNGDVIKSGBKIFLDIRSKYLDPSYPKLLAAYQSNKSEP
FIYHRFQAQCDAIAANGLYSILFG

**W6AP62** *Bacillus pumilus*
Appendix III

NKERFLTLYHQMKNDANGYFSKEGIPYHSVETLICEAPDYGHMTTSE AYSYWLWLEVLYGHYTGWDWSKLEAAWDNMEKIFIIPVNEGDHDEQP HMSSYNPPSSPATYASEKPYPDQYPSQLTGARPAGQPDPIDHELRYSTYG TNEYLMHWLLDVNWYFQRRGQPQESVWEAIPHPSQDDKTAGKPNPEGFMSLFTKEDQAPAAQWRYTNATDADARAIQ AMYWAKELEYNGSAAYLDKAKKMDFLRGYMYDKYFQTGIGKGG NPYPGNGKSAHCHLMAWYTSWGGLGGEYANWSWRIAGASHCHQGY QNPVAAYASLSEDKGGKLPSSTGASDWEKTLRQELFYVWLQSEKG AIAAGGATNSWNGDAYSAPAGRSTFYDMAYEDAPVYHDPPSNWNG MQAWPIERVAELYYIFAKDGDKTSENFQMAKSVIKVWNYSLDYFI GTRPVSDQDGYFLDGQGQRILGGANAVATTSGEFVPGNIEWSGP DTWSGFQSATGNPNLTAVTKDPTQDTGVLSGLIKALTFWAATQKE TGNYTAIRAKTAAQILLEAVAWNYNDGVIGVTEEDREYHRFTGK WENGAPSFTYHRFSQVDMATAYAEYHRLIN

10BR01 *Paenibacillus mucilaginosus*

DTTNKTRFLTLYNQIKDPANGYFSPEGIPYHAETLSEAPDHGHMTT SEASYWLWLEALYGHHTGNWTRLEQAWDNMEQYIIIPNASEQPTM SGYNPSATYAPHRQPDQYPSQLGSSVTAGKDKPLDAELKATYGSN QTYLMHWLVDNWYWGFGLNPSHTATYINTFQREGQESVWEAIP HPSQETFQFGKPEGFGATLFVKDSGAPAKQWRYTDATDADARVQV MYWAKSLGYNPSYIEKAKKMDLYLRYMYDKYFQIGSAADGTP SPGTGKDSSHYMAWYTAWGGGIGNSWAWRIGASHNHQAYQNPMA AYALSEGGLAPSKTAKQDWESELRQELFYWTWQSEGGGATN SKGGTYAPAYAGVSTFYGMAYEDAPVYHDPPSNWFGQAFQAWPVERI AEYYYAMAKGDTASENFKAMKRVMDQVVKWALAYTTPAGQF YILGGQETWQDPSWKGFSSFTGNPNHYVIAGKQSDTGVSLSYIKL LTAFYAAQTQAEENGTLSAAAGAQKATCTAEQLLNVAWNHNDGIGIAV EKRGDYRFTKEIYFPSGWSTYGQGNALPGTAVPSDPAKGGNG VLYSYESLRPKIKQDPKWPLENLYKSSYDPVTKKWTNGEPTFTYHR FWAQVDMATAY

Q8KKF7 *Paenibacillus barcinonensis*

DGINEARFLQLYAQLKDPANGYFSAEGIPYHSIETLSEAPDYGHMST SEASYWLWLETMYGHYTGWDWSKLEAAWDNMEKYIIIPVNEGDGNE EQPTMSKYNPNATYAAEKPFDPQYPSNLNGQYAAKDKPLDAELK ATYGNQTYLMHLDVDNWYGYGNLLPSHTSTYVNTFQREGQES VWEAIHPHPSQDDKFSGKAGEGFMSLFTKESQAPAQWRYTNATDADA RAVQVLYWAKEEMYNNSEYLDKAKKMDFLRYMYDKYFQKVGS
AKNGTPTPGTKDSNMYLMAWYTSWGGGLQGWDWAWRIGASHT
HQAYQNPVAAYALSDPAGGLIPDPTAKADWNATLKRQLEFYTWLLQ
SHEGAVGGGATNSIGGSYAAYPAGVSTFYDMAQYEAPVYRDPDNT
WQFQAWPLVERAEMYYLAESEGDLLTSENFPQKAKVITKWDWTW
DYVTTPAPGEFWIPQGQWQGQPDKWNGFSTYTTNNPNHAITKDPPV
QDTGVLQSYIKALTFFAGTEAENGTLSAKQGEAQDLSLLDSTAW
DYNDGVGIVTEEERKDYFRYFAKEIYLANWSGTFQGQNTIPGTAQVP
SDPAKGGNGVYIGYSDLRPAIKQDPAPAWAYLDNKYKSYNPTTKQWE
NGAPFTFYHRFWSQVDMATAYG

**W4CUL3 Paenibacillus sp**

ASVEKTRFQLYQLDPKDPASGYFSAEGIPYHAVETLLLSEAPNYGHMT
TSEASYWMWLEVLGYNGNTGDWHLESAWDNMEKYIIPNEDQGV
QEQPTMSSYNPNSPATYASELPQDPQWPSALSNGKYPGKDPLDAEKLK
STYGNQTYLMHLVDNYYFQGFNNLPHTASYVNTFQRGQVE
SVWEEAVGHPQDNKTFGKSNEQFMFLFTKESVPSAQWRYNTNAD
DARAVQAMYWAKDLGYNTVLYNKKVMGDFRLRMYGMYDFQKIK
GSAANGSPQPGTQGKDDSTLLWOTAWQGGLGTDGWAWRIGASH
AHQGYQNVVAYALSTAAAGLIPSSATAGTDWQKSLTRQLEFNYWL
QSAEGIAGGATNSYGGSYAYPQKSTFYGMEAYEAVHPQPSN
NWFGMQSWSMERMAELYYILASSGDTSDNFKMAKRVINETYWNWS
KDYFYILGQWQETQPDPSWKGFSFTGNPNHYIAGKTSQDTGL
GSYIQLLTFYAAQTQAENNNTLSAAGAQAKTTAEQLLNVAWHNHD
GIGIAVRKRGDYRSQFKEFIYPFSGWSGTYGGQNALPTGAVPSPDA
KGNNGVYLSYSELHPKIQDQPKWPYLELYKSSYDPVTKWNGEP
TFYHRFWSQVDMATAYG

**G7VQK6 Paenibacillus terrae**

ATPESTRFQLYKQLKDPASGYFSKEGYIPYHSVEKLLSEAPDYGHLTT
SEASYWMWLEVLGYNGNTGDWHLESAWDNMEKYIIPGKEEQPTM
SNNPNPSTATYAEQPDLPQPSLRDSLQYSAKQDPDSELKATYGN
QTYMHWLVDNYYFQGFNNLPHTASYVNTFQRGQVESVWEAIP
HPSQDNHKFQSSFMFLFTKESVPSAQWRYNTNADARAVQA
MYWAKELGYDNSVYLDKAKKMDYLYRMSYDMYDFQKAGSASKGS
PRIAGTNGDASFYLMAWYTAWGGGLQGSGNWAWRIGASHAHQGYQ
NVVAAYALSDKGGLIPSPQTPAQDWDATSLKRQLEFYTWQLSDEGA
IAGGATNSWDGAYKAPSTGTYGAMYMTAGPVYQDPSSNWFMG
QAWPVERAELYYLAKKGDTSSEQFQMAKQVTENWIAWSKSYVK
GEFWLPSDLEWSGKPDTWSGFANHKGNTLSHVTUKPVQDAGVLG
Appendix III

SYAKALTFFAAGTKAEGKFYSELGEAKDSLKLALLDAAWSYNDGIG
ITTKEAREEEYYFSKEVYIPNGWSGKTGQGNTIPGDATPDPKGGG
NGTYSTYSDIRPNITKDQPWSYLKDKYTTTWSNSQTQKWDKGAPQFT
YHRFWAQVDMATAYA

E0RLD5 Paenibacillus polymyxa

ATPESTRFLQLYQKDPASGYSKEGIPYHSVETLMSEAPDYGHLLT
SEXASYWMWLEVLYGHTGDWGHLESAWDNMEKYIPYNEDGKGE
EQTMSYNPNSPATYAAEYSQPDQYPSRLSGQYAGKDPDSLSEKL
TYGNQNYTLWHLWLLDVWYFGEGNLNPSTHAAYNTFQRGEQES
VWEAVPHPSQDNQKFGKLANEGFMSTFSTKENAPAQWRYNTNAD
DARAVQAMYWAYLGYDNSVYLKDACKMDFLRYGMKDQYFKQ
TGSASNSPIAGTGKDALMYLAWAYTGGLQGQSNWAWRIGMASH
HADQGGYQNAXYALSNRDGGGLIPNSTAQGWAYATLSKQLEHYT
WLQSDGIAAAGATNSWDGAYKAYPSGTSTGMYATGAPVYNDP
PSNWFQGMAWQPVERVA
ELYYILAKKGDTSSEQFKMAQVNTENWIAAWSKSYVTAAKGEFWL
SXLNWSKPMETWLSVENHKGANTNHLHVVTNPGQDAGVLSYVKAL
TFFALGKTAEKGFSELYSKEAKDSKLALLDAAWGYNDIGITKKEA
REDYRYFTKEVYIPNGWSDGKQGNTIPGDATPDPSKGGNTYS
SYSIRPNITKDQPWSYLKDKYTTTWSNQTKWWAKGAPQFTYHRFW
SQVDMATAYA

R4LQA1 Actinoplanes sp

EGSADARFAQLYNDIKNPANGYFSPEGVPYHSVETLIDEAPDQGHT
TSEAFSYWMLWEAEHRVAGDWTSFNAAQTMEKYIIPSHADQPTN
DASKDPATYAAEHPPLSQYPSQDSSVSSVTDAPLNEKSTYGT
DIYGMWLLDVWYFGEHCGDKTSRVTYNTFRQGQPESTETVPQP
SCDTHAAGPGNYDLLFTKDSYAKQWYTDAPDASRAVQAAYW
ALTWATEGKASQVSASVANAAMDYLRSFYDKYFKNPQGCTST
GCAAGSGBKSSNLMSMYWWYAWGAATDTSAGWAWRIGSSTSHFGYQ
NPMAYYLSNNASMTPKSPKATADWQASLNRQLEFYQWQLSSSEGGI
AGGATNSWNGYSQAPSGTPTFYGLYVEAPVYEDPPSNRFQGMQT
WSSLRLEAYYLSGDTKAKAVLDKFWALDNSTIGATSFESIPSDL
WSGKTPWNPSPPAANTGLHVTSTKQGDLGVAGSFACKLLTYAAK
SGNTRAKDAAKGLLAIWAYSKGVSVTETNRMDDVYNAS
TGQGITYIPPGWEGKMPNGDVIKPVSFLDISWYKNDPFDKVQSYL
DGPPAPTFNYHRFWAQVDVATAYA
**A0LS10 Acidothermus cellulolyticus**

NDPYIQRLFMYNKHDPANGYFSPQGIPYHSVETLIVEAPDYGHETTSEAYSLWLEATYGAVTGNWTPFNNAWTMTMTYMIPQHADQPNNASYNPNPSAPYAPEEPLSMPYPAIDSVSVPGHDPLAELQSTYGTDPYGMHNLADVDNIYGYGDSPPGCELPSSAKGVSYINTFQRGRQESVWETVTQPCTDNGKYGGAHGVDLIQPQGSTPPQWWKYTDAADARAVCAAAYWAYTWSAQGKASAIAPTIAAKAKLGDLRYLFDKYGKQVGNCYPASSPCGATGRQSETLYLGWYAYWGGSSQGWAWRIGDAAHFHYIQNPLAAMSVTPLIPLSPTAKSDWASLQROLEFYQWQLSAESGAGGATNSWNGYGTTPAGDSTFYGMAYDWEVPYHDPSSNWFGFQAWSMERVAEYYVTGDPAKALLDKWAVWVNPVTTGASWSIPSNSWSQGPDWTNPSNPSTANLHVTITSSQGVDVAAALAKTLEYYAAKSDGTAASDNLAKGLLDIWNQDSLGVSTPETRTDSRFTQVDPTTDGDLYIPSWGWTGMPNGDQIKPGATFLSIRSWYTQDKQWSKVQAYLNGGPARTFNHYRFWAEDFAMANA

**W7VNI1 Micromonospora sp**

DNAYIKKFLDQYGKIKNSGYFSPEGVPHYSIETLIVEAPDHHGTTASEFSLWLEATYGAVTGNWAFNNAWTVMEKYYIPSHADQATAGSPGTPQYAAEHNLYPSQYPSTLDANVPVQDPQLRSLQSTYGTGDYMGHWLLDQNTYGYCRCDGTTRPAYINTFQRGTQESVWETVPQSPCDTFKHGPNGYLDLFVKESSNAPAQAISKYNAPDADARAVCAAAYWALTWAKAQKAGDVATVAKAAKMGDYLRYALFQYFKIKNCVCTACPASGRDAHLLSWYAYWGGAYDASQNWSRIGSSHSHFYQNPFAAWVMNVAELKPKSTPAASDWQKSLDRQLEFYTWLQSAEESIAGGATNSWQYQAPATFYGMFYDVPVYNDPQNSQWFQMCAWMSQRIAELYLETGNAAKAKLLDKWVPWAIANTLTGTDWSIPSDMKWTQPAWNWSSPPQPNLHVETTVKGQDVGAVAGAYARTLIAAYAKSGNTAAKDTAKGGLDACADASKGVSTPEKRGDYKRFDDVYADQGLYIPNGWGTKMPNGDVIAPGKSFIIIRSFYKNDPDWPKVQAYLDGGPEPVNFYHYRFWAQADIAMA

**A4X938 Salinispora tropica**

DNAYVARFLQTQYGKIKNSGYFSSEGVPHYSIETLIVEAPDHHGTTSEAFSDLWLEATYGAVTGNWTPFNNAWTMTMTYMIPQHADQPNNASYNPNPSAPYAPEEPLSMPYPAIDSVSVPGHDPLAELQSTYGTDPYGMHNLADVDNIYGYGDSPPGCELPSSAKGVSYINTFQRGRQESVWETVTQPCTDNGKYGGAHGVDLIQPQGSTPPQWWKYTDAADARAVCAAAYWAYTWSAQGKASAIAPTIAAKAKLGDLRYLFDKYGKQVGNCYPASSPCGATGRQSETLYLGWYAYWGGSSQGWAWRIGDAAHFHYIQNPLAAMSVTPLIPLSPTAKSDWASLQROLEFYQWQLSAESGAGGATNSWNGYGTTPAGDSTFYGMAYDWEVPYHDPSSNWFGFQAWSMERVAEYYVTGDPAKALLDKWAVWVNPVTTGASWSIPSNSWSQGPDWTNPSNPSTANLHVTITSSQGVDVAAALAKTLEYYAAKSDGTAASDNLAKGLLDIWNQDSLGVSTPETRTDSRFTQVDPTTDGDLYIPSWGWTGMPNGDQIKPGATFLSIRSWYTQDKQWSKVQAYLNGGPARTFNHYRFWAEDFAMANA

183
Appendix III

TFTHGQYGFGLDISVQEQNAPAQQWKYT
NAPDADARAVQAAYWALTWAKQQGRAAEVAATVAKAALKLDYL
RYAMFDKYFKQIQNCVGASTCPAGSREGSASHYLLSWYYAWGGAYE
SGQNWSWRIGSSHNFHYQQNFIAAWALTTVPELEPSPSATTDDWAR
SLERQLELYTWLQSAEGAIAGGATNSWGGYRUAQPPAGTPTFYGMYF
DEKPVYHDPPSNQWFGVMQVWSMHIRAELEYLETGDARAEALDRWV
PWAIANTRLGDWSIAPELTWTGQPNTWNPNPEPTNDLHEVTETG
QDVGAAAYARTLIYAARSGNVTAKTAKGLDLLDALHAAASDALGV
STVEKRGDYERFDVVDYDASTQGQLYPGWTGTMPNGDVIEAGRSP
VEIRSFYLNHDWPQVAYLDGGAEPFRYHRFWAQADVAMAYA

Q9XCD4  Thermobifida fusca

TSSYDQAFLEQYEIKKDPASYGREFNGLLVPYHSVETMIVEAPDHG
HQTTEAFSYYLWLEASYDRGVTGDWKLHDAWESMETFIIPGTDQ
PTNSAYNPNSPATYIPEQPNADGYPSPMLMNVPVQDPLAQELSSTY
GTNEIYGMHWWLDVDNVYFGFGCDFGDGTDDAPAYINTYQRGARESV
WETIHPSCDDFTTHGPGNYLDLFTDDQNYAKWRYTNAPADARA
VQVFMWAHEAKEQGKNEIAJMLDKASKMGDYLRAMFDKYFK
KIGNCVGATSCPQGGQKDSAHYLLSWYYSWGGSLDTSSAWAWRIGS
SSHSGYQNVLAAYALSQVPELQDSPTGVQDWATSFQRLEQLWQ
LQSAEGGIAAGGATNSWKGSYDTPPTQLGSQFYGMYYDWQPVWNPPS
NNWFGFQVWNMERAQLYYVTGDARAEALDKWVPWAIQHTDV
ADNGGQNFQVPSDLEWSQPDPTWTGTYTGNPNLHVQVVSQDVQV
VTAALAKTLMYYAKRSGDATLATAEGGLDALLAHARSDIGIATPEQP
SWDRLDDPVDGSEGLYVPWGWTMPNGDRIEPEGATFLSIRSFYKN
DPLWPQVQEHNLNPQVPAVIPERHRFWAQVEIATAFA

W7S125 Kutzneria sp

TSDYQVFENEKLJEKINKDPNSGYSFRFNGLLLPYHSVETLMVPEAPDYG
HETTSEAFSYYLWLEASYGRITQDWAPFNAAWTSLETAFPSADQQP
TNSGYASKPATAYAEYPSTQPSQSLQGAVGSDPIAGELKSTY
GTTSSYGMHWWFVDDNIVYFEGHCEDGNTTPAFINTQRGQSSESVWET
VTPQSCDMKYYGKNGYLDLFQGDSTSSYAKWQKYTDAPADARAV
QVAYQAEWAKAQGKSSAVADVKKASKMGMDFLYSFDKYFKKI
GNCVGPSTCPAGSGBKDEHYLSWYYAWGSGATDSNAWAWRIGDG
AAHGGYQNPALAALSTGDPLKLSATGSDWATSLGLRQLEQLWQ
QSSEGGIAGGATNSWDGQYGTTPSSGDPDTPFYGMYYDQQPVPWHDPPSN
QWFGFQTFWGMERAVYYQTTKDRACKVLKDQVWPAIANTTVG
AGSFQIPSDDLWTGTAPDTNATSPGNTGLHLHAVKNYTQDVVGAS
LAKTLMMYAAAGSGDTKSRVTVAEGGLTALTAHEDSLGIAVPETRTDYNRFDDTYDAAADQQLVYPPGWGTMPNGDPNINSTFSLRSFYKNDP
DWSKVQSYLNGGPAPTFTYHRFWAQSDIATAFA

**R11FN8 Amycolatopsis vancoresmycina**

TSDYQLAFLTQYNKIKDPNNNGYFRKFGNILVPYHSIETLIVEAPDHGHHETTSEAFSYYLWLEAAYGRVTGDWSPFNQAWTSETYAIPSADQPG
NSGYNASKPATYAAEYPSKPQYPSQLQSGVSVDPIAELKAAYGSADVYGMHWLLDVNIYKFCHCEDGTNTAPAFINTQFRGSQESVWET
VTPQSPCDLKFGKNGNYLDLFTGDSSYAKQWKYTPADAPDADARAVQ
VAFQAEKWAAAOQKSADVSAVVKKASKMDYLYSLFDKYFKKIG
NCVGASSCAAGTGDEHYLISWYYAWGGSMDSSSAWAWRIGDGA
AHQGYQNPLAALADPGKVTSATGAQDWATSGLQLQFLQWLSQSEGGGLAGGATNSWDGQYGTPPSGTPTFYGMFYDYQPVWHDPPSN
QWGFQFTWGERAIYYYATNDARAKKILDKEWPAIANTTVAG
GSQIQPSDLTWSGATPNATSPGANGLHLHVTKVNSQDVGVAASL
AKTLYYASGSSNAQAKTVGQQLTALTANADSKGIAVPERTDYNR
FDDTNTATTQQLQYVPSGWSGTMPNGDPINANSTFSLRSFYKSDQP
WPKVQSYLDGGAAPTFTYHRFWAQSEIATAFA

**T1V3R1 Amycolatopsis mediterranei**

TSDYQLAFLTQYNKIKDPNNNGYFRKFGNILVPYHSIETLIVEAPDHGHHETTSEAFSYYLWLEAAYGRVTGDWSPFNQAWTSETYAIPSADQPG
NSGYNASKPATYAAEYPSKPQYPSQLQSGVSVDPIAELKAAYGSADVYGMHWLLDVNIYKFCHCEDGTNTAPAFINTQFRGSQESVWET
VTPQSPCDLKFGKNGNYLDLFTGDSSYAKQWKYTPADAPDADARAVQ
VAFQAEKWAAAOQKSADVSAVVKKASKMDYLYSLFDKYFKKIG
NCVGASSCAAGTGDEHYLISWYYAWGGSMDSSSAWAWRIGDGA
AHQGYQNPLAALADPGKVTSATGAQDWATSGLQLQFLQWLSQSEGGGLAGGATNSWDGQYGTPPSGTPTFYGMFYDYQPVWHDPPSN
QWGFQFTWGERAIYYYATNDARAKKILDKEWPAIANTTVAG
GSQIQPSDLTWSGATPNATSPGANGLHLHVTKVNSQDVGVAASL
AKTLYYASGSSNAQAKTVGQQLTALTANADSKGIAVPERTDYNR
FDDTNTATTQQLQYVPSGWSGTMPNGDPINANSTFSLRSFYKSDQP
WPKVQSYLDGGAAPTFTYHRFWAQSEIATAFA

**D2B809 Streptosporangium roseum**

DNEYVKRFTTYMNKLDPANGYFSQGVPYHVSVEFMVEAPDHGHE
TTSEAYSYLYWLLEAAYGKVTGDWSRFNDASMEKYIPATADQPT

185
Appendix III

NSFYNPSKPATYAGEWDDIKQYPSKLDDGVSVDPSIANELKTAYGT
NDVYGGMHWWLDVDNTRYGFGRCGDGTTKPAYINTYRGPEESVFETI
PQPSCDTFKHGGKNGYLDLFTGDSSYAKQWKYNAPDARADAVQA
YWAHTWAKEQGKEAQVASSVTAAKMGDLRYYAMDKFYYFQQG
CTSTTCAPAGTGBKDSAYLLSWYYAWGGANDTSAGWAWRIGSSSHN
HSGYQNPMAAWALASSVDAKPKGATAVQDWDSTSLKRLQEFYLWQLS
SEGAIAGGATNSWQGHYAAPPSTLPFTYMGAMAYDQPVYPYHDPPSNQ
WFGFQAWSMERAELYYATGNAADAKLVLDKWKATDNVE
DGTFRIPSTLVWTQDQPDFWNSGPNPAGLHVSIRYTDVSAGVAGSY
AKVLTYYAKSNKTAKAAYAKLGDGLWKNQDAKGVSVPETKADYNRLNDPVYVPPGWTGMPGNGVDSSSSTFSIRSFYKNDWPDK
VDAYLKGTPGPSNYHRFWAQVDVAVALA

D9WN6 Streptomyces himastatinicus

AKTYDARFLELYNKITAPSAGYFSPEGIPYHSVETLIVEAPDQGHETTS
EAYSYLIWLQAMYGKVTGDTWTKFNSAWDIMEKYMIPTHADQTNSF
YNASKPATYAPEWDQPSQYPSKLDDGNPVQGPDIAEKLASYGTDDI
YGMHIQVDNAYGYGNSPKCEAGPSDTGPSYVNTQRGPQVESV
WETVTQPCTDGFKYGGKNGYLDLFTGDASYSKQKFNTADADAR
VVQAYAYWASEWAKAQGKSQISIGNIAAKAKMGDLRYAMDKYF
KKVGNCAGETCPCAGSGKNSYLYLSSWYYAWGGATDSAGWAWRI
GSSHAHGYQNPMAAWALSAYADLKPSATGASDWDSTSLKRQLEF
YRWLQSSSEGAIAGGATNSWQGRYATPPAGKSTFYGMYKEPKYVH
DPPSNQWGFGQAWSMERAELYYNRTGDASAKALVDKWKTVTALSK
TTINPQTYQFPSNLQWSWQPDFWNSASSPGANTGLHVTVDYTNDV
GVAAAYATKLSSYAAKSHITAANKTAKALLDGMWDHHQDALGIA
VPESRADYNRFDPVYVPSGWTGMPGNGDIDSKG5FSIRSFYKND
PAWSKVEAYLKGAVPSFYHRFWAQADIALAMG

D7C16 Streptomyces bingchenggensis

SKAYDARFLDLYNKITAPGAGYFSPEGIPYHSVETLIVEAPDQGHETTS
SEAYSYLIWLQAMYGKVTGDTWTKFNSAWDIMEKYMIPTHADQTNSF
SFYNASKPATYAPEWDLPSQYPAQLGNGNSVGNPDIAEKLASYGT
DIYGMHIQVDNVYGYGNPSGKCEAGPSDTGPSYVNTQRGPQES
WETVTQPTCDGFKYGGKNGYLDLFTGDASYAKWKFNTADADAR
RVQQAYAYWAAEWAKAQGKSQISIGNIAAKAKMGDLRYAMDKYF
YFKKVGNCVGETCAAGSGKNSYLYLSSWYYAWGGATDSAGW
WRIGSASHAhGQNPMAAWALSAYADLKPSATGASDWDSTSLKRQ
LEYFYLQSSSEGAIAGGATNSWQGRYATPPGAAFITFYGMYDEKPV
F3NPZ3 *Streptomyces griseoaurantiacus*

STAYDARFLDLGYKIDTPANGYFSPDGVPYHSVETLIVEAPDHGHETT SEAYSLLWLQAMYGKVTDWDQDFAWMEKYIMPHADQPTN SFDYASKPATYAPEYDTPDEYPSALDTGASVGRDPAAELKAYSAGT DVGMIHWQIDVNDVGYGYNAGPKCEEGGPTTSGPSYINTFQRGPQES VVETVPQPTCDAFKYGGGRGNYLDFTGDSSYSKQWYKTDAKDADA RAVQAAYWADVWAQKQKGDSVTSYGKAAKMDYLYMDK YFKKIGNCVPSSCAAGSGKSHLYLSCWYYAWGGATTDASGWAW RIGSSHVHGYYQNPLAYALSSYADLPKSAQTSAGAWSLDRQELF YRLWQLSDEGAIAGGATNSWQGRYATPPSGTFTYGMYDEAPVYH DPPSNQWFGFQAWSMERVAEYQQTTGDADKAVLKWVDWALSE TTINPDIRTYVSPLQWSGKPDTSNASAPDSGLHVTADYTDVDGVAAAYAKTLTYYAEGDTEAKSTAKALDGMWDHYQDDLAV PETRADYRFEDSVVPGWTGMPNGTDSSSTFASIRSFYKDDPA WSKIESYLKKGAAPVFTYHRFWAQADIALAMG

D6K6C0 *Streptomyces sp*

SKAYDARFLDLGYKIDTPANGYFSPDGVPYHSVETLIVEAPDHGHETT SEAYSLLWLQAMYGKVTDWDQDFAWMEKYIMPHADQPTN SFDYASKPATYAPEYDTPDEYPSALDTGASVGRDPAAELKAYSAGT DVGMIHWQIDVNDVGYGYNAGPKCEEGGPTTSGPSYINTFQRGPQES VVETVPQPTCDAFKYGGGRGNYLDFTGDSSYSKQWYKTDAKDADA RAVQAAYWADVWAQKQKGDSVTSYGKAAKMDYLYMDK YFKKIGNCVPSSCAAGSGKSHLYLSCWYYAWGGATTDASGWAW RIGSSHVHGYYQNPLAYALSSYADLPKSAQTDGAWSLDRQELF YRLWQLSDEGAIAGGATNSWQGRYATPPSGTFTYGMYDEAPVYH DPPSNQWFGFQAWSMERVAEYQQTTGDADKAVLKWVDWALSE TTINPDIRTYVSPLQWSGKPDTSNASAPDSGLHVTADYTDVDGVAAAYAKTLTYYAEGDTEAKSTAKALDGMWDHYQDDLAV PETRADYRFEDSVVPGWTGMPNGTDSSSTFASIRSFYKDDPA WSKIESYLKKGAAPVFTYHRFWAQADIALAMG

M1MJV0 *Streptomyces hygroscopicus*
Appendix III

SKAYDARFLDLGYKITNPANGYFSPEGIPYHSVETLIVEAPDQGHETT
SEAYSYLLWLQAMYGKVTGDWSKFNGAWEIMEKYMIPTHADQPTN
SSYNAKSPATYAPELDTNEPYPAKLDSVSVGPDPIAGELKSAYGTD
DIYGMHVLQDVNVYGVGFNPSGKEACGPTATGSPYINTFQRGQGES
VEWTVPQPTCDAFYKGGRNGYLNLFTGDSSYAKQWKYTDADPDADSR
AVQAAYWDAWVKAQGKSDVSATVAKAAMGDYLRYAMYDK
YFKKIGNCCTSPCACPAGTGKDASHYLLSWYYAWGGATDTSAGWA
RI GSSHVHHGYYQNPLAAYALSSVADLKPKSATGATDWSKGKSLQRQLEFY
QWLQSESEAGAIAAGGATNSWLGRYAAAPPAGASTFYGMYYDWQPYVHD
PPSNQWFGFQAWSMERVAEYQQTGAASAKAILDKVWKAWLSTTT
INPDYTIRIPLQWSQDPDTWNASSPGNSGLHVTADYTVGAVA
AYAKTLTTYAAKSGDAAKSTAKALDDGMWNNYQDSLGIAPET
RTDYSRFGDSVYVYPSGWGTPNGDAINSSSTFASLRSFYKSDPNWS
KIEAYLKGGAAPVFTYHRFWAQADIALAMG

S5UZR1 Streptomyces collinus

SKAYDARFLDLGYKITNPANGYFSPEGIPYHSVETLIVEAPDQGHETT
SEAYSYLLWLQAMYGKVTGDWSKFNGAWEIMEKYMIPTHADQPTN
SSYNAKSPATYAPELDTNEPYPAKLDSVSVGPDPIAGELKSAYGTD
DIYGMHVLQDVNVYGVGFNPSGKEACGPTATGSPYINTFQRGQGES
VEWTVPQPTCDAFYKGGRNGYLNLFTGDSSYAKQWKYTDADPDADSR
AVQAAYWDAWVKAQGKSDVSATVAKAAMGDYLRYAMYDK
YFKKIGNCCTSPCACPAGTGKDASHYLLSWYYAWGGATDTSAGWA
RI GSSHVHHGYYQNPLAAYALSSVADLKPKSATGATDWSKGKSLQRQLEFY
QWLQSESEAGAIAAGGATNSWLGRYAAAPPAGASTFYGMYYDWQPYVHD
PPSNQWFGFQAWSMERVAEYQQTGAASAKAILDKVWKAWLSTTT
INPDYTIRIPLQWSQDPDTWNASSPGNSGLHVTADYTVGAVA
AYAKTLTTYAAKSGDAAKSTAKALDDGMWNNYQDSLGIAPET
RTDYSRFGDSVYVYPSGWGTPNGDAINSSSTFASLRSFYKSDPNWS
KIEAYLKGGAAPVFTYHRFWAQADIALAMG

B5HPK7 Streptomyces sviceus

TKAYDARFLDLGYKITNPANGYFSPEGIPYHSVETLIVEAPDQGHETT
SEAYSYLLWLQAMYGKVTGDWSKFNGAWEIMEKYMIPTHADQPTN
SSYNAKSPATYAPELDTNEPYPAKLDSVSVGPDPIAGELKSAYGTD
DIYGMHVLQDVNVYGVGFNPSGKEACGPTATGSPYINTFQRGQGES
VEWTVPQPTCDAFYKGGRNGYLNLFTGDSSYAKQWKYTDADPDADSR
AVQAAYWDAWVKAQGKSDVSATVAKAAMGDYLRYAMYDK
YFKKIGNCCTSPCACPAGTGKDASHYLLSWYYAWGGATDTSAGWA

188
WRIGSSHAGGYQNPAGALSSYADLKPQKSAQTGQSDWAKSLGRQL
EFYRWLQDQISEAGAIAGCATNWSAGRAYAPPAGKSTFYGMYYDEQPVY
HDPPSNQWGFGQAWSMERYAEYLQQTGNAQAALKAVLDKWVKWALS
KTINPDGTYRIPATLQWSDQPTWNASSPGANGLHVTAVADYNDB
VGAAYAAYAKTLYYAAKSGDTAATKATTAKALLDGMWNYQDSLGIA
VPEEDRTYNDQFDSSVYVPSFSGTMPNGDTINSSSTFASLRSFYKSDP
AWSKIEAYLKKGAVPSFETYHRFWAQADIALAMG

**L7EZA5 Streptomyces turgidiscabies**

SKVYDARFLDLGYGKITNPAGSYFYSGIPYHVSSTIVLLEAPDQGGHTTS
EAYSYLLWLQAMYGKVTGWDWTFAWSEMEYQYMPITHADQPTNS
FYNASKPATYAPEWDFSQYPADLDTGVSIGDPTIAEKLKAYGTDD
VGMMHQWLDQDVNYQGNYSPGKCEAGPSTDGQSY_INITFQRGQESV
WETVPQPTCDAFNGGTNGYLDFDASYAKQWKFNTNDADNADAR
AVQAAYWADVQAKGQAGDREVTVGKAAMGDDYLYRSMDYK
FKKIGNCVDPTAACAAGTKDGAYLMSWWYAYAGGATDTSAGAWA
RIGSSHTGYQNPAGALLSSADLKPQKSAQTGQADWSKLSGRQIEF
YRIVALQNSMEAGATNQRAWAYGTPAGTFYGMUYDEKPVYH
DPPSNQWGFGQAWSMERAQEYYQQTGNAAKSVLDKWVDWALSK
TSINPGTYQIPSTLQWSGAPDTWNATTPGANGLHVTAVADYNDB
VGAAYAAYAKTLYYADRSGDTEAATTAKALLDGMWNYQDTLGIAV
PETERDYNRFDDAVYVPSGTGMPNGDTSNASSTFASLSFYKNDP
NWSKIEAYLLAGAAPSTFYHRFWAQADIALAMG

**K4QTE6 Streptomyces davawensis**

SGEYDARFLDLGYGKITNPAPANGYGYSPGYHVSSTIVLLEAPDQGGHTTS
EAYSYLLWLQAMYGKVTGWDWTFAWSEMEYQYMPITHADQPTNS
FYNASKPATYAPELDTQNEPARYALDTGVSIGDPTIAEKLKAYGTDD
VGMMHQWLDQDVNYQGNYSPGKCEAGPSTDGQSY_INITFQRGQESV
WETVPQPTCDAFNGGTNGYLDFDASYAKQWKFNTNDADNADAR
AVQAAYWADVQAKGQAGDREVTVGKAAMGDDYLYRSMDYK
FKKIGNCVDPTAACAAGTKDGAYLMSWWYAYAGGATDTSAGAWA
RIGSSHTGYQNPAGALLSSADLKPQKSAQTGQADWSKLSGRQIEF
YRIVALQNSMEAGATNQRAWAYGTPAGTFYGMUYDEKPVYH
DPPSNQWGFGQAWSMERAQEYYQQTGNAAKSVLDKWVDWALSK
TSINPGTYQIPSTLQWSGAPDTWNATTPGANGLHVTAVADYNDB
VGAAYAAYAKTLYYADRSGDTEAATTAKALLDGMWNYQDTLGIAV
PETERDYNRFDDAVYVPSGTGMPNGDTSNASSTFASLSFYKNDP
NWSKIEAYLLAGAAPSTFYHRFWAQADIALAMG
**L1KHJ0 Streptomyces ipomoeae**

SKTYDARFLELYGKITNPANGYFSPEGIPYHSvetliveapdhghetts
eayssyllwlqamykvtgdwskfnawemexkymipthadqptns
fytankpataypehdpgeyqaqlntgvsvgsdpaaelksaygtdd
iygmvhwlqvdvnygygngspkgceagptdgtgsynftqrgpqeysv
wetipqptcdqfkyggkgnyylslftgdasayakqwkftnapdadar
avqaaywadiwakqgkgsdvsatigkaakmgylryamdkyf
krigncvgsccapgtgdshylyswyyawggatdtsgawawri
gssthggyqnlplaayalanyaplkpksttgqadwakslrdqiefy
rwlqsneqgiaggatnwswagryatppagtptfgyngfdekvypvhydp
psnqwfqfwqawsmervaeyyyqttgnaaaktvlkdkwvdwalsktt
inpdtystripqtwsgapdtwnasspganallhvtvadtytdvgv
aaaayakltlyyadrgsdadaaarvakaallldgmwdhhqdlgialavp
etradynrfddrvyvpsgwtgmpngdainsstfsirsfyeddpa
wskieaylaggaapsftyhrfwaqadialamg

**M3ECC0 Streptomyces bottropensis**

SSTYDREFLEMYGKITNPANGYFSPEGIPYHSvetliveapdhghetts
eayssyllwlqamykvtgdwskfnawemexkymipthadqptts
fytankpataypehdpneypaqqlntgvsvgpdpiaaelktygtdd
iygmvhwlqvdvnygygngspkgceagpadtgsynftqrgpqesv
wetipqptcdqfkyggkgnyylslftgdasayakqwkftnapdada
ravqaaywadkwakaqgkgsdvsatigkaakmgylryamdkyf
ykkiigncvgsccapgtgkdshyllswyyawggatdtgaagwsy
rigsshttggyqnlplaayalanyaplkpksttgqadwakslrdqief
rwlqsneqgiaggatnwswagryatppagtptfgyngfdekvypvhy
ppsnnqwieqfqaqawsmervaeyyyqqttgnaaaktvlkdkwvdwalsk
ttvnpdgyripqtwsgapdtwsasspganallhvtvadtytdvgv
gvaaayakltlyyadrgsdadaaarvakaallldgmwdhhqdlgialavp
vpetradynrfddrvyvpsgwtgmpngdainsstfsirsfyeddpa
wskieaylaggaapsftyhrfwaqadialamg

**B5HJV6 Streptomyces pristinaespiralis**

SKEYDGRFLELYGKITDPANGYFSPEGIPYHSvetliveapdhghetts
eayssyllwlqamygritgdwtkfngawemekymipthadqatgsf
ydpnkpatyapehdqpsqygapelqpsvtsgdpiaaelksaygtddv
YGMHWLQDVDNVYGYGNEPGKCEAGPSATGPSYINTFQRGQPESV
WETVPQPTCDRFAYGGTNGYLDLFTKDASYAKQWKYTNAPEADAR
AVQAAYWADLWAKEQKGGSQVSGTVAKAAKMGGDLRYAMFDKY
FKKVGNCVGTCPAGTGDSSQHLLSWYYAWGAADTSAGWAW
RIGSSHAHGGYQNPLAAYSLAYAPLKPSATAQDDWAKSLDRQIEF
YRLQADEGAIAAGGATNSVGGRYEAPAAAGTPTFYGMAYDEKPYHV
DPPSNQWFGFQAWSMERVAEYYQQTGDAQAKEVLDKWVDWALSE
TTVNPDTFRIPSLQWSGKPTDWNAAANPNGANAGLHVTWADYTDQDV
GVAGAYAKVLTYYAARSGDTEAASVAKALLDGMWDHDDALGIA
VPETRTDYRSFDDPVVYPSGWGTMPNGDSSSTFASLRSFYQDDDP
AWSKIESYLAGGAAPAETFYHRFWAQADIALAMG

**M3C0N9** *Streptomyces gancidicus*

AKTYESRFLELYDKITDPANGYFSPEGIPYHSVETLIVEAPDHGHETTS
EAYSYLLLWLQMANYGRITGDWTRFNDAWATMERYAIPTHADQPTTSF
YPDPSKPATYAPEHDTPEYPSQLDSGVSVGRDPIAELKSAAYGTDDV
YGMHWIQDVDNVYGYGNSPKCEAGPSDTGPSYINTFQRGQPQESVW
ETVTRPTCDACKYGRNGLYLDTFKDASYARQWKFTNADAR
VQAAYWADLWAKEQKKGGEVAGTVAKAAKMGGDLRYAMFDKWY
FKKIGNCTSTSCAPAGTDKASHYOUSSLWSYAWAGATDSASWSWRIGS
SHAHGGYQNPLAAYALATYAPLKPSSATGADAWKSYDRQLEFYR
WLQSDEGAIAGGATNSWAGRYTTPPSGTPTFYGMYYDEKPYHVHDDPP
SNQWFGFQAWSMERVAEYYQQTGNAQAKAVLDKWVDWALSCTTV
NPDSFRIPSTOPSWGATDWNASSPGANRGLHVEVVDYTDNVGVA
GSYAKVLTYYAARSGDTEAADTAKALLDGMWANNQDDLGIAYPET
RTDYQRFDPPVHVPSSGWGTMPNGDSSSTFASLRSFYQDDDPAWS
KIESYLLEGGSAPFVFTYHRFWAQADIATAMG

**D9XJA9** *Streptomyces griseoflavus*

EKTYDARFLELYGKITPNANGYFSPEGIPYHSVETLIVEAPDHGHETTS
EAYSYOLLWLQMANYGRKVTGWDSDKNGAWDIMEKFMIPTHADQPTNS
YPNASKPATYAPEHDTPEYPSQLDPSGVSVGPDPIAELKSAAYGTDD
VYGMHWIQDVDNVYGYGNSPKCEAGPSDTGPSYINTFQRGQPQESVW
ETVTRPTCDACKYGRNGLYLDTFKDASYARQWKFTNADAR
AVQAAYWADLWAKQKGKADVSATVAKAAKMGGDLRYAMYDK
YFKKIGNCTSTSCAPAGTDKASHYOUSSLWSYAWAGATDTSASWSWRIGS
SSHAHGGYQNPLAAYSLANYAPLKPSATGADAWKSMQRFLEF
YRLQADEGIAAGGATNSWAGRYTTPPSGTPTFYGMHYDEKPYHV
DPPSNQWFGFQAWSMERVAEYYQQTGNAALAKSVLDKWVDWALSSET
Resurrected exoglucanase sequence

DNAYDQRFLTMNKIKDPANGYFSPEGVPYHSVETLIVEAPDHGH
ETTSEAFSYLYLWEAMYGRVTGDWSPFNNAWDTMEKYIIPSHAD
QPTNSSYNPSKPATYAPEHPDSPQYPSQLDSSVPVGQDPIANELKS
TYGTNDIYGMHWLDDLVDNVYGFNGSPRCEGDSTTPAYINTF
QRGPQESVWETVPQPSCDTFKYGGPGYLDLFTGDSYPAKQWK
YTNAPDADARAVQAAYWAHEWAKEQKASEVAATVAAAKMG
DYLRRSMFDFKYFKKGNVCVGASSCPAGTGDASHYLLSWYAW
GGATDTSAGWAWRIGSHSHFGYQNPMAAWALSNAELKPKSP
TGASDWATSLKRQLEFYQWLQSEGGIGAGATNSWNGSYATPPA
GPTFTYGMYSDEQPVYHDPPSNQWWFGMQAWSMERAEEYYAT
GDARAKAVLKDWPQAIANTTINUADGIFSQIPSDEWSQPDW
ASSPGANTNLHVTVTNYQDVGVAGSLAKTLTYAASKGNTTAK
DTAKGGLDNLNYQDSKGISVPETRADYNYRFDGVVYVPPGWTG
MPNGDVIKSGATFDERSFYKNDPDWPKEAYLNGGPAPFTTYH
RFWAQVDIATAYARSCC
Appendix III

List of beta-glucosidase proteins from the species used in the construction of the phylogenetic tree

U5DQJ4 *Rhodococcus equi*

MTMRQVSLTSGADFHTTHPPVPGLPAIMLTDGPHGVRKQAATAA
GYSPESVTPATCFPTASALATATWDLVEVEGVVALGTEARTEGVSVLHGAPNIKRSLACGRNFYFSEDPFLLSSRMMAAWILGVQSTVGASLKHFAVNNQEFRRVSDAVVDARALREIYLAGEFHEAVDVAPATVMAAYNVGGTHCAENRWLLTDVLREQWGDGVLVSDWGVAVTRRSRCL
AAGLDLEMPSYGGGLDDVLAAVGAGKLDIAAIAREAVETVEGVDARLKL
LPEGHAADIAAYEVCRRPVVLANGAPVTMPPHVDDVAAIVECYLG
QGAAGSIAARIILGDAEPPGRLAESPLHTSDNPVHVWPAGPSVVEY
RESIYVGYYDAEELDVFRYFGHGLSISFAWTELAVDFDSTSE
DIEQRLDVSVRVTNTGDPRGSEVVQVYVRDVESTVFPRPDQELAAFAK
VFAPGESRRVTLHVDRAFSFDWTITTDDWIESGDEIRVGASSRDIR
QSATVTLTSRSGFDAGPLAYHGSPVFERAAPAEYVGKPLPNDVDA
PRHYVDTPLADIRHPAAAALLTRAMRRKVAATAPALDEDDPSRLIE
RSLQELPPMRPMLMTQGGVTPAAAQAFVICNGHTVRGGWALVAAL
RRK

F5XL24 *Microlunatus phosphovorus*

MAVGLDIPRLGTLAEEKASLVSGLSFVFTQPFVLGPAIMVSDGPH
HGLRAOPPGGDHVGLGSSLPTATCFPTASIALASAWNPELLHRIGQAL
AQEARACNLVILGSGINMRKSLPCGRNFYFSEDPLAGELVAGIVGD
GIQSCGVGTSAKHW.AANNQETDRLRCDSSQVDERTLSRETPAFARVV
EKSPWTMCSYKNKVTSAESENTWLLDTLVEEFFGLVSDWGV
AVYHPVPAVQACDLMDDPKGRSEAAIYVIAVGSGLSDVDARV
RTVLELVAKGMHAELDESFDIDAHHALARQAAAESVVLKNDGLL
PLTVEANIAVGIFARTPRYQAGSSQVVPTRLDTVLEELQTVYELPP
FAAAYVGDSNDAVLLVEAEQVAAADIVMILGLPAEESEGFD
RTHLNLDPNLQTLAAVAEANPVVVVLARGTSTVLGDLVRASAL
AEAWLGGQAAGGGIVDVLGAVNPSSGRLAETIPNLDSSYLNFGP
EEVSYYRGEGIFYRGYDRQHLDVAFPGFGLYSFESLSDKVRMR

193
**Appendix III**

GSVADETLGATVEVTNTGDVDGSEVVQVYVQDVSTVVARPVQEL
KGFKAVAPAGAVQVSIELDQRAFSYWSPRYRRWVEAGDFEISV
GSSSRDLALSQSVTVEAAATLPLTRLDSLQEWLADPAGRQQLVERE
VAAAGPAGAMQEGLLEVIG
NFPNALANLNLSDLHDSLDLVARAEEWAGQQR

**F6FVZ5 Isoptericola variabilis**

MTTDTPAPGPTAPTALTDEVPRLVAELTLEEKASLCGQDFWHTQAV
ERLGIPAVMVTDPHESTKQAGATDHVLNLQESVPATCFPPAAGLGST
WDPELWQGEGAEILGENAVLQLPQGAVNMKRSLPGCRFYELAA
EDPFLAGELAALVEGIQSKGVTSLKHFHANQETDRMRVQDAXQV
ERTILEYLFRAFEKVTQAPQWTVMAYNKVNTSASHQWPWTLLE
LREDEWGKEGLVSDDWAVVDRAGVKAELDELSPSSGLNDRARV
EAVRSCHLSEADRUTLVTRVLQVARSQAAALAPEDEFAEAIHHA
QEAATRAAVLKNENGVLPSGDALGDVGGVGEAMARTPRYQGAGS
SQVNPRTLVALSALDAERGLDVPFPYRPYALPAAGKGQQDHPPDDEL
RAAEVEAAGTAVVFLGPLQAIDESEGYDREHMLPSHTALLEV
AAERVIVVLSNGTAIVGQWDQADAILETWLGQAGSATTAVL
LGDAAPSRLAESIPVRLEDVPAQLNFPENGVYRGEFIGYFIRGLD
ATRAEVSYPFHGHTTFGFSGLADVVAEVETQATADVVVRVVAL
TVNTREGEGAVAVQYVDRPASTVARAPRELRGFARVSLAPGASER
VELALTNRDLSHWDTLTMAHVEPGAPELAVAGSRLDPKATVE
AAPALPRPLHRYSTIGEWRDQPEAWAALREKLGGFALFEDSPDPA
MAAFLVEMPVIKVPMGGMASLSIDDFTLLARFGRP

**U1GC79 Propionibacterium avidum**

MTPAEITDLISQLTLEEASLCGQDSWHQLQTVTVERLGIPGMVTDPH
GLRKVADGSMAGIYDSVPATCFPTAAGLASTWDPELVHDIGVALGEE
TRAESVSVLPGINMKRSLPGCRFYFESDVPLAGTLATLEEVRGIG
SQGVTSLKHFAANQQETDLRNLVNEAIDERTLREILPAFEMVVKQA
DPWTVMCSYANGVYSSQNRWWLTTLLHEEWGYKGLVISDWGAV
VNRVEGLRAGDLDEMPGDAPRNDARIVEAVERGLDDLEELDTAVARI
LSLARASAAMDPGYSIEHHQLARRAAASVLLKNDGVLPLID
DPDDVVVVGIEFARTPYQGAGSSKNPRTVDTALSDLRTQWDVDPF
APGFTLTDHPDQPLADEAVSLRSRGKTAVFLGLPAPAEGYDRTNT
DLPTQIDLLARVYKASHASTVTLLNSGSSVTAEWDQADAIWECW
GGQASGSGVDDVTGLKNSGPGLAECTRKLDSIPALNLNPFGFQHV
TYEGRYIGYRGLDATERVAYPFHGHLSTYTFAYCDLDDLVRPVSDE
TAMQETTVLTFTTVSNTGDRAVATPVQYLVGFPDATVDCVRELKA
FLRVELDPGESRSTIDLTTRDLSDYWDILLSHWTVEPGTLRVEGSSR

194
S3WX10 *Propionibacterium* sp

MTPAEITDLISQLTLEEKASLCSGGD M2XAG3 *Rhodococcus qingshengii*

LTDVDPDDVVLIEEAFILAMPPAYKMLQMAPIMTPPEELDMLGD

M2XAG3 *Rhodococcus qingshengii*

MTENYVDGLSLEEKAASLGDSDFWHSQVPGIESILLTDGPHGVRKQP

EGGDALGLGHSIPATCFPPAVGLGSSWNLIDLRIQVGEALGDEAAKEQ

VSVLLGPGINIKRSPLCGRNEYVSEDPFLSGRVAALITGIQSRGVGT

SLKHFAANNQEHDRMVRVSADVERTLREIYLAGFEYAVKTAAPTTV

MCSYNKINGVYSQNHWLLTEVRLREQWGDGLVSDWGVANDRVA

ALAAGLDLEMPPTGTDTQVDAVRGGDLDRESVLTATAERLTLVART

AAARTEGHTYDVERHHELARAAAESAVLLANDGLLPLTPGGQTV

AVIGEFACSPRYQAGSSQVVTPLDNALDAILDREGADVTFAPGF

TFDGTPDDDMVTEAVADAARADVAVLGLPSATESEGFDRTDIEPL

ADQIALLEAVHGANPNTVVVLANGTVVSIEPWKDAAAILEGWLGG

QAGGSAIADLLFGRTETPHRLQDNSYLHPGSQHQVRYG

EGLYVGYRYYDSALREVAYPFGGLSYTTFDTRIDTSVEAGENSAEV

VTVRNSGDRSGTVQVYVHADASASIDRPQAELKGFAKVLDPDES

ATVTITLTDRAFYWSVAQDWAIEPGYEIRVGFSRDIATTDTI

AGNIVGVTLDAMSTIGEWLASHPVGAVLGAAMAAAGDGACAVSP

EMMALAGSMPLGKLATFLGIGEEQVQVLAAAAQPAS

T1VRJ4 *Rhodococcus erythropolis*

195
Appendix III

MTENYVDGLSLEEEKASLTSGSFWSQSVQVAGIESIILTGDHPGVRKQPEGGDA
GLGHSIPATCFPPAVGLSSWNLDLIRQVEALGDEA
QVS
VLLGPGINIKRSPLCRGNFEYVESDPFLSGRVA
LITGIQSRGV
TSLKHFAANNQEHDMRMSADVINTERLREIYLA
FEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RGID
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVNDRA
AALAGLDLEMPPTGTQIVDARVGGDLD
VTTAARL
RTAARTEGHTYDVERHELARTAAAESAVL
LGDDELPLTPGQ
TVAVIGEFACSPRYQAGSSQVTPKLDN
ALDI
LEGADRVVFAPGFTGDGPNDDMVTEAVDAARRADVAVLFLGLPS
ATESEGFRDRTDIE
LDQAIAL
EAVHGANPNTTVVLANG
VSTP
KD
AISLVLVA
REIYLAGF
EEYAVKTAAPTT
VMSYNKINGVYSSQNHWLLTEVLREQWGDGLV
VDWAVN

E8MQE8 Bifidobacterium longum

MEEPRTTARQSGRIGANAYRTAQLKCLKERGIMSESTYPSV
EKASLTSGSWSHQLQVESKIGPYMITDGPGLRKSASSTGETD
LNNSVPATCFPPAAAGLSSWNPELIHKVEAMAEECIQEKVAVLPG
VNIRKRNPLGRCFEYWSEDYPYLAGHEAVGIVAGVQSKGVGTLSKHF
**W6F3F9 Bifidobacterium breve**

MSESTYPSVKDLTLEEKASLTSGGDAWHLQGVESKGIPGMYITDGGPH GLRKSASSTGTDLDDSDPATCFPPAAGLSSWNPHELHIKGEAMA EECIQEKAVAVILGPVNIKNNLPGCFFEYWEYSEDYPYAGHEAIGVEG VQSKGVTSKLHFAANQNEDTDLRVDARISPRALREIYFPAFEHIVKK AQPWITMCYRINDVQAQNHWLLTDVLREDWGFIVMSDWGA DHARDQSLANGLNLEMPSYTDDQIVYAVRDGLITPAQLDRMAGQM IDLVNKTRAAMISIDYRFVDAHDEVAHQAIESIVMLKNNDAILPL NAGPVALNPATPQKIAVIGEFARTPRYQGGSSHIPTKMTSFDLTILA ERGIKADFAPGFDTDLEDAPDALESEAVETAKNADVLMFLGPLEAA ESEGFDRDLDMPAKQITLLEQVAAANQVVVVLNSWGVITPAWA KNAKISLWLLGQSSGPPALADVIFGVQVSPSKLAQSPIPLINDDPSM LNWPGEHGYGEGVGYRYYTDYKAVDPFGYGLSYAITEIT GVAATGANTATNVANTVTNSTDVAAETVQYVVPGBKADVPK HELKGFKVLKSEGKTVTIDLDEREAFAYWSEKYNHVEAGEYA IEVGVSRRD IADTVAYALDGDGKTQPLTEWSTYGEWEADPFAGKIVAAAVAAAGEA GELPKLDPNNMRMFLNSMPINSLPLTGEGKKIAQFMVDEYTKLS K

**J9XU17 Bifidobacterium pseudocatenulatum**

MSEKTYPSINDTLTLEEKASLTSGGDAWHLQGVEAKGIPGMYITDGGPH GLRKSNSATTGEVDLNNCPATCFPPAAGLSSWNPHELHIVGEAMA EECIQEKAVAVILGPVNIKNNLPGCFFEYWEYSEDYPYAGHEAIGVEG VQSKGVTSKLHFAANQNEDTDLRVDARISPRALREIYFPAFEHIVKK AQPWITMCYRINDVQAQNHWLLTDVLREDWGFIVMSDWGA
Appendix III

ADHDRVASLNAGLNLEMPSPYTDQIVYAARDGRIPQAQLDRMAQG MIDLVKTRAMSIEYRFDEIAHDEVAHQAAIESMVLLKNDAILPI AGDADKVTIGEFARTPRYQQGSSHITPMTSFLDLTERGVDKA FPGFTLDLEPALEAEAVDAKAGADVVMLGLPEAAESEGDRE TLDMPAKQIALLEAAVENKNVVLNSGVSTVTVPWAKNAKILE SWLLEGQGGPAALDVFGKVPSPGLAQIPAITYPDINDDSPINWPGEEG HVDDGEGFVGFYRTYNKAVDYPGFGGLSYATFESVDTAKVT GACTASVSAVKNASNDAETVQYVAPGKADVVRPKHELKGFK KVFLKAEGSEAEVSFELDDRAFAYWESRFNDWHVESGEYTGIEVTSSR DIAAGASVLEGDAKQPLEWNSFMWEWRLKDPSVGKVEILRAELE AGRMPVVPDNDMTRLFDSLMPINSMSVLMGADGKQIFEYMLEKYAE LTK

W4NB76 Bifidobacterium moukalabense

MSESTYPEVNDLTLEKASLTSGGDAWHLQVESKIGPGYMITDGP GLRKLASSTGETDLNSVPATCFPPAAGLSSSWNPHELIQVGEAMA EECIQEKAVILGPVGNIKNPLGRCEFYWSEDPYLAGHEAIGIVSG VQSNGVTSLKHFAANNQETDLRVSANISQRALREIFYFAFEHIVK EAPQWTMCSYNRINGVHSANQHWLTDVLRDEWGFEGIVMSWDGA DHDRVASLNAAGNLLEMPSPYTDQIVYAARDGRIPQAOQLDRMAQGM IDLVDKTRAAMSVEGYRFVDAHDENVQAAVESVMVLLKDAILP VSSDAKAVIGEFARTPRYQQGSSHITPMTSRMTGFLELTDARGVDR FAPGFTLDLEPAAMAASAEVETAKGADVVMLGLPAAESEGFD RETLDIPAKQIELLEAAAVENRNIVVVLNSGSVSVAPWAANAKILE SWLLLGQAGGPAADVIFGHHVSPSGLAQTPMDINDDPSMINWPGE GHVDYEGEVGFVGYRVYDTYGKAVDYPFGYGLSATFESVDSVKVAR TGDNTASVSAVKNSVDVAETVQYVAPGKASAVPVLHELKGF RKVFLKAGESAEVSFDLDERAFAYWSEKFNGWHVESGAYTVEGTS SREDIAGTG VELDGDGKSEPLESTEWSTFGEWSEDPTGSKIVASVYAEGEAGNLPKMP DNDMMRMFLKSPINSMPNLMSGGKKITAFMLDEYAKVAK

D2Q5N4 Bifidobacterium dentium

MSESTYPEVNDLTLEKASLTSGGDAWHLQVEAKGIPGYMITDGP GLRKLASSTGETDLNSVPATCFPPAAGLSSSWNPHELIQVGEAMA EECIQEKAVILGPVGNIKNPLGRCEFYWSEDPYLAGHEAIGIVSG VQSNGVTSLKHFAANNQETDLRVSANISQRALREIFYFAFEHIVK EAPQWTMCSYNRINGVHSANQHWLTDVLRDEWGFEGIVMSWDGA DHDRVASLNAAGNLLEMPSPYTDQIVYAARDGRIPQAOQLDRMAQGM IDLVDKTRAAMSVEGYRFVDAHDENVQAAVESVMVLLKDAILP VSSDAKAVIGEFARTPRYQQGSSHITPMTSRMTGFLELTDARGVDR FAPGFTLDLEPAAMAASAEVETAKGADVVMLGLPAAESEGFD
R5T443 Clostridium hathewayi

MTDKIKDLVTRMTAMEEKALLCSGKGNFWQMEGIERLGPSVMTDGPHGLRKGAGEADHLGLNQSVKATCFPAPVTSASSWDKAALYDMGQAI
GEECVQEVEAVVAVGPTNISRSLPGCRNFLFSEDPYLAGEMAAWie
SGVQSKGITSKHFAMANNQKARLVSNSVDERALREIYLAPFEKAV
KQAQPWTVMCSYRNQGSYSCENEWLLTEVRNEWFGQFLVMTD
WGMNDRVKALKAGELEMPGDPYPNDDKKIVDAVRNGELDEAVLD
RAAERLTVIMRAGEVHKKEYDAAMAHNLARIAAESVALGNDG
MPLKDESCRIPTION

R7K9E6 Acidaminococcus sp

MDNKENSQIQTADNAVANVPDFDEILKQMTLEEKLCSGKTFWLT
KEIKRLGVPSVLMDPGNTRLKKEAKGRNTMISEPATCFPTAVT
LCSTWDPSEAKAIADAEQEGQSTVLPGVNIRKSLPGCRNFLFSEDPYLAGEMCHGQWSQENIVGSLHYCANNQEHIMSVTIA
DERALREIYLAPSEKQDEQRTTVMASYRNKLYCDLNMRLTEV
LREDWGFKGIVWVAVDGEVGKAGLDEMGNGINDKLIIVE
AVRNGTLDEADLDKVLARMKFAECKAKEVENHADFEAHTLAR
EIGAGAVLLKKNEENLIVKSGEKIAVGLAKVPVRQGAGSNNP
YKPKPVFSFIALSNANREYTAADGVLTGKNGYDSLLKAKVIAKADK
VLFLIGLTDSEYESEFDRKHLMNPNGHEILFDAQVNSNVARVFSG
SPVDREIAPAKGGLLCAYLGGQAVGEAVMDVIFGVDVPSGKLAET
WPLRLHDNIAKSVFPMGTKVEYRESVYVGYRFDTDVRDFMPFG
YGLSYTTFEYSDLKSLKEFKKIDTSVEISFKIKNTGSIDGAEVAQVYIS
Appendix III

R6PJW Clostridium sp

MKNPKKEILAQMTLEDKAALCDGADFWHLKMGKMKGPSIMVCDGPH
GLRKKDKYNKTSSLSSCVPSICFPATATTACSWDPDLHEMGVALGK
KLKVEEVTLLPGPNKMFSLCGRNFYFEDPVLAGEAILAAGFIEG
VQSMGVGTSIKHCANSQETRRMTCSVVDERALREIYTAEIAVAK
KAPWNTVMSYNKINGAHGSENKHTQIEIRLDEWGFMDVVSDDW
AVNNRVLKNGNDLEPSASGAKKIVEAUKNGELDEAVTNER
ALNVLNLIKAADGAKPGYEYNHADDDQPLARKIAQGSMVLLKNDG
LPLKKEGKIAVIGDFAKFRHQGQAGSSQINPTKMDAYDELKELYG
VEFVQYQYSAKAAKNAAHIDAAAAALAAKCDAAIVFGVLTDDYES
EGFDRTTHMTLPEAHNKLVEEIVKNVNVNFLAGSPELWEDNSVR
AVLNSLQGGQAGAGAVDISSKVNPSGKLAETPVYYSIDPAVNPN
PGNAPATVEYRESVYIGYRYEYAKANGAVRYPFGFGLSYTTFYGYSDIKL
DKSEAMDITDLSKFGKNTGSVGAGIAQVYVSDKVTIIYRPPVKE
LKGFKWVLEPGGEKEVTVELCRAAFAYNVKINDWCVESGFEFILVG
SSSADIRLS
ACVNVNAPDNPDNYTYTGVQNVPAQFEAVLGRPLPPT
V KRKDIPIYVTNDNENAAHKTNKAKLYNTLKLVEPFAQAIALQTQP
FRDFISMSGSGSEDMAAGLLKILNEKGGVRAILKCVPKAIKGIGPL
LKNI

R5XIE Anaerotruncus sp

MKHPEIVSKMSQAKFVSVGDYWHOLEEAPGLPKIMITDGPHGL
RKQNTDKKASSGILQGNSVPSATCMPPAATSACSDLNWENLRRESSGAL
EEELQEKVSVILPGTNKSPRCGVRNFYESFDPLAGKMSASLING
CQSKGTVGNSLKFHCNSQEQAFRMVLSVETERTMREYFPAFIAVKE
SPQWTMVNSYRNLVQAQNEWLQEKVLREDWFGKGLLVTDW
ASVDRVPGLKYGTLEMPSSPGPLNTKRIAAVNGEDEAILDKRV
NVDDLILSKPALKQNPQNGYKFDVEAHHALSRIAGEMAIVLVKNDKIK
LPLKKGQKIAVIGEMAKSPRFQAGGSSVINPTKLANYDELYKLGD
VTAYAQQYISAPGGKDKDRSDAQSLVEAVAAKAADAVVFVGL
TREFEFEGYDRENINMPHINNGKLVSEIAKANANTTVVLAGGSSVYI
WLDEVKGLSNGLGQQAISVANLTTGAIVNSPGKTAEETVVPKYE
NPTFIGNYPPGIPVEHKESVYIGYRYTDAEKELPFPGYGLSYTTF
YSMDKLSASDIKDTDKLVLKSKNTGDVGAEIVQIVYADKESTIF
RPKKELRAFTKVLKAGEEKEITLELGKRAFAYNVKLGKDWHVESG
EFEIÅAASSRDEKLKASVNYTSTVEAEVDPYDRDIAPSYTADIKVD
DKQWGAVYGSELPARERDKNAKIDLYCCLNDARHGTKWKGGKLRCRIE
KIMSNGMSAENGDGKMLAAMATQIPIRNFVQMSMGVSPKMAEGLL
KMLNDDESSFVGFNAIFWRLGGALTRLPSLKKSI

R5N0X1 Eubacterium sp

MKHEKEVEKMSLEQKAAVFSGYDYWHLEEAPELGPKIMITDGPHEL
RKNAPDKKSSSTGGILGNSVPATCFPAPATSSCSWDPHELLEQEGAM
GEECLKEKVSTILGPNTIKRAPVGGNRFEYFSEDPLLAGECAAVIN
GVQSKGVTSLKHFAANSQEARFMMVNYEVDERTLETYLTAIFEIA
VKKQAPWTPMNYSNINGVYASENEWLQKVLKEWGFEGLIVTD
WGASVDRPLGKATGDLMPCSMLNTNIIAAVKGDTLDEKILDER
VDTVVDLVKSFLPALEKTHVVDAHHAIQKIAEGSMLKNDGGI
LPLKGDQKVAVIGEMAKAPRFGQAGSSVINPTKLNASFDLQKLGVD
ISYAGQGYIKSAPKDKTPRKTSAELIAEAEASKADAVAVVFVLGL
EEFEGEGYDREGIEPAEHENLAVAAEANPNTVVVIAGGSVILMPW
KNVKGLNSLGGQAGGIAVANLTGKNPSGKTSETYPEAFIPDNT
YGNFPGPVPVSETHEBRYIGRYYDAADIDVEFPGFGLSYTTFEYSD
IKLSKKDKMDTDTVTFKIKNTGSDGAEIAYVADKESTIFRPPP
ELRGFKKVFLKAGEEKEKVESLNFVAYNVEYGDWVETGDFIDIL
VGSSSRDJKLTSMTVSFTASIPDRYKAPNNYNVANITRDFIAA
VYGELPNPEIDKNKIDLYCCLNDARHGTKWKGGKLRCRIEKIMSMGMS
DANGDGKMLAAMATQIPYVRNFISMSMGAFSPKAEGLLKMLNDDES
SFVGFNTIFWRLGGALTSLPKLKKSI

K8E314 Carnobacterium maltaromaticum

MKYQSLIEQMTLEKEASLSMGENFWNTKAIERLNIPSIMTDGPHLR
KQGGKADHLGLNASLPACTYAATLANSWDRELLNEIQGFLEEAC
VSENVILPGPGLKNPLCGRFNFYFSEPDYLGASQQVKGIQS
QGVAASPKHFAVNSQEHMRMTIDDEVDRSRELLEYLEGFRVVKQS
KPTKIMSYNQINGIYAEHPHLMNDILYGEWQFVDMVTDGWGN
DRVAGLRAGNQLMPSTNGTDQIEVVAIQTEGELSEAILDQANQQL
KLVFETSSVTIDQYKHKEAVDAKRMVLLKNOQIEPLRA
SEKIALIDFANKPRYQGAGSSLNPTQVPNFVELVKESPLSIIQYQAQG
YQRMMQVRVKTKLNEAIELAKKVDKLFFVGLDESREAGIDRHDL
KLPFNQNLHLEIKNVPNVVILAGGVLLEPFERRVQGLIYSLAG
QGVVAEALKEILLGTNSPGKSLSETYPLAYESVPSATYYPGKEATSEHR
EGLFVGYRIFYDFTANPVKYFSFGYGLSYSYTKFAYSITREDLSVFTVSN
SGKVAEEEVAQLYIEQKESIIKRAKRELKGEKJYLEPGQSKVTVIQLT
EHDFSYYSLVNQDWAIEAGNYNIQIQSSIEDIRLLTMIEQVGGKNEEYY

201
Appendix III

QLKKYPTYAQANVQRPDAEFYQULLYQPAPPALLWDWKPKPLGMNDTI
AQARNNKWLGMSSSYINIVFKNVLKIVKQPLWSNNYVFVNMPFRQ
MERFTGKISKKNVQRYLWVNG
KN

E0RVH9  Butyrivibrio proteoclasticus

MKYQDIIEKMTIEEKAFLSGKEQWQDRFLGISIFSDPGPHIRK
QAGAGDHGLNESCVPATCFPTAACATIANSWNEELEGEELGLKMGEA
AEGVNVLLPGPNIKSPLCGRNFYFEDPYLAGKMAASYVKGQS
QGVYACPKHFAVNSQELRMAMNSVLDERTLREIYLTGEIEAVKEGK
AKTISAYNENGYTANENKHLLTDILRKDWGFHVITDVGASND
HALGVAAGSNLMPNPGDSRELIAAVESGSKISDEVARDVDELLD
AVMTLYNSQANKNDNKPAHHAARAKAATTLENLKNEGSILPLK
PAGAKVAVGDFANPRQAGASSLNPTVETISEVIGSYDVQIGSS
RGYSRTGEADAATKREADLISARADIVLFFGLNDESESEGDRTHM
RIPQNQINLLQELQVQKVNLGIIAGSAIEMPWHYFKAILHCYLNQ
QAGAGAVMDILTPRPVNSPKGLSETIFRRHEHTPAYRYPSSRTEY
RESLYVGYRYTDADIPVLYPFAGLSYTFKQFYDNSWEVDGBE
NVGVEAGKEVAQLYVSLPGAKVFPRKELKGFAKSVLEPGESKRVE
AFDDKTFRYWNVKNKDEVEVGEGYQIRIGASSADIRLEGKILKSATT
DVLPISEAEPSYYSKGKTQVDEAEFEKPGVPSNLQICQLYAKASGLRFVYKRLTMKKADESGKPDNLNILFYNMFRAM
AKMTGGAVSMDDMIDVGLDVNGHHF
GGGLGKIISSGYFRNSKLKNKQYESRIKK

N4WNP3  Gracilibacillus halophilus

MKHKEIINKMTELLEKASLMSGKGDFWQTNIDRLGINSMFADGPGHI
RKQAEADADHLGLNESIPATCFPTAACATVANSWPDLGEGKDLYLREA
VAQNVNLGPINMKRDPLCGRNFYFSEDYPHYAGKLLASYIRQG
SHGISAVKHKFAANNQEARRMTIDTVDERTRLEIYLTAEAIQEGE
TKTVMSYNKLNGYTMENLHLMQRIWLDEGWGYNVVTDWGGSN
DRISGLIAGNELEMPTTAGETNKDIVQAIAKNTIKEELVDVCRDLE
LMKSTEVSQGNNQNDQFDKHQVFAQQAEEISVLLKNEHILPLQ
KDVKVATIGFAENPRQAGSIVNPNTLDLTTEFMKDYNQSIY
AQGFERYGKRSKRIKHACQLAKNADIVLLYMLDETEAEGLLDRH
DMKIPENQIELLHALYEVNPNNVVLSCGSAVEMPWIDKVGLVHHG
LSQGQAKAKAILRVTGEVNPGLAETTYPIQDTPANYHPFGEAS
VEYRESMYIGRYYHTAHDALFPFGYGLYTTFEYSDLQVTQQGV
TFITTTNGEVAGSEVACLQYVSSSGIEFRPSKELKGFSDKVFLQPGAKK
VEIAFDKTFRYNVITRNWEVESEAYMIQGAVEDIRLKTNHIHEGS
QAPQPYQKNQLPSYYSSEVNDVRQEEFQQLLGRRVPVSHDRTRPL

202
**K0J833 Amphibacillus xylanus**

MKYNTINQMTLAEKVSLMSGKDFWQSENIIDRLGIPSMLADGPHGI
RKQAAADHLGLNPSIPATCFPTAATVANSWDLVNLKMGQYLGIEA
ANQKVNLLGPGINMRPLAGRNFYEYSEDYPDLAGKLAASLGMIKQG
SKGIAASVKHFAVNNQEERRMAIDTVDERTLREILTAIEAIKGA
KTVMSAYNLGNYANENPHLLNTILRNEDYDGVVTVDWGNND
DRVAALKUVGELEMPTNGETDQDIYQAIKSHELKEVLDEAVDRL
ELIMSTSAKESNASVDEKHHQLAQKIAESIVLKNKNEIPLNN
DNVNAVIGDFAEREYPQAGGVNPTLENTLDLSAQSGIKSIFEPG
FNYRGLKSNRIAKACAEAEVLLLHGLDEVTAEAGLDRESIKIP
QNIQIDDDLEIYVQNVNHVIVVLSSAAGAVEMWPWIKGLLHQLSGQA
GAKAILRILTGEVNPSSQAEPSYIPYEDTPSYHQQGEVSEVYEYREG
LFIGYRYTDVANAVRFQFYGLSYTTFEYSDELVDRAGATFTTTNTG
NLPMEASASYLQVQCGSRSAFPRKPKLGFKVSLQPGESKTVTIPDFD
KTFRYFNKTNQWEIEEADYIMIGSSVDIQLSVFLVEGTGAPLP
YDQTDIPSYYSQGQSVSNVEFETLLGRKVPDAKWDRTKPLGYNDTIA
AQCYAKGLFARFAYQLIKSHFSLWKRIGKRSTANLIMMSVYHMPFR
GLARMHGIMNNPMVDGLVMIVNGQFYKGLKHLRERKQMIAAK
VNN

**W4QD03 Bacillus hemicellulosolyticus**

MKYNDLIIKMLEEASLMGKDFWQTESIERLGININSMLADGPHGI
RKQAAADHLGLNPSIPATCFPTAATVANSWNEELVEKVGEYLGEEA
VSKQVNLLGPGINMRPLAGRNFYEYSEDYPDLAGKLAASLGMIKQG
SHGISAVCKHFAVNNQEERRMSIDTIDERTLREILTAIEAIKGA
KTVMSAYNLGNYANENPHLLNTILRNEDYDGVVTVDWGNND
RVAGLLAGNELEMTTAGEIDKEIAAIKGHISEDICVLDLDDL
LFTTEKVFAKETKFDIEHHQQMAQKVASEISIVLKNKNEIPLKQNE
KVAVIDGFAQERQYQAGQSSVINPTLNDTLESGLTYIGFEKGFD
RYGKKSQKIDKACELADEVVLILYGLDEVTAEAGLDQRSMSIPENQIELIHALHKVNIVAVLSGCAVGAVEMPWIKGVGLLLHSYLQAG
SKAIALIZGEVNPSSKGALAYPKHEDTPAYHHFQGEVKSVROYREG
FIGYRYYDANTANVNLFPFSGFLSYYTFYSEVSVDGTVTFAINTGD
YAGSEVAQLYVGCSQDTNRPKKLKGFVTCLNPGETKSVTIPDDDTFRYFVNKTNQWEIEASYNMIGASCSDRLVETLVVEGTAFLPYD
KDVLSYYSGKANQVESIEFALLGKYKVPSTWRTVLPYNDTIAQC
CYAKGLFARFARATLSSHFLWKIGKRSTANLIMMSVYHMPFRGM

203
Appendix III

ARMTGGIMNMPMLDGVLMIVNGKFFKGLRHVLKERKKMRKAQKESQQLINNRRNKGEVL

**R7RMQ6 Thermobrachium celere**

MNIDEIMNELTEEEKCSLLSGADFWNLKSIERLGIRKIMMTDGPHGLRKQDMDAEIGLEKSVPATCFPSGAALASTWNKNIKEVGKAIAKECLNDWDVLLGPANVNIKRSPLCGRNFYYSEDPVLSKSKIAKYFKGKVQSGVACIKHAANQFERRRTTDDVRVDERAFREYLKSFEEAIKEAKPD
CVMCAYNKINGEYASDNKLLLNDLREWEYGEEVGSWGVAVNRVKSLAEAGMIDEMPSCGFVNDRIVYEAVKSGKIKIEVLDRAVRLKLIKHLHSNKSQTRCVDYEYHNHLASKVAAEAVILLKNDDILPLNKSRIAIIGEFAKNPFRQGGGSSHVNPATKLECPDIEIKKYANSVYARGFNSN
NDEIDEAIKEAVNLAKISDVVVLFLGLPERYSEGDFRAMDLIVNYQ
NINLVDEYVNKINVSLSGISCVEMPWLDKVAVNLGYLYLLQAGGSAINANFQTSVPSGKLSETFIKRLEDNPSEYINFPGSNDRVYGYGESVFVGYRYYDYKNIYDIVQFPSFGHGLSYTFRYEYSNLKIDNDYFDDECIDISFKLNVKGYKAKEVVQVYISKPNSSSIPIKELKEFKIEKEDVGEEREVNLKIKVDDLSYDFEQFSMVFSVDEEGYKLIISRSSRDIRLEGSIKVKRNQKV
KRMYHINSTIMBDVITKKGKELKLVQMLDVENDVQAKMMEFLNYMPKRLGLIYYNGKDYDFEKLEIIDLLNNEA

**F0HG16 Turicibacter sp**

MRDKQKQLINQMTQLKAGMCSGLDWFRLKSVERLGPKVMVSDGPHGLRKQKDGAADVNSIKACVFPGCAIACSFDRDLLYNGLILGCCECAENVSLILGPVNKRISPLCGRNFYSEDPSYSLSEMACNYIQGQSQGVGTSLKHFANNNQESSRMATASAQIDERTLREYILASFENAVKVGKWAVMCYNRNEFASFENKLLTDILHREWFEDGVYMSDWGANVRVEGLKAGLDLMPGSHGTNDKLIIEANKGILDETTLEDAERIVTKKIKVFVNRQDAIFNRVDHHEAASKIQSVALKNLDGVPNKEAKIAFIGAFKMSTRPFQGGGSHINTFYKVNSALEAVSNVTSVSYAQGYEDVDQINEELVQEAIVKSAVIFGALPDFFESEGYDRTMQLPNCQNELIKEVVKQPNTVVVLHSGVEMPIHDKVNGLEMYLQQAVGLATIDLLFGRNPNSKLAETFPLKLSDNPSLNYQVVDDILIHYREGIFVGYRYDDEKPVFPGYGLSYTTFETYHDLVLNKRKMLTDLEVSLKLTNTGAVAGKEVVQYVSDLTHLTLTIRPIKELKDFVKVELQAGETKEVQMTDLKRAFAWYNETISDHYWGTGEYELGKSREILVLDVVKVQSTVLFKVTANTTFGDLMKHETLRPIIESLAKQIDSNVQGEIDFNELELQDTPLRELRTIQUIDNAMIEYLIQTFNYL

**E6UA77 Ruminococcus albus**
R6U2U6  *Ruminococcus sp*

Ruminococcus sp

T4VHL0  *Clostridium bifermantans*

Clostridium bifermantans
Appendix III

VMCAYNKLNGEYCSENYRLLTEILRNEWGFEGFVVSDWGAVNDRDKGLYAGLELQMPADGGMGDALIVEAVKSNRLESGVLDAVERILNITFKAENKRESYIYSEKHHILARIOGECVLLKNEEKILPKKEEKIAVIGELATKVRYQGGGSSHINPTKLDTYEIVNFAGSENIRYARGYIDLSIIDDITYLEEAKQLAIADKYLFPGLPERYESEGFDRTHLNIPKQNYDVALKSVPNENIVVLNSPPIEPMFVSVDVAILEAYLTGQASGKAICDILGVEPNSGKLAETFALKLDNDPSYNLNPGEVDKVEYEGIFVGYRYYDKKAMDVLFPPFYGLSYTNYFENSLKISKNEIDDETKVTVSVSIKNIDVFGKEIVQLYISDKESSVIRPEKLGFEKIGLEPNGKEKETFILNKRSEAYYNVLDGDWHVESGFEILKSSREIVLKEIVTNTTSPKTIYTKNTALGDISHLPEVQQIMDAIQSFGRDTSGLGEGNMFAEMMKFMPLRALATFNPDDQQLVDRIESINS

U4MU82 Clostridium thermocellum

MAVDIKKIIKQMTLEEKAGLCSGLDFWHTKPVRLGIPSIMMTDGPHGLRKRQREDAEIDINNSVPATCFPSAAGLACSWDRELVERGVAALGEECQAEVNSILLPANIKRSPLCGRNFYFSEDPYLSSELAASHIKGVQSQVGACLKHFARANQEPHRMMTVDITVDERLREIFYFASFENAVKRAPPVVVMCAYNKLNGEYSCNYLLTEVLKNEWMHGFVSVDWGAVNDRVSGLDAGDLDEMPTSZHITDKKIVEAVKSGKLENILNRAVERILKVIFMALENKKEAQLYDKDAAHRRLARQAASEMVLLKNEEDDVLPLKKSGLTAICVFQKPYQSGSSHHTPRLDIYEIKKAGGDKVNLVYSEGYRLENGIDEELINEAKKAASSSDDAAVFAGLPEYESEFGDRTHMSIPENQNLIEAVAEVSNNIVVLLNPSVPMDKVKVSLAYLGGQALGGQALADVLFGEVNPSKGLAETFPVKLSHNPSYNFPGEDDRVEYKELFVGYRDITYDKGIEPLFPGHGLSTKFEYSDSVDKADVSDNNINVSVKVNQVMAGKEIVQLYVSDKVSSVRRPEKELGFEKVFNLNPGEKTVFTLDKRFAFYNTQIKDVDHVESGEFLILGRSSRDIYLKESVRNSTV RKRFVTNSAVEVDVMSDSSAAAVLGPVLKEITDAILQIDMDNAHDDMAANIKNMPRLSLVGYSGSQQRLISEMLEELVDKINNVE

G8LXR2 Clostridium clariflavum

MSRDIKKIAEMTVEEKASLCSGFGNWHTKAVELQIPPMIMDVGPHGLRIGFKNADLSDTQNSLPATCFPTAVNMASTDWDRNLVEEIGKAIGEERAAEESVILLPANIKRSPLCGRNFYFSEDPYLSSOMAASHKGVQSQVGVSLSKHKFCANNQEHRRRTLTDVKVVERDLRTLEIYLASFEEAVKQAKPWTVMCSYNVNGESENNYLLTHILRDEWGFEGFVSVDGAVNERVKGLKAGDLDEMPSGGERDKQIADAVKNGELPPEVLDKAVELKLIIFKAIIDNKSSGTTFDKKAHHHELARRASMVLLKNEGDGILPLKQKGKIALIGAFAKNPRFQGGGSSHVNTLSNAYDAIVDLAGQKAE
LYSPGYDLETDVDKEKLIDELETEAKEAAAADAVAVIFAGLTDSEYSEGYD
RAHNLIPENHRLLIEEAEEVQGKTUVVLNSGSPIDMWDKVKAELTY
YLGQAVGEAYADLFGEFSPGCGLAEFTFPEKLISHPYNLFPGEDS
VEYREGLFVGRYRYDAKDIKLPLPFGPSGLVDTCYEYSDKIUSKKEIND
PELLTVSKKNTGMGKKEIEVQLYVRDVEKSVRDPKLELIGFEKIEL
DAGEEKTIVTSLDKRAFAYTVNTEIKDHWSEGEFELIGKQSSIDVLK
EVVTNOSTIMIKFHFMTLGDITPGLDLKLNQYMEYLKKHG
MDKGIKHKMKSMPDTEMVKYTPLRCLSFGNVEKTVIILLEDLNS

**L0EBH0 Thermobacillus composti**

MPRLKLKLISEMTEEEKAGLSGLNFLWRTKPIRLGIPSIMMTDGP
LRKQQEDDDGDHLVIGSDVATCFPSAGLAAASNRELVEKVEALG
RECRAHGVILLSLAPANNKSLPLCGRFNYFSEDYLTAKMAASHIRG
VQSGVGTAAKHYAMNNQEHRNMVDAIVDERTQREILYLAFEGA
VRQAPWAVMAANYKVTGNYATENKTTLDILREWGEFVVDSD
WGAVNDRVAGLAALDEMPNGGYRDKKIVAEVKSSLLEELLD
RAVERILNFKAADAQYNVSFDPRHDLALFAEVMVECKLLKNE
DGILPLKHGRIAVGEMMACKPRYQGGSSHVPTKLDIPFDEMEVA
IQGAEELVYAELYKHEKDPNEVMVEAVRAVES(i)AVFAGLDR
YESEGYDRIHKMRPDHNHRLEAAVAVQNPVVVLNSGSPVEMPW
GQVKGVEAYLGGQAAGGAIADLLFGESNSPGLKEFTKSRHNSPS
YLNFGESDREYERELGFVGRYHEARGIEPLUFAGHGLSYTFYE
ITSLDKREMRTDRTLRVCVKNTGNGKEVQLYRDTASSVR
EKEILKFQKVSLPGEKETVEFALDKRAFAVVTDLRNWHVETGEF
EILAGGSSDRIALAKVHVHTSEAVKKTFTPSTLDLLADPAQA
QAMQRMSGWVP(V)D(T)EETQMMMAVMLDMLRSLVAFSGGAFTEE
IMNGILRQLNGVNS

**D9TTJ4 Thermoanaerobacterium thermosacchara**

MKDKIKLKLISEMTEEEKASGLNLSGLNWQTKEKPIRLGIPSITMDGP
LRKAAKSDGLGGLDDSVATCFPSGNAALASWDRLIKKVGEAEGE
CAEDVHIILGAPANNKSLPLCGRFNYFSEDYLTASELALAANYKGVQSKG
VGTSIKHYANQEDYRTMDVKVNERALREYLTLGFEGAKQSQP
WTVMASYNKVNGVYATENHILINRNEWKDFGIVSDWGVAND
VAALKAGLDIEMPGSGGEEKKIVAEVKQGQISEEYLNSAIERLINIF
KAYENKKNQNYDAEKKHIQLARQVASECMVLKNEEILPKKQG
KIAIIGELAVKPRYQGGSSHVPTKLDIPYDKTKIANGKAEKTYPG
ELEKDEVNKLIEEAAYAKNSDVAFAGLPSYSEGRAYDRHR
ESHNKLIQAIAYQPNVVLNLNCGSPIEMPVWQVKGILEAYLGGQA

207
Appendix III

SGGAIADILFGEKKNPCGKLAEFTPRELRNNPSYNFPGEGKNVNYGEG
IFVGYRYYDTKGVQELPFPFPHGLSYTTFETYDSTKNNKITEETIIVER
VKVKNNTGKRAGEKIIQLYVRDQISSVTRPHKELKGFQKIYLPGEK
VFVFNLKRSFAYYDVSSDKWYVETGDEIVLGGSSKNILLKTVIHITS
TTSVVKEYTRNSTDINDVNQYGRQIYNIKRIKSDTDNDSDNMILK
KQLETNSQEIPESSSSQKDMMNALKNMLPLALTILSNIFTEEMVEN
DGLNTISN

O08331 Clostridium stercorarium

MRQDIIKIIISQMTLEEKASKALCSALDAWTLKGVERLGIIPSIMVSDPGH
LRKQRDQPDTPGKKTVPATCFPATVAGLASSWNRVEKVG PALGEE
CQAEGIAVLGLPQTNKLSLPSGRFNEYFSEDPSLSEMARMHKVGQVS
RGVTGSLKHHANQNHEHRMSVDAVIDRTLREIYLASFGAVKKA
KVPWTIMCSYRNGNEYASENFLLTDVRNEWGFEGIVSVDWQAVN
ERVKGLEAGLDLEMPSSFQGIQDKIKVEAVKKGELPEEVLRTVERILN
LIFKADVRNKENAGYDREAHHKLAARECMVLLKNEKDIPLRLK
QGTIAVGFEAKPRPYQGGSSHVNMPTMDSYPYEIKKSAQNADVIY
AQGITYIEKDPEKLEAEKQTALKADAVIFAGLPFPEYCEGDRT
HMTRMPESHCTLEIEAETVTNVYLCRCSPFEMPWIDYKVLGPLEEA
YLGQGAM GGGLPAPCSETPIQPLKAFEIKFQPSLDPNFSYNLERDREV
YREGIIVGYRYYDKKNNELPFPFQGLSYTTFEYGLKISRKEISDNE
TVTTSVIVLKNTGDMAGKEIVQLYVRDIETSVIDRTEGEKFNLPQ
GEEKTTVFDLDKRAFASYTHRYKDVHEVFGDFISSDFIRDLKDK
IFVKTSTTIKRWTVNTLVGDLLSDRVLEPVFREIFINEKTWYDLLL
DEDNHLSVWMRTPLRLS ANSTGGELEEEKLNRLIDTNANIK

F8FL27 Paenibacillus mucilaginosus

QRQDIQLISQMTLEEKAGMCGLDGFLHKLKGVERLGIPSVMVTDGPH
LRKQKAADHLGLFDSVPATCPSAAGLACSWDRERLIRVVAIGGEE
CQAENVAVLGLPARKSMLCGRNEYFSEDPSLSEEMASHIAGV
QSGQVGTSLKHFAVNNQHEHRMDSVADAVERTLRLEIYLASFAVK
QSPQWTVMSSYNRVNGITYASENELTTDLKNEWHEFGVVFVSDWGA
VDERADALAGLELEMPASGGVGERKVEAVQSSRLEMEALDRAV
ERLTLIFIFADVHRKPQAYDPAEHHLRAEVAREMSVLKZMSSS
PLSKGSLAVIGANADKPRYQGGSSHIVPTQLDPVVEIIRKLADA
VYTAQGIRLESDMAEALTEEAKRASAAADTAVIFAGLPYSEREGY
DRTHLSLPA NQIRLEEVAAVQPVKPV VVVLNPSVEMPWSAGAQLAV
EYGILGGQAVGGAADLLYGEANPCGKLAEFPQLDSLPSYLFPPGE
GDRVEYREGIIVGYRYYDKTNNVKPLFPFQGLSYTTFEYTSMTLDR
QDTESIITVRVTVKNTGAAAGKEIYQLYVKAIAESTVIRPAELKGFAK
VFLQPGEERTVSFLDKRAFAYYNVDLKDWHVETGFEHILAGSSQHVILQDSVVVESTAALRKETYTRNTTGLDLLQDAAAAREKAQGLLQAFOQ
EASGFADDHADMMAAMMKYMPLRALVGFSQGALTEQTLLEDLLKELNH

**W4BGN2** *Paenibacillus sp*

MSEQRDIEKIIINQMTLEEKAGMCGLDFWNKGVVERLGIPSIMVTDGPHGLKRQQRGADHLGIFDSVPSTCFPSAAGLASSWDRELIRQVGVALGEESCAEDVAVLLGPANIKRSVCRGFYFSEDYLSSELAASHIQ
GVQSQVGSXSLKFAANQEHRRMTTDAVIDERTLREIYLASFEQAVKKAPWTVMCSYNQVNGTYASENPRLLETILKDEWIGFQDQDSW
GAVNQRDDALAAEGOMELEEPSSNQRGKRVHVAVQGKLTIEEALDRAVARILRIIFMAVDHIKKEVYEQHKA overshooting the word limit
LVPSMALQNFQVPLSSVLQVSGVQKTLDDHSLKLAKAVDRIEYLSLTVGCGVVSGSYQWYKGNL

**E5YXP0** *Paenibacillus vortex*

MSEQRDIEKIIINQMTLEEKAGMCGLDFWNKGVVERLGIPSIMVTDGPHGLKRQQRGADHLGIFDSVPSTCFPSAAGLASSWDRELIRQVGVALGEESCAEDVAVLLGPANIKRSVCRGFYFSEDYLSSELAASHIQ
GVQSQVGSXSLKFAANQEHRRMTTDAVIDERTLREIYLASFEQAVKKAPWTVMCSYNQVNGTYASENPRLLETILKDEWIGFQDQDSW
GAVNQRDDALAAEGOMELEEPSSNQRGKRVHVAVQGKLTIEEALDRAVARILRIIFMAVDHIKKEVYEQHKA overshooting the word limit
LVPSMALQNFQVPLSSVLQVSGVQKTLDDHSLKLAKAVDRIEYLSLTVGCGVVSGSYQWYKGNL
Appendix III

KGFAKVDLEPGEHKTVSITLDKRSFAYYNVELKDWHVESGDFEVLIG
KSSQEIVLKDLTQVQSTVRLEQQYTLNSTIGELLSDPVTAEETGQOLLK
KFQEASPMSGMAEDDSHSELFAMAAMKDMPLRNLLAFGGGAVKEET
LLQLLDGLNRRR

L5NBC9 Halobacillus sp

MDASIDHLKKMLTEEEKAGLSGRDFWNLKGLERLDIPSVMVTGDPH
GLRKYQAQAGAHLLGNASVPATCFPSAAAGLASTWDQDLICRVGEALG
EEAKTEEAVLLGPGTNKRSPLCGRENFEYFSEDPYLSSSMAASHIEG
VQSQKIGTSLKHFAANQHRRMSVDARVDERLREIYLASFEHAVK
QAAPWTVMCSYNQVENGEYASESCLLTLTEILDNEWGWDGVVVSDDW
AVNEVGDGLNAMDLEMPSTNGHIDREVEAVKGDGLTEATIDTV
GRVGLQIKAVNQQQTSYDKEAHHLRAAADGMVLKNEGRIL
PLDKASAVLIGSFVPQRIQGGSSHHNPRVDDVEEVKVTGTR
VTAEGYPLENDAIDEAMIEEACTTAADAADVAVLFAGLPDRESEGY
DRKHLEPENHRALIERVAQVPNIVVLSENGAEPEWPVLDEVPAVL
EGYLGGQAFQGAVADLLFDKTSGKALETFVPVLVDNPINTNFGE
GDVEYEKGFLVGYRHYDQAKIEPLFPGHGLSYSTFEYSSLMDSHE
IDDTVEEVSTVDMNTGSVEAETQLYVRDESSVIRPEKELKGFA
KVSLEPQEVKTIISFTLDRRSFAYNYVNLGWDQVETGDIFILGKS
IVLKDTHHRSTTVISMPIHRNTLVGDLKLPKATASVMEFLANPF
GGMETEDEGMMDMAMNPLRALVNFSGAFTEKHLQELIGS
LSRRI

C2Z9J7 Bacillus cereus

MKRDIIKKISQMTLEEAKASLCSGLDFWNKIGIERLGPISGIVIMVTDGPHG
LRKQAEGADHLLGIYNSPSTCFPSAVGLASTWNKDQNLQVGVALGEE
CQAENVGVGLPGANIKRSPLCGRENFEYFSEDVPILSSMAANHVKG
VQSQGIGTSLIKHAANQEHRRMSVDVDAVERTLREIYLASFEDVKE
AQPWTVMAYNKGHEYASENNYLLNDLKDWEWGFEGVSVDWA
VNERASLANGLELMPSSFGIEKIKIVDAVGELSVEKLDQSVER
LLNIIFKAVDQENAVSYDHAQILAREVASEVMLKNEDSIPL
KKEGTVAIIIFGAKPQRQGGSSHHINPTKLESILEEIIEMVSNGEKTNLIF
EGQYNLASSDVENDMINNEAKIAESADTVLFXVGLPDRESEGFDK
HLQMPNHVQLIEAIEAVQSNIVVVLSNAPIEMPIWIGVKGILEGY
GGQALGGAIADILFSGDANPSGKALETFPKVLSDNPYNFPGEQGDVK
EYKEVFVGVYIYDKKNEVPLPFGFGLSYTNEYSNLSDVDKEIKD
TETSVTVNVKNIGSTVGEIVQLYIKDVESTMIRPEKELKGFEKVEL
QPGEEKTVNTLNKRSFAYYNVELKDWVETGEFEILVGGSSREIVL
QDNMYVQSTTIQQIVHRNTTLDIFADPILAPIAKGLMEKALKDSPFG
SMAEGDSVSEMMDAMLNYMPLRALVNFSAAGFTEEMLSKIGTLD
AQMN

**W4R0J6 Bacillus weihenstephanensis**

MKRDIIKSQMTLEEKASLCGLDFWNTKGIERLGIPSIMVTDGPHG
LRKQAEGADHLGIYNSIPCTCFPSAVGLASTWKNKELIKQVGVALGEEC
QAEHVGLLPGANIKRSLCPGCRNFYEFSEDPYLSSQMAINHVKGQV
SQGIGTLKFAANQHRRMSVAIDVERTLEIYLASFEDIKAEQ
PWTMSAYNKVNGEYASENYLLHDILKEWGFEGVSVSDGAVN
ERVASLANGLEMPSSFEGIKKIDAIHCGLSVEKLDAQVERLLIY
IFKAYDNQLENATYSKDMHQLAREVESMVMQLNEDSILPPLKE
GTAVIGEFAQPRYQQGGSSHINPTKLASIFEESEMVSGEKNTLFA
QGYDLASDDDVNEALKIAESADTVLFGPLDNEYSEGDFRH
LQMPENHVQLIAEAIQVSINVLGSNGAPIEMPIWKVQGQELYLG
QQALGGAADLLFGANPSKLAETFPEVLSDNPSYNFPGEDKVE
YKEGIGFVGYRYYDKRIEPLFPGFGLSYNFEYEYNSNKKEITD
TETVSVSNVNTGRAGKEIVQYIKDESSMTRPEKEKGKVELK
QPGEKETKSTLNNKRSAAYNVNVELKDWHTGFEIFVGGSKSSKEIVL
DMSMYVQSTTTQPKVEHRNTTLDIFDPILAPIAKGLMEKALKDSPFG
SMTEGSASSEMMDAMLNYMPLRALVNFSAAGFTEEMLSIIGILD
AQMN

**U5ZNE7 Bacillus toyonensis**

MKRDIIKSQMTLEEKASLCGLDFWNTKGIERLGIPSIMVTDGPHG
LRKQAEGADHLGIYNSIPCTCFPSAVGLASTWKNKELIKQVGVALGEEC
QAEHVGLLPGANIKRSLCPGCRNFYEFSEDPYLSSQMAINHVKGQV
SQGIGTLKFAANQHRRMSVAIDVERTLEIYLASFEDIKAEQ
PWTMSAYNKVNGEYASENYLLHDILKEWGFEGVSVSDGAVN
ERVASLANGLEMPSSFEGIKKIDAIHCGLSVEKLDAQVERLLIY
IFKAYDNQLENATYSKDMHQLAREVESMVMQLNEDSILPPLKE
GTAVIGEFAQPRYQQGGSSHINPTKLASIFEESEMVSGEKNTLFA
QGYDLASDDDVNEALKIAESADTVLFGPLDNEYSEGDFRH
LQMPENHVQLIAEAIQVSINVLGSNGAPIEMPIWKVQGQELYLG
QQALGGAADLLFGANPSKLAETFPEVLSDNPSYNFPGEDKVE
YKEGIGFVGYRYYDKRIEPLFPGFGLSYNFEYEYNSNKKEITD
TETVSVSNVNTGRAGKEIVQYIKDESSMTRPEKEKGKVELK
QPGEKETKSTLNNKRSAAYNVNVELKDWHTGFEIFVGGSKSSKEIVL
DMSMYVQSTTTQPKVEHRNTTLDIFDPILAPIAKGLMEKALKDSPFG
SMTEGSASSEMMDAMLNYMPLRALVNFSAAGFTEEMLSIIGILD
AQMN
Appendix III

**K0FNN0 Bacillus thuringiensis**

MKRDIIKIISEMTEEEKASLCGKDFWNTKGIERLRGIPSIMVTGPHGRKQAEADHLGIYNSTTSTCFPASVGLASTWNKDLHEVGVALGEEQAEEHVGLLPGANIKRSLPLCGRNFYFSEDPYLSQMAINHVKGVQSQGVTSLKHFAANQEHRRMSVDAIVDERTLREIYLAISFEDVIKAPTQWTVMSAYNKPEGFVSVADWGVNEVNSLANGLGEMSPSSGIFGEEKIIDAINCGELSVKLNQAVERLYIIKAYENQLENATYSKDTTHQLAREIVASESMVQLNQEDSIELPLKGEGTVAVGEIAKQPRYQGSSSHINPNKLESIIFEELEMSGKNTNFAQYDLASDDVNENLINEAKIAESADTAALFVG0LDPYFDFGFRKHLQMPENHVQLIEIAIAVQSNIVNLSSNGAPIEMPGKVKGILEYLGGQALGGIAADLLFGDANPSGLAETFPEVLSDNPSYLNPGEQDKYEYKEGTVFVGRVYDANQIEPLFPGFLSYTNFEYSKLSISKNEIKDTDTVSVLNVKNSGAIKGEIQLVYKIDVESSMRPEKEIKGFKEKVELQPGEKTVSTFTLNNRSAYNYENLKVDHVETGFEIVLGKSSREIVLQDNIFVQSTTIKHTNRLGDFADRMPIAIKELMEALKDSPFASMAEGGSDESEMDMLNYMPRLNVFSAIFTEEMELSEILLNDAQMNQ

**Resurrected beta-glucosidase sequence**

TDNIKELVQMTLEEKEKLCSGKDFWHTQSIERLRGIPSIMVTGPHGRKQAEADHLGLINESVPATCFPTAALASSWDPELLHEVGEALGEECCRAEVSVLPGVNIKRSPLCGRNFYFSEDPYLAGEMAAAWISGQSKGVTSKLKHFAANQEHRRMTVDAVDERTLREYLAIFENAVKQAQPWTVMSYRINQVYSEKKWLLTEVFREDEGFEGLVSYVGWAVNDRYKVGLKAGLDLEMPSSGGLNDDQKIVEAVRNGLDEAEDLRAERILQIAARAKQNYTVEAHHALARRIAAESAVLLKNDGILQLKKEAKIAVIGEFATEPRYQGAGSSQINTPTKLDNALDELRERGADVYAPGYELDGRDTDAALLEAVEVAKNADVVVFAGLPDSYSEEGFRTDLNPENHNALIEAVAPVNPVVVLNSGSPVYTQMPWRDKVAILESYLGGQAGGSAIADILTGEVNPSSGRLATEFPURLDNPSELYNFPGEQPQHVEYESIFVGYYDTEACKHAFPFYGYLSYTTFEYSDLKISAKADDESWVTSVTNTGDRAGSEVQYVVGDAESTVFHPVEQLKGFAKVFLEPSREVTLDRAFSYYNVKINDWTVSEGDFEIRVYGGSSRDIRLKAVTVLNSSTTPLATFTVNTTIDMASPAFGKALLGALVQAVSAGSGAKDVSRRMMAMLQDMPLRSLAMFTGGAIIPEMLEELVEMLNG
Appendix IV

List of sequences used in the cellulosome

Scaf1: Histag-CBM-COHESIN7-XDOCKERIN

MGSSHHHHHHHHSSGLVPRGSHMANTPVSGLKVEFYNNSNPSDDTNSINPQFKVTNTGSSAIDLSKTLRYYTYTDQKDFQWDCHAAAIGSNNSYNGITSNVKGTFFVMSSSTNNADTYLEIISFTGGLEPGAHVQIQGRFAKNDWSNYTQSNDSLFSASQFEWVDQVTAYLNGVLBLWKEPPGGSVVPASIGTAVRIKVDTVNAKPGDTRVRIIPVRFSGISGSKGIANCDFVYSDPNVLEIEIEEPGELVDPNPTKSFDTDVFADKMIIFLFAEDSGTGAYAITEGDVFATIKVKSAGPNGLSVIKFVEVGGFFANNDLVEQKTQFFDGGVNVGTNSKPVIEGYKVSGYILPDSFATVAPLKVAGFKVEIVGTELYAVTDANGYFEITGVPAASAGYTLKISRA

Scaf2: Histag-CBM-COHESIN7-linker-COHESIN7-XDOCKERIN

MGSSHHHHHHHHSSGLVPRGSHMANTPVSGLKVEFYNNSNPSDDTNSINPQFKVTNTGSSAIDLSKTLRYYTYTDQKDFQWDCHAAAIGSNNSYNGITSNVKGTFFVMSSSTNNADTYLEIISFTGGLEPGAHVQIQGRFAKNDWSNYTQSNDSLFSASQFEWVDQVTAYLNGVLBLWKEPPGGSVVPASAVRIKVDTVNAKPGDTRVRIIPVRFSGISGSKGIANCDFVYSDPNVLEIEIEEPGELVDPNPTKSFDTDVFADKMIIFLFAEDSGTGAYAITEGDVFATIKVKSAGPNGLSVIKFVEVGGFFANNDLVEQKTQFFDGGVNVGTNSKPVIEGYKVSGYILPDSFATVAPLKVAGFKVEIVGTELYAVTDANGYFEITGVPAASAGYTLKISRA
Appendix IV

TYLDRTVIANVVVTGDTSVSTSDQPMWVGDIVKDINLLLDVAVIRCFNATKGSANYVEELINRNGAINMQDIMIVKHFGATSSDYDAQ

LFCA-Dock: LFCA-linker-DOCKERIN-histag

MASMTGGQQMGRIRTPVETHGVLSVKGGQLVDENGKPVQLRMGSSHGLQWFGDFVKNKDSMKWLDDGIVNFRVAMYTAEG
GYITNPSVKVKEAVEAAIDLGMYVIIDWHLSDNDPNTYKEQAKAFFQEMAKYGNYPNIVIEICNPNGGVTWSNQIKPYAE
EVPIAIRANDPDIIVGTPTWSDQVDHAADNPLPSNIMYALH
FYAGTHGQSLRDKIDYALSKGVAIFVTWGTSDASNGGPGFLN
ESQKWDMDNSRNISWANWLSLDKSETSAAALMGASPQTGGWT
DSNLSASGKFVREQIREFPnplsdlsqptpsnptpslppqvvyGDVNGDG
NVNSTDMLKRYLLKSVTNINREAADVNRDGINSSDMTILKRYLIKSIPHPLEHHHHHH

LFCA-CBM: LFCA-linker-CBM-histag

MASMTGGQQMGRIRTPVETHGVLSVKGGQLVDENGKPVQLRMGSSHGLQWFGDFVKNKDSMKWLDDGIVNFRVAMYTAEG
GYITNPSVKVKEAVEAAIDLGMYVIIDWHLSDNDPNTYKEQAKAFFQEMAKYGNYPNIVIEICNPNGGVTWSNQIKPYAE
EVPIAIRANDPDIIVGTPTWSDQVDHAADNPLPSNIMYALH
FYAGTHGQSLRDKIDYALSKGVAIFVTWGTSDASNGGPGFLN
ESQKWDMDNSRNISWANWLSLDKSETSAAALMGASPQTGGWT
DSNLSASGKFVREQIREFPnplsdlsqptpsnptpslppqvvyTSMANTP
VSGNLKVEFYNSDPTTNSINPQFKVTNTGSAIDLKLTLRY
YTYTVDGKQKDTFWCDHAALGNSYNGIATNVSNGTFTVKMSS
STNNADTYLEISFTGTLPGAYHVIQGRFAKNDWSNQTQSD
YSFKSASQFVQWQVTAYLNGVLVWGKEPGGSVPLEHHHHHHHH

214
Cel8A: **CEL8A-linker-DOCKERIN-histag**

MGVPFNTKYPYGPTSIA DNQSEVTAMLKAEWEDWKS KRITSN
GAGGYKR VQR DASTNYDTVE SGM GYGLL AVCFNEQALFDD
LYRYVK SHF NGNL MHWHID ANN NVTSHD GGA ATDADE
DIALALIF ADKL W GSSGA I N YGQE ARTL INNYNHC VEHG SYV
L KPGDRWGG S VTNPSYFAPA WYKV YA QYTGD TRWNQVAD
KCYQIVEE V KKY NNGTGL VPDWCTAS GTPASG QSYDYKYDA
TRYGWR TAVD YSWFGDQR A KANCDMLTKF ARDG AKGIVD
GYTIQG S KISNNHNAS FIGPV AASMTGY DLNFAKEL YRETVA
VKDSEYYGYGNSLR LLTY ITGNF pnplsdlsqtppsnptpsllpqqv
vyGDVN GDGN VNSTDLMLKR YLLKSV TN INREAADVNRDGA
INSSDM TILKRYL KSIPHL PYLEHHHHHH
I would like to start thanking Raul Perez-Jimenez for giving me the opportunity to carry out this thesis under his supervision. During the last years I have learnt things that I could not even imagine four years ago from and with him. Many thanks to Txema Pitarke for for giving me the opportunity to develop my career in CIC nanoGUNE.

I also want to acknowledge the colleagues I had during my Thesis. Muchísimas gracias Borja, has sido un gran apoyo. Eskerrik asko Aitor, zure laguntzagatik. Mila esker Leire Barandiaran, laguntzagaitxik, aunque sea duro lo hemos pasado bien Dinosaurios. Muchas gracias David, por toda tu ayuda y por los fittings. Eskerrak emon nahi nizkioke ere Leire Aldazabali, beti hor egotiaigaitik. Merci Marie. Muchas gracias Alvaro por tu ayuda. Danke Jörg. Gracias también al resto de mi grupo por todo: Patricia, Bábarba, Anne, Susana… Thanks also to the rest of nanopeople, bereziki, Aitziber, Iban, Unai, Paulo, Cesar, María. Eta nola ez zuri Miren, mila esker hor egotiaigaitxik.

I also want to give thanks to Mariano Carrión Vazquez, Albert Galera and the rest of the group for the cellulosome experiments. I am very grateful with the people from Weizmann Institute of Science for giving me the opportunity to do a stage
there. Especially Professor Edward Bayer, I learnt a lot in his
lab. I also want to give thanks to Melina for having care of me
as I was your little sister.

Amaitzeko, gertukuez gogoratzia nahi dot, kuadrilakuez, beti
hor egotiagaitxik, beti irifiarre bat ataratziaigaitxik ta emondako
animuengaitxik. Marta maitxia zelan ikasi dozun nere proiektua
esplikatzen ta zelan lagundu dostazun deskonektatzen. Azpiri,
beti hor egotiagaitxik ta babesa emotiagaitxik, beti zauzielako
hor bixok bihar dan guztixandako. Eskerrik asko zuri be Amaia,
beti hor egotiagaitxik ta zuri be Ale, entzutiaigaitxik ta
laguntziagaitxik, portadia primeran geratu da. Baita eskerrak
emon nere taxista gogokuenei, astelehen goizak ez diralako
berdinak zuek barik. Ta zelan ez etxekuei, beti zauzielako hor ta
zuek barik hau ez zalako posible izango. Mila mila esker ama,
laguntzagaitxik ta beti animuak emon ta entzutiaigaitxik.
Eskerrik asko aitxa ta Jon, bihar doten danandako hor
zauzielako. Eskerrik asko bebai beste etxekuei, mila esker Itsaso
zure laguntza ta animuengaitxik. Ta zelan ez zuri maitxia,
eskerrik asko Gaizka, onduan euki leiken pertsona honena
zaralako, zurekin hau ta nahi dotena eitzia posible dalako,
ezebez esan barik be dana esanda dauelako…

Mila esker danori nerekin egotiagaitxik!