6-1-2012

In silico identification of potential biomass and cell wall degrading enzymes in the microbial community of the Red Sea Atlantis-II brine pool using metagenomic approach

Norhan Mohammed Magdy Mofeed

The American University in Cairo

School of Sciences and Engineering

In silico Identification of Potential Biomass and Cell Wall Degrading Enzymes in the Microbial Community of the Red Sea Atlantis-II Brine Pool using Metagenomic Approach

A Thesis Submitted to

The Biotechnology Graduate Program

in partial fulfillment of the requirements for
the degree of Master of Science

by Norhan Mohammed Mofeed

Under the supervision of Dr. Hamza El Dorry
May/2012

Dedication

I would like to send my extreme love to my family; especially my mother, father, brothers and my fiancée. If it were not for their care, support and continuous encouragement, I wouldn't have been here today and wouldn't have been able to achieve this work. I'd like very much to dedicate this work to them.

I'd like as well to extend my thanks to all my friends and colleagues at AUC and outside it for bearing with me all the difficult times and being supportive.

Acknowledgements

I would like to thank my mentor, Professor Dr. Hamza El Dorry for advising me and supervising my thesis, Dr. Mohammed Ghazy for co-advising and for his continuous help and support throughout my work. I would like as well to thank Dr. Rania Siam for her support and guidance throughout my studying. I would also like to thank Dr. Ari Ferreira, Mr. Hazem Sharaf, Mr. Mustafa Adel and Ms Mariam RizkAllah for their help in the computational work. And I would like to extend my thanks to Mr. Mohamed Maged who helped me with everything.

I would also like to thank KAUST for providing us with the funds necessary to do this work.

And I would extend my thanks to the Alfi foundation which provided me with a fellowship and enabled me to complete my thesis.

TABLE OF CONTENTS

List of Abbreviations

BAC	Bacterial Artificial Chromosome
PCR	Polymerase Chain Reaction
WGS	Whole Genome Shotgun
E. coli	*Escherichia coli*
HSP	Heat Shock Protein
UCL	Upper Convective Layer
LCL	Lower Convective Layer
HMM	Hidden Markov Model
BLAST	Basic Local Alignment Based Tool
CAZy	Carbohydrate Active enzyme
NC-IUBMB	National Committee of International Union of Biochemistry and Molecular Biology
IUB	International Union of Biochemistry
GOLD	Genomics OnLine Database
MDa	Mega Dalton (one million Daltons)
SLH	S-Layer Homolgy
ORF	Open reading frame
Nr	Non-redundant
ATII	Atlantis II brine pool
ATII-LCL V0.2	Atlantis II Lower Convective Layer Version 0.2 database
RBS	Ribosomal Binding Site
C	Contig
CT	Contig Terminated
P	Potential
FL	Full Length
D	Domain
psu	One gram of salt per 1000 grams of water is defined as one **P**ractical **S**alinity **U**nit or one PSU (http://podaac.jpl.nasa.gov/SeaSurfaceSalinity)
pfam	Collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs) (http://pfam.sanger.ac.uk/)

List of Bioinformatics Tools

Program Name	Link	Description*
NCBI Blast	http://blast.ncbi.nlm.nih.gov/Blast.cgi	The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families (Johnson et al., 2008).
Artemis	http://www.sanger.ac.uk/resources/software/artemis/	Artemis is a free genome browser and annotation tool that allows visualisation of sequence features, next generation data and the results of analyses within the context of the sequence, and also its six-frame translation.
BProm	http://linux1.softberry.com/berry.phtml?topic=bprom&group=programs&subgroup=gfindb	Algorithm predicts potential transcription start positions of bacterial genes regulated by sigma70 promoters (major *E.coli* promoter class).
SecretomeP	http://www.cbs.dtu.dk/services/SecretomeP/	Prediction of non-classical protein secretion: The SecretomeP 2.0 server produces *ab initio* predictions of non-classical *i.e.* not signal peptide triggered protein secretion. The method queries a large number of other feature prediction servers to obtain information on various post-translational and localizational aspects of the protein, which are integrated into the final secretion prediction (Bendtsen et al., 2005).
PSORTb	http://www.psort.org/psortb/	PSORTb program for bacterial protein subcellular localization prediction: PSORTb v3.0 now handles archaeal sequences as well as Gram-positive and Gram-negative bacterial sequences (Yu et al., 2010).

SWISS-MODEL	http://swissmodel.expasy.org	SWISS-MODEL is a fully automated protein structure homology-modeling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make Protein Modelling accessible to all biochemists and molecular biologists WorldWide. (Arnold, Bordoli, Kopp, & Schwede, 2006)
TMHMM	http://www.cbs.dtu.dk/services/TMHMM-2.0/	Prediction of transmembrane helices in proteins (Krogh, Larsson, von Heijne, & Sonnhammer, 2001)

*Description of the tools are copied from the corresponding websites

List of Figures

List of Tables

Abstract

Atlantis II is the largest brine pool in the Red Sea. It lies at 2,200m deep with an area of about 60km^2. The lower convective layer of the Atlantis II brine pool (ATII-LCL) is characterized by extreme conditions, the temperature reaches 68.2°C, salinity of 270 psu, and there is high concentration of heavy metals. Microbial communities inhabiting this harsh environment are expected to have enzymes and proteins that are adapted to these conditions. Such proteins and enzymes would be very attractive candidates not just to understand the structural alteration that lead to their adaptation to these abiotic factors, but also for their potential use in industrial and biotechnological applications. In this work we established an ATII-LCL metagenomics dataset of potential biomass and cell wall degrading enzymes. Out of 1,337,597 pyrosequencing reads, a total of 28,547 contigs were assembled using Newbler GS assembler version 2.6. A total of 58,124 predicted open reading frames (ORFs) were identified using Metagene Annotator program. We searched the 58,124 ORFs for domains that matched to cell wall and biomass degrading enzymes using the Pfam database Version 26. The 53 matched sequences were confirmed by BLASTx search against NCBI nr database. Upstream regulatory elements, ribosome binding sequence, and secretory signal peptide sequences for secretion were checked for their presence. Additionally, halophilicity based on high prevalence of aspartic and glutamic acids was checked. Out of the 53 potential ORFs, only 14 presented a full-length coding sequence. We selected 4 ORFs with high similarities to cellulases, alpha galactosidase and cell wall lytic enzyme for further investigation. The four proteins have traditional signal peptide for secretion, and high occurrences of aspartic and glutamic amino acids when compared with non-halophilic orthologues. Moreover, the 3D structures of the four proteins were predicted and the relevant acidic amino acid residues were located on the surface of the molecules. Based on these features, we believe that the four proteins should have unique properties regarding stability in high saline solution, high temperature, and elevated concentration of heavy metals. Thus, this work established a dataset of the most abundant glycosyl hydrolases present in the microbial community of the ATII-LCL environment, and selected the most promising candidates for further molecular and catalytic characterization.

Introduction

Even though they are too tiny to be seen, microorganisms have been found as an essential part in human's life and in Earth's history generally. The physiological, biochemical and metabolic capabilities of microorganisms have played an important role in the climatic, biogeochemical and geological evolution of Earth (Newman & Banfield, 2002), (Xu, 2006). Currently, the microbial life nearly exists virtually in all the biological niches on Earth from seas and oceans to deserts, from the Arctic and Antarctica to the tropics, from underground mines to the top of high mountains and from the surface of hot springs to the underwater hydrothermal vents (Xu, 2006). These microbial communities are involved in synthesizing half of the photosynthetic biomass and producing vast amounts of oxygen (Field, Behrenfeld, Randerson, & Falkowski, 1998). They are also engaged in the sulfur, nitrogen, carbon and phosphorus cycles. It has been estimated that about only 1% of microorganisms are readily amenable to culturing by traditional laboratory techniques (Streit & Schmitz, 2004). Many methods that were meant to discover microbial diversity were not successful because of the bias occurring due to the culture methods limitations (Streit & Schmitz, 2004). Therefore, metagenomics was developed to address the uncultured microorganisms.

Metagenomics

Metagenomics is a powerful tool that has paved the way for studying microbial communities that cannot be cultured. It's based on the analysis of the extracted DNA from environmental samples that allows the identification of novel metabolic and physiological processes of the different microorganisms inhabiting this environment. By incorporating all the data obtained from this powerful genomics tool, the mood of living and interactions of these microbial communities can be explored and determined. Furthermore, new biological molecules can be discovered that can be of a great contribution to the therapeutic and biotechnological applications (Schmeisser, Steele, & Streit, 2007; Sharma, et al., 2008). In addition, Metagenomics have a variety of other purposes.

Metagenomics is being used to study different microbial communities in different microbial niches. Metagenomics projects with different purposes construct libraries of cloning vectors having the environmental DNA as inserts. These can be screened afterwards for genes of interest. Direct sequencing of the collected DNA is another alternative approach in order to obtain metagenomics datasets. Such techniques would provide us with an insight about the biochemical pathways of these

microbes. A summary is shown in Figure 1 to indicate the principal steps of some metagenomics projects.

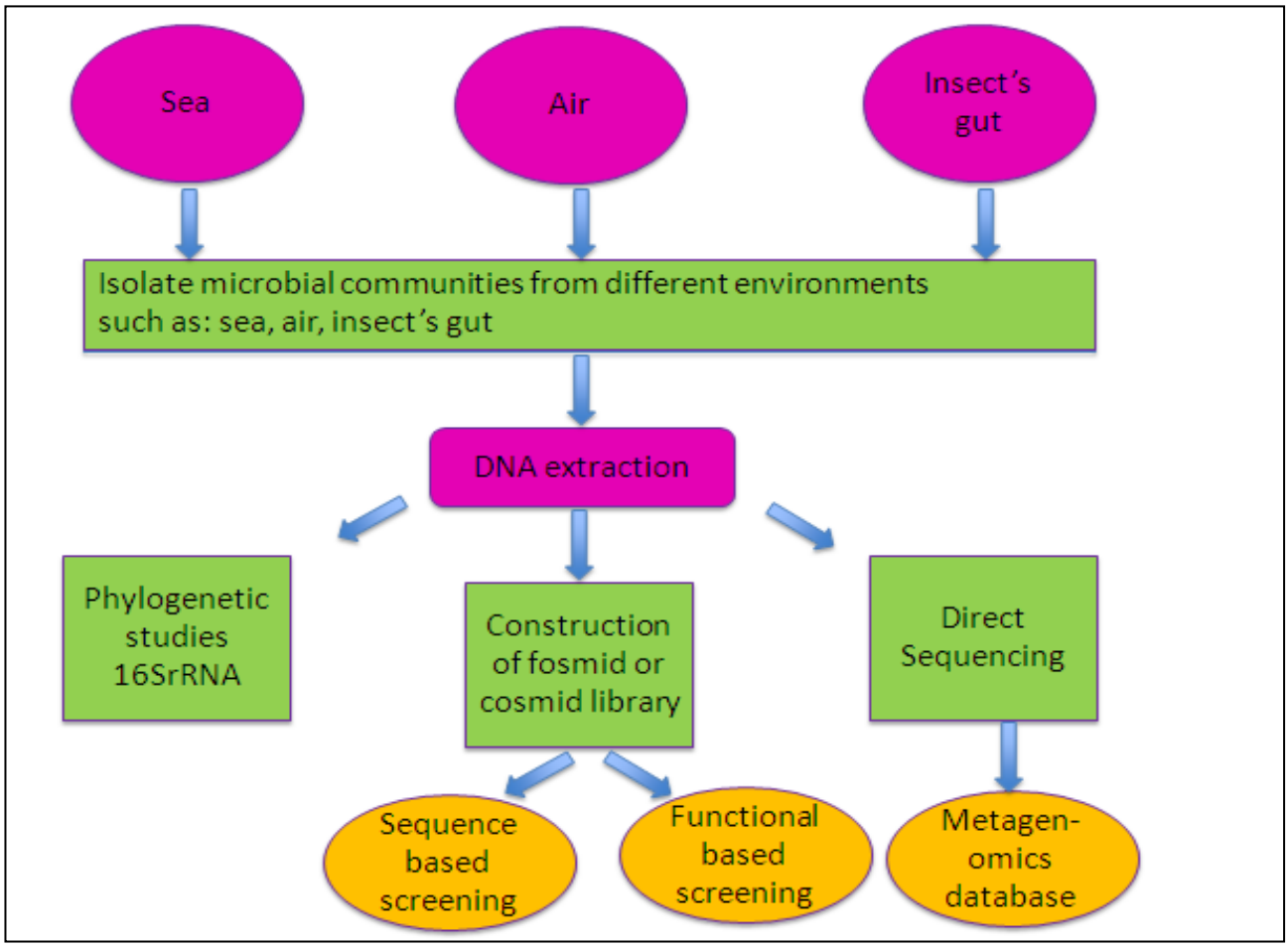

Figure 1. Summary of steps used in metagenomics approach.

Microbial communities are collected from environments of interest such as air, sea, soil or even an animal's or insect's gut. The DNA of the microbial community is then extracted and variety of metagenomics approaches is used to address different objectives as summarized in Figure 1.

Phylogenetic studies

Prokaryotic ribosomal RNA (rRNA) is an essential gene that is found at least in one copy in the genome. The 16S rRNA gene is a powerful molecular marker because of its ubiquity, evolutionary properties and universality that allows us to study microbial ecology (Case, et al., 2007). PCR amplifications of variable regions of the 16S rRNA genes are used to map uncultivable bacteria in different microbial niches (Y. Wang & Qian, 2009).

Primers are designed at the conserved regions in the 16S rRNA gene that flank a target variable region which will be used in the phylogenetic studies. Additionally, primers that target specific taxa can be designed as well (Teske & Sorensen, 2008).

Whole genome shot gun (WGS) metagenomic approach, dissimilar to the 16S rNA analysis, would give us an insight about the metabolic potential of bacterial community present in a certain environment. This is due to the analyses of the complete genetic information of these microbial communities. This is considered to be of great importance as it allows us to understand different ecosystems (Dinsdale, et al., 2008).

Construction of metagenomics libraries

Metagenomics projects tend to construct metagenomic libraries to screen for genes of interest, in which environmental DNA is cloned into vectors such as cosmids, fosmids or bacterial artificial chromosome (BAC). This technique was proposed by Norman Pace and colleagues, where they were able to construct a library using 16SrRNA genes from oceanic samples (Pace, Stahl, Lane, & Olsen, 1985). In extension to this idea, DeLong and his team were able to construct the first fosmid library using entire metagenomic DNA from oceanic samples as well (Stein, Marsh, Wu, Shizuya, & DeLong, 1996), where these F-factor based cosmids are meant to propagate large DNA fragments of uncultured microorganisms.

After the purification of the extracted DNA, it is cloned in the desired vectors. Since plasmids and small capacity vectors are not beneficial in screening for functional operons, large capacity vectors such as BAC and fosmids are preferred (Jo Handelsman, Mark Liles, David Mann, Christian Riesenfeld, & Goodman, 2002). Such vectors are then transformed into bacterial cells, such as *E. coli*, to be ready for the screening process. An alternative approach for constructing the library in such vectors is using expression vectors to directly obtain expressed genes.

The next step is the library screening for the desired gene or protein product. This screening process can be done by two alternative methods: either sequence based screening or functional based screening.

Sequence based screening is the screening for genes using conserved sequences of homologous genes already present in the database. This method uses hybridization probes or PCR primers to detect the desired genes through the conserved sequences (Schloss & Handelsman, 2003). This method could be used in the phylogenetic studies, where universal phylogenetic markers as 16SrRNA genes are used in relating a particular fragment of gene to a particular taxon. A famous example of this is when rhodopsins were successfully related to γ-proteobacteria after it was thought that it was only limited to Archaea (Beja, et al., 2000).

However, this method has some pitfalls: first, it is limited to known genes and no new genes are discovered. That is because the primers that are designed for screening are designed using already existing sequences that will favor the detection of previously known sequences. Functionally related families of genes that resulted from evolution, are not necessarily detected by the same set of primers (Riesenfeld CS, Schloss PD, & J, 2004).

The second pitfall is that only a small fragment of the desired gene will be detected using the primers, which will require further steps to retrieve the full length of the gene. This could be achieved by using primer walking or inverse PCR to obtain the full length genes (Mishra, Singla-Pareek, Nair, Sopory, & Reddy, 2002).

It is worth mentioning that there is another method, also depending on hybridization, which is microarray based method and known as “metagenomics profiling” in which the metagenomic library is hybridized with labeled DNA from reference strains or communities. This method allows for the rapid identification of clones that do not belong to known microbes (Sebat, Colwell, & Crawford, 2003).

Function based screening is initially done by identifying the clones that express certain activity or function. It overcomes the pitfalls of the sequence based analysis where it does not require the presence of the reference gene conserved sequences to be detected. Therefore, it can detect both previously known and novel genes with desired traits. Moreover, the results of the functional analysis is considered to be unambiguous, in contrast to the sequence based analysis which can give vague

results during annotation and could relate our sequence to poorly similar sequences or protein of unknown function in the database (Riesenfeld CS, et al., 2004).

The disadvantage of the functional screening is the probability of having a toxic product to the host which is *E. coli* in most cases. Besides, some genes can not be properly expressed due to the lack of the suitable regulatory elements that are related to transcription, translation or even the protein folding (Jo Handelsman, et al., 2002). This shortcoming could be solved in some cases by using alternative hosts other than *E. coli* such as *Streptomyces, Bacillus* or *Pseudomonas* (G. Y. Wang, et al., 2000).

As a result, and in order to extend the screening spectrum, other hosts are being used such as *Streptomyces lividans* and *Pseudomonas putida* (Martinez, et al., 2004). Also, it is expected that the incorrect folding of the desired protein is not achieved due to the absence of the required chaperones in the host strain. Unsuccessful expression of the desired gene may also be attributed to different codon usage, leading to low activities and missing clones containing the gene of interest. Currently, many laboratories are trying to find solutions to these challenges by finding new vectors and strains and by figuring out more sensitive methods for the screening process (Wolfgang & Ruth, 2004). Screened genes resulting from either approach are then sequenced to obtain the sequences of the genes of interest.

Direct Sequencing

Recently, the advances in the sequencing technologies made shotgun sequencing of environmental DNA feasible. The majority of the metagenomics projects apply the whole genome shotgun sequencing for the cloning and sequencing of environmental DNA samples. In this method, small inserts of genomic libraries are being sequenced by Sanger conventional technique to yield reads of lengths about 600-900 bp. The drawbacks of this approach are the cost and the possible bias that may occur during the construction of the library (Kennedy, et al., 2010).

A more recent approach is done by using 454 pyrosequencing which has decreased costs of sequencing compared to that of Sanger. This can be employed by directly sequencing the extracted DNA from the collected environmental samples with no need to the cloning step (von Bubnoff, 2008). Improvements are being made to the 454 pyrosequncing methods in order to achieve a higher read length with high accuracy leading to vast information generated by this method of sequencing (Wommack, Bhavsar, & Ravel, 2008).

Marine metagenomics

The marine environment is considered to be the largest habitat on earth as it constitutes about 70% from the earth's surface. The conditions of this habitat vary from sunlit surface to very deep trenches, and high pressures. The temperature ranges from ice waters in the Polar Regions to high temperatures at deep hydrothermal vents. Microorganisms inhabiting this environment originate from the three domains of life; Archaea, Bacteria and Eukarya. All these varieties rendered the marine microbial communities among the first attractive communities to be investigated by culture independent approaches (Giovannoni, Britschgi, Moyer, & Field, 1990).

Marine metagenomic studies aim to explore the microbial communities present at this environment. 16SrRNA analyses are used in studying the abundance and the diversity of the marine microorganisms. As an example, studying different phylogenetic markers, such as RecA/RadA and heat shock protein 70 (Hsp 70) at the Sargasso Sea surface, done by Venter and his team, showed nine different bacterial phyla and two archeal phyla (Venter, et al., 2004).

The different environmental conditions and challenges present in such an environment has led to diversity within the microorganisms residing in it, including signaling pathways and metabolic capabilities (Ellegren, 2008). In a study done by Dinsdale *et al*; 2008, nine different biomes were explored, most of which were marine environments showing us the metabolic capabilities of the microbial communities and the different lifestyles adopted by these communities. For example; they revealed that there is a higher abundance of genes that are involved in respiration in the coral associated microbes than commensal ones living in association with terrestrial animals (Dinsdale, et al., 2008). This can be referred to the diurnal respiration mode adopted by these communities where their environment is saturated with oxygen at the day, and becomes anaerobic at night (Shashar, Cohen, & Loya, 1993). A second example is the presence of sulphur metabolism in microbes associated with aquaculture fish. This is owed to the presence of organic and inorganic sulphur as food supplements to aquaculture fish (Iwanicka-Nowicka, Zielak, Cook, Thomas, & Hryniewicz, 2007). These examples demonstrate the high power of the metagenomics studies which can show the emergent biological properties of different environments.

Since the deep sea waters provide a huge microbial biome, it was important and interesting to explore its community structure. In favor of that purpose, a total of 200Mbp were obtained by WGS from the microbial community at Station ALOHA at the Pacific ocean. This data was then compared to other

WGS sequenced data to have an insight about the lifestyle and the functional activities of this microbial community (Konstantinidis, Braff, Karl, & DeLong, 2009).

A similar study was done where a fosmid library was constructed from 3,000 m deep Mediterranean plankton and analyzed by two methods; direct sequencing and phylogenetic analysis of 16SrRNA. It was observed that the assembled genes belonged to organisms of classes like alphaproteobacteria, planctomycetes, acidobacteria and others. This data was compared to the data obtained from ALOHA and a similarity was found. This suggests that at deep water columns, and in the absence of light, temperature is an important factor where several chemolithtrophic metabolic pathways are adopted to degrade organic matter found in such habitats (Martin-Cuadrado, et al., 2007).

Similarly, microorganisms present in extreme conditions in the marine environment are adapted to extremes of pressure and temperature resulting in novel metabolic pathways and may provide novel biomolecules that may be suited for industrial applications.

Metagenomics applications

Metagenomic libraries have been used to mine for novel genes that can be of a significant importance in industry or the pharmacological field. Scientists have been successful in finding a number of biotechnological products. These include lipases, metalloproteinases with high optimal temperatures as well as esterases (Chu, He, Guo, & Sun, 2008). These esterases have high tolerance for high salt concentrations, high pressure and organic solvents, rendering them as enzymes with interesting characteristics that can be used in industrial applications. A number of antibiotics such as Turbomycin A and B which exhibit an antibiotic activity against a wide range of gram positive and gram negative organisms were also discovered from soil metagenomes (Gillespie, et al., 2002).

Metagenomics facilitate the study of extreme environments and the organisms present in such environments helping us to understand their biogeochemical cycles and lifestyles. Marine microorganisms as well are receiving a great attention since they can produce novel compounds that could be promising in the field of medicine and pharmacology. Antimicrobial metabolites, amides of isoleucine derivatives that were isolated from *Acremonium furcatum*, were shown to have antimicrobial activity against *Bacillus subtilis, E. coli* and *Staphylococcus aureus.* This is of great importance to overcome the ability of pathogens in resisting antibiotics (Gallardo, et al., 2006).

Microbes are also known to have symbiotic relationships with numerous marine organisms such as sponges, squids, tunicates and others. Symbiotic relations are to provide advantage to both partners. For example microbes in relation with sponges provide them with food, process their waste and provide them with secondary metabolites that help them in defense mechanisms. These secondary metabolites are pharmaceutically valuable compounds that can be used as pesticides and herbicides (Kennedy, Marchesi, & Dobson, 2008).

Additional example where marine metagenomics is successfully employed is the identification of two alkane hydroxylase genes from the Pacific. These two genes were identified from a metagenomic library and were expressed in *Pseudomonas fluorescens* strain. The identification of these new genes would probably increase the applications of the alkane hydroxylases and emphasizes once more on the importance of the metagenomics applications (Ahn, et al., 2003).

Red sea and Atlantis II brine pool

Red Sea is a part of the large rift valley of the continental crust between Africa and Asia. It lies in the fault depression between two Earth's crust blocks; the Arabian and the North African, formed million years ago (William B.F. Ryan & Schreiber, 2012) (Cochran, 2008). This basin contains approximately 25 depressions deeps that contain brine pools. Some of these brine pools contain hot and salty water making its water denser than the normal seawater. This creates distinctive layers between the brine pools and the sea. The hottest and the saltiest waters are always among the deepest depth and closest to the sediments. It is believed that the waters acquired it high temperature due to the perfusion of the waters inside the rock fissures, dissolving minerals and gets heated up by the magma activity lying beneath. By the current flow effect, the water is driven up again to the brine pools (Cochran, 2008).

Deep hypersaline anoxic basins are regarded as one of the harshest environments conditions. Nevertheless, it's the least explored. These pools are frequently anoxic, from the restricted oxygen diffusion and renewal through the brine pool. In addition to the hypersalinity, brine pools form characteristic layers of sharp gradient in the temperature, pH, salinity, density and oxygen amounts. These layers are well separated and stabilized by the salinity gradient and destabilized by the heat underneath. So, certain brine pools consist of layers with distinct temperature and salinity changes starting from the sea-brine interface (Antunes, Ngugi, & Stingl, 2011).

These distinctive layers within the brine pool allows for the presence of different biological niches for microbial growth. Besides, the difference in the density gradient acts naturally as a trap for inorganic and organic decaying matters which provide this area with nutrients (Antunes, et al., 2011).

Atlantis II brine pool is a submarine basin and the largest one found in the Red Sea where it consists of 60 km^2 depression. It is sited at latitude 21°23′ N and longitude 38°04′ E (figure 2). It is about 2,194 meters deep (André Antunes, et al., 2011; Atlantis II Deep," 2011). It is located in the axial rift of the Red sea where it consists of a main basin and a smaller basin that are connected through a narrow channel. Other brine pools such as Discovery and Chain deeps are also shown to be connected to Atlantis II through subsurface connections (Faber, et al., 1998).

It is one of the most remarkable areas in the Red Sea as it contains water of high temperature reaching 68°C in its deepest layers and pH 5.3 (Antunes, et al., 2011). It is characterized by high salt levels of 270 parts per thousand (270 grams/1000 grams of water), which is about 7.5 times more than normal sea water (In the open ocean the range of salinity is generally from 32 psu to 37 psu; http://podaac.jpl.nasa.gov/SeaSurfaceSalinity). It has a thermodynamic nature owing to the increase in its temperature from 55.9°C in 1965 (Ross, 1972) to 67°C in 2000 (Winckler, et al., 2000).

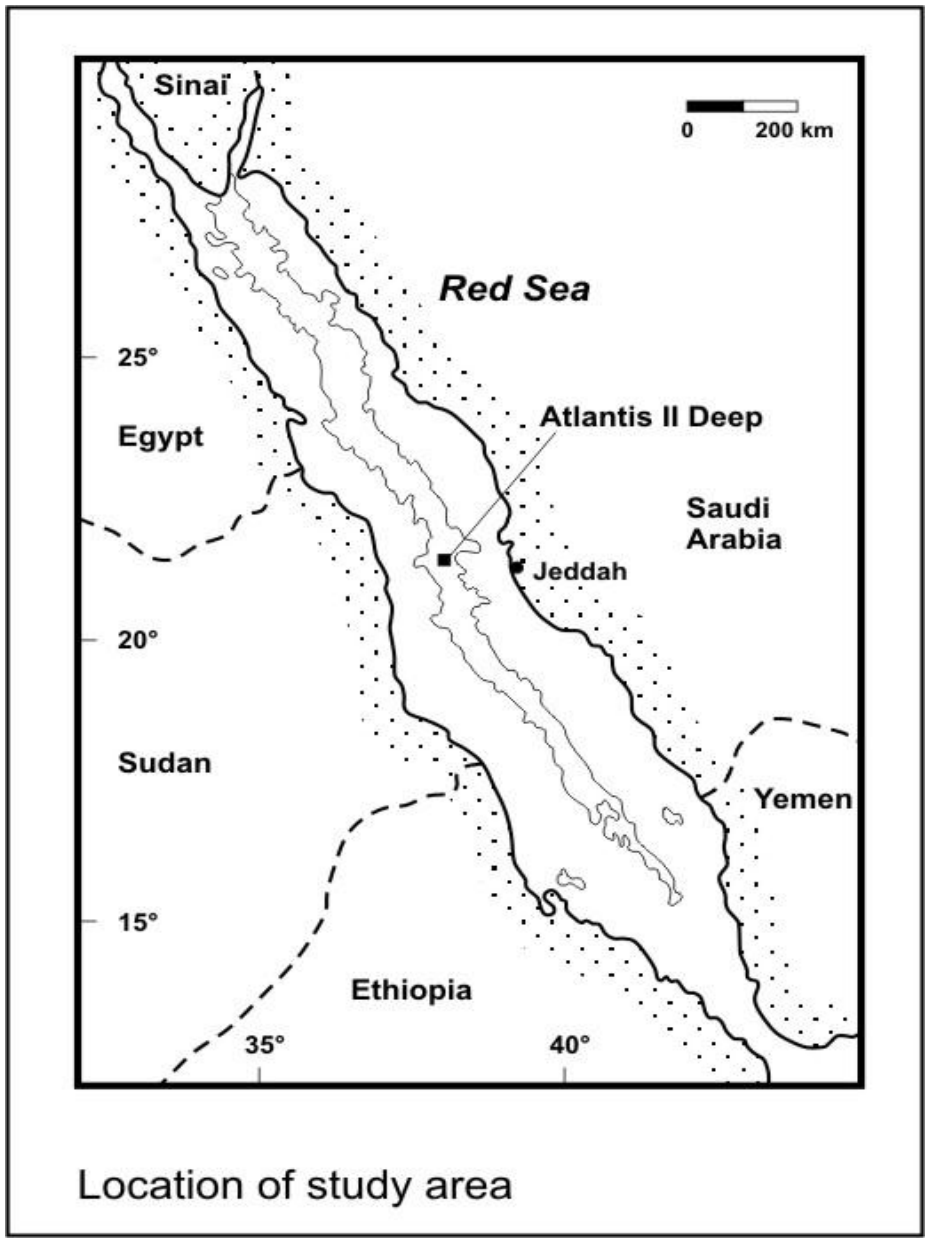

Figure 2 Atlantis II brine pool location in the Red Sea, the area of study, latitude 21°23′ N and longitude 38°04′ E. Picture adapted from www.iml.rwth-aachen.de.

The Atlantis II brine pool consists of interface, upper convective layers (UCL) and a lower convective layer (LCL). Metals such as copper, zinc and cobalt are present in concentrations of thousand times more than those present in normal seawaters ("Atlantis II Deep," 2011). This is thought to be as a result of the interaction between the oxidizing, weakly alkaline water of the Red sea that led to the deposition of the heavy metals in the brine pools (Antunes, et al., 2011). This vast amount of trace metals is enough to inhibit the activity of enzymes that are not well adapted to these conditions. The

studying of these unexplored areas, although still limited, give an insight about the large biodiversity of microbes present (PY Qian, 2011). Therefore, screening for enzymes in such harsh conditions seems appealing to find novel enzymes with extraordinary properties. These enzymes are expected to be suited for industrial purposes as they will be thermal-resistant, salt-tolerant and uninhibited by high concentrations of heavy metals.

Glycosyl Hydrolases

Glycosyl hydrolases, some of which are also known as glycosidases, are a tremendous group of enzymes that catalyze the cleavage of the glycosidic bond between two glycones or between a sugar and non-sugar moiety. The glycosidases are classified as EC3.2.1 according to the International Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). They are involved in more than one process such as the degradation of biomass like cellulose and hemicelluloses, in antibacterial activity like lysozyme and in normal cellular functions. They are found essentially in all domains of life (Gillespie, et al., 2002).

They are classified according to their action as exo- if they act on the terminal ends or endo- if they act in the middle of the molecule. They also can be classified according to the stereochemistry of the resulting product into inverting or retaining enzymes (Davies & Henrissat, 1995).

Regarding the sequence similarity, protein folding and catalytic amino acids, glycosidases have been grouped into more than 100 families (GH-number) in the cAZY database which is a specialized database dedicated to display the structure and biochemical information of the carbohydrate active enzymes (Cantarel, et al., 2009). This helps in the prediction of the catalytic machinery and mechanism of newly sequenced proteins. It is different from the International Union of Biochemistry (IUB) classification system, which is based on substrate specificity. Enzymes within this classification having different substrate specificity may be members of the same family, indicating some sort of evolutionary divergence that occurred to obtain new specificities. On the other hand, enzymes that act upon the same substrate may fall into different families (Gillespie, et al., 2002).

A higher level of classification is proposed, which is based on the tertiary structure of the enzyme. These groups are called clans. Members of the same clan are believed to share a common ancestor (Henrissat & Bairoch, 1996).

As stated by the Genomes OnLine Database (GOLD), 334 metagenomics projects are finished or at different stages of the sequencing process. Mining diverse environments is interesting as it results in obtaining novel biocatalysts that can be used in industrial applications. Cellulose is known to be the largest biomass available in nature, and is considered a source for biofuel production. Cellulolytic enzymes were successfully isolated from different environments throughout metagenomics projects, such as soil, lakes, rabbit's cecum and so on (Li, McCorkle, Monchy, Taghavi, & van der Lelie, 2009).

In most metagenomics projects, family GH13 was the most abundant. This family includes many activities such as α-amylase, isoamylase, α-glucosidase, oligo-1,6-glucosidase, pullulanase, cyclomaltodextrinase, maltotetraose-forming α-amylase, dextran glucosidase, trehalose-6-phosphate hydrolase, maltohexaose-forming α-amylase, maltotriose-forming α-amylase, maltogenic amylase, neopullulanase, malto-oligosyltrehalose trehalohydrolase, limit dextrinase, maltopentaose-forming α-amylase, amylosucrase, sucrose phosphorylase, branching enzyme, cyclomaltodextrin glucanotransferase (CGTase), 4-α-glucanotransferase, isomaltulose synthase, trehalose synthase ("Carbohydrate active enzymes," 2012).

The second most prevalent family was found to be GH23, containing lysozyme type G and peptidoglycan lyase. that GH2- β-galactosidase and β-manaosidase- and GH3- β-glucosidase , α-L-arabinofuranosidase and xylan 1,4-α-xylosidase; were also found to be widespread in many environments (Li, et al., 2009).

Cellulases

Glucose is the most commonly used constituent in food industry as well as in the production of ethanol and other chemicals. Cellulose, together with sugar cane and starch are considered to be the main sources for glucose production in industry. It is almost exclusively found in plants' cell walls, although some animals produce it such as tunicates, as well as some bacteria (Mba Medie, Davies, Drancourt, & Henrissat, 2012). Although, it differs and depends on the plant type, typically plants have high content of cellulose constituting their dry weight. That is why there is an increased demand for the utilization of the agricultural land to provide us with our tremendous needs for these valuable products. In order to eliminate the increased violation of the agricultural land for the sake of food industry, the use of forestry wastes in biofuel production has been proposed (Jeng, et al., 2010).

However, cellulose does not exist in a pure form except rarely as in cotton balls. It is usually associated with a mixture of hemicelluloses, pectin and lignin (Yarbrough, Himmel, & Ding, 2009) and this is the reason why it is difficult to be utilized in the production of bioethanol. Cellulose is a β-1,4- linked polymer that is aggregated by hydrogen and van der Waal's bonds to form parallel chains. Hemicellulose is found in the form of branched polysaccharide chains with hydrogen bonds to the surface of the cellulose fibrils. Pectins consist majorly of galacturonic acid and in some conditions it has more complex structures when rhamnogalcturonan and xylogalacturonan are present with variable side chains (Mohnen, 2008, Mba Medie, et al., 2012). Lignin fills the space between cellulose, hemicelluloses and pectin in the plant cell wall (Mba Medie, et al., 2012). In addition to that, the plant cellular structure itself plays a role in the utilization of the cellulosic biomass as well (Lynd, Weimer, van Zyl, & Pretorius, 2002). These are considered additional limitations over the structure of the cellulose itself to its exploitation.

Cellulose is hydrolyzed in three steps using three kinds of cellulases: 1-**Endoglucanase;** which breaks the crystalline cellulose units randomly. 2- **Exoglucanase;** which cleaves the cellobiose units from the cellulose polymer chain. 3-**β-glucosidase;** which breaks up the β-1,4-linked cellobiose into smaller monomers of glucose (Jeng, et al., 2010) (figure 3).

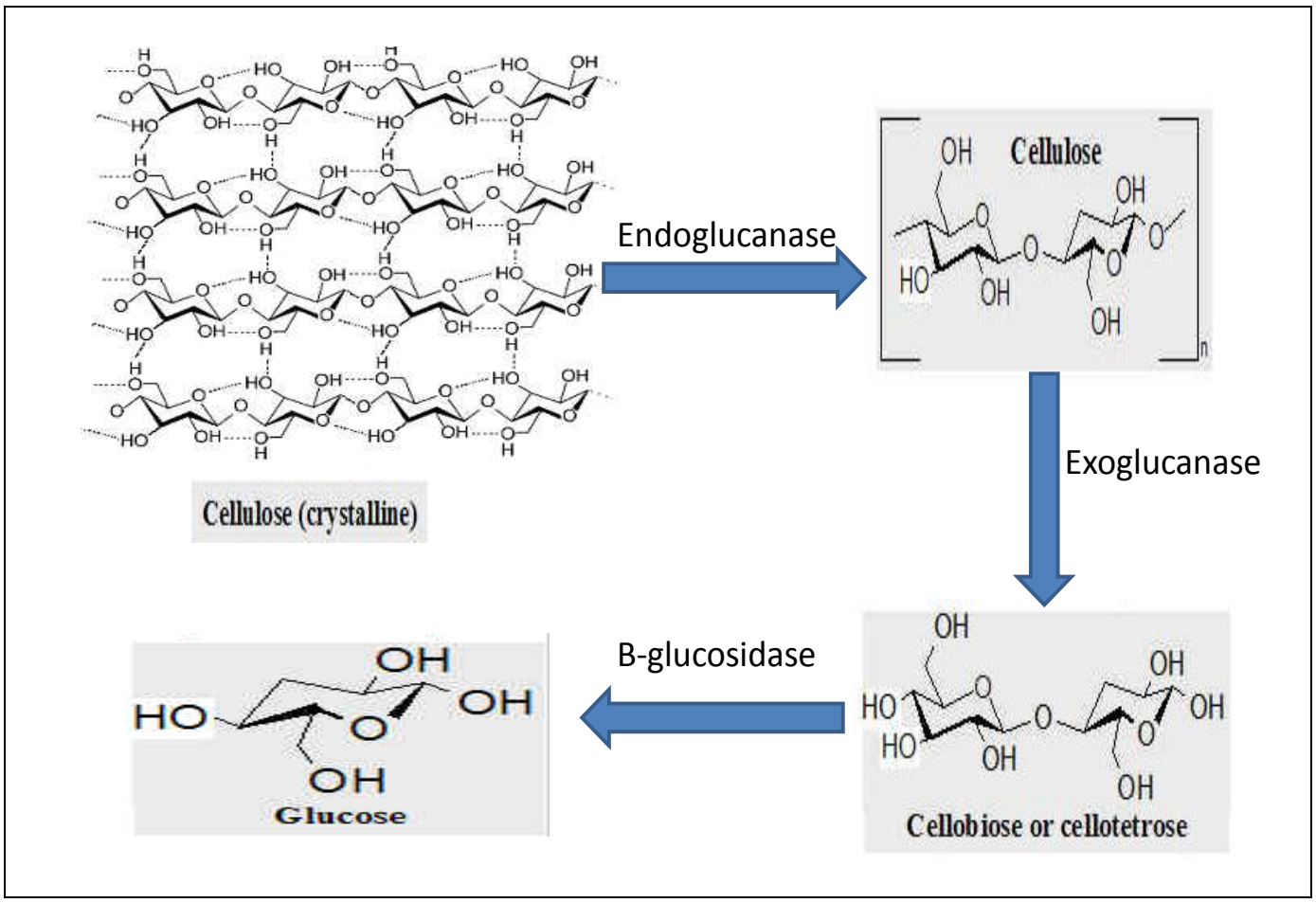

Figure 3. Cellulase enzymes action on cellulose. First, endoglucanases act internally on the amorphous cellulose yielding new ends. Then exoglucanases act upon these new ends resulting in cellobiose which is a substrate for beta glucosidase yielding glucose units.

The production of biofuel from lignocelluloses usually requires four steps: pretreatment, saccharfication, fermentation and finally separation. The potential of saccharification is its low energy consumption, low disposal of wastes and no corrosion occurring to the equipment (Z. Wang, 2010). The enzymatic hydrolysis of cellulose into glucose is considered the major limitation in the conversion of lignocelluloses to glucose. During the conversion process, insufficient activity of β-glucosidase will lead to the accumulation of cellobiose product and inhibition of endoglucanase and exoglucanase enzymes. In the overall route of cellulose degradation, the accumulation of glucose leads to decreasing the effectiveness of enzymatic saccharification. Therefore, it was proposed that

saccharification of lignocelluloses followed by fermentation into bioethanol would solve the problem of the products inhibition caused by the accumulation of glucose (Lynd, et al., 2002).

Among bacteria utilizing cellulose, there are different strategies in aerobic and anaerobic bacteria. Anaerobes -except few of them- act through a complex system known as cellulosome (Shoham, Lamed, & Bayer, 1999). Several anaerobic species do not release measurable amounts of cellulases into the media; instead they are anchored to the cell surface.

The challenge associated with the degradation of cellulose is the nature of the cellulose itself, where it is insoluble and crystalline in structure. A general feature that is known to most of the cellulase enzymes is the presence of carbohydrate binding module (CBM). This helps in the binding of the substrate and its solubilization by bringing the enzyme's catalytic domain in close proximity to the insoluble cellulose. In addition to that; CBM is suggested to have another non-catalytic function known as the "sloughing off" where it removes cellulosic fragments from cellulosic surfaces, consequently enhancing the enzymatic activity (Ding, et al., 2001).

Microorganisms secrete multiple enzymes in order to be able to digest a specific substrate. This is called an enzyme system. An example to that are the previously noted cellulase enzymes- exoglucanase, endoglucanase and β-glucosidase in order to digest cellulose. It is worth mentioning that this system is also active on hemicelluloses. The cellulase systems produce higher activity when they act collectively than when each enzyme acts individually. This phenomenon is known as synergism. There are four systems of synergism: a) endo-exo: between both endoglucanases and exoglucanases b) exo-exo: between exoglucanases acting upon reducing and non-reducing ends of the cellulose polymer. c) exo-β-glucosidase which hydrolyzes the cellobiose. d) intramolecular synergy between the CBM and the catalytic domains (Ding, et al., 2001).

There is more than one system for the mechanism of action of the cellulase systems. It is suggested that in *Trichoderma reesi*, a filamentous aerobic fungi, and aerobic bacteria such as *Cellulomonas* protruding hyphae aids in peeling off cellulosic structures and delivers cellulosic enzymes locally. This system is called non-complex system and the protein complexes are not high molecular weight. Anaerobic bacteria have limited amounts of ATP and as a result do not have the luxury of spending large amounts of energy secreating enzymes into the media. Instead, the enzymes stay anchored onto the cell surface in high molecular weight complexes, known as cellulosomes (Lynd, et al., 2002).

Non-complex systems: Aerobic fungi have been extensively studied and are used tremendously in industrial applications of cellulases especially *Trichoderma reesi*. The inducible cellulase system of *T. reesi* has two exoglucanases, five endoglucanases, and two β-glucosidases where all act in synergy to effectively degrade the cellulosic biomass. It produces about 0.3 g of protein per gram of substrate. The exoglucanases, also known as cellobiohydrolases, have tunnels in their three dimensional structure which accommodate the crystalline cellulosic cleavage from the reducing and the non-reducing ends giving rise to cellobiose (Divne, et al., 1994). Both exoglucanases are essential for the hydrolysis of microcrystalline cellulose. However they are not efficient in reducing the polymerization of the cellulose. Endoglucanase have been thought to have the ability to digest the cellulose chains at amorphous regions, yielding cellulosic chains that can be further processed by the action of cellobiohydrolases (Teeri, et al., 1998). The reason for the presence of five endoglucanases in *T. reesi* and the synergy between them has not yet been clearly explained. Part of the difficulty that faces this step is the inability of the endoglucanases to act on purified cellulose in the lab. But it is thought that the presence of CBMs is not essential for their activity or for the synergism (Z. Wang, 2010).

T. reesi produces two β-glucosidases which act on the hydrolysis of cellobiose and other small oligosaccharides into glucose. Large fractions of these enzymes are found to be bound to the fungal cell wall despite the ability to isolate them from the culture supernatant. They limit the release of glucose units to the media following the hydrolysis of the cellulose due to their close proximity to the fungal cell wall. β-glucosidases of *T. reesi* are subjected to inhibition by glucose and they are produced at lower levels than other fungi such as *Aspergillus* species. *T. reesi* β-glucosidases are sufficient to act upon cellobiose but not enough for extensive cellulose saccharification. On the other hand, *Aspergillus* β-glucosidases are more glucose tolerant, they are not inhibited by the release of glucose, and they have the capability to act extensively on cellulose saccharfication. On large scale industrial applications for cellulose saccharification, supplementation of *T. reesi* β-glucosidases with β-glucosidases from *Aspergillus* is done to obtain the desired results (Reczey, Brumbauer, Bollok, Szengyel, & Zacchi, 1998).

Complex cellulase systems: Organisms with complex cellulase systems (cellulosome) are usually present in anaerobic environments in consortia with other microorganisms that are cellulolytic and non cellulolytic. The cellulosome allows the consorted enzymes to act in proximity to allow the

synergy effect between them. It also minimizes the space through which the hydrolyzed products have to diffuse leading to the efficient uptake of the products by the bacterial cell (Schwarz, 2001).

Cellulosomes are protrusions on the cells that are growing on cellulose. These protrusions contain the enzymes that are bound hardly to the cell wall surface but at the same time have enough flexibility allowing them to bind to the microcrystalline cellulose (Schwarz, 2001) (figure 4). They are relatively large, stable complexes that range from 2 to 16 MDa in some species. In others, the size may reach up to 100 MDa in case of aggregation of cellulosomes into more complex polycellulosomes (Schwarz, 2001). They are extensively glycosylated, where these glycosyl groups may play a role in the cellulosome protection against proteolysis. Cellulosome mainly consists of:

- Scaffoldin subunit: contains multiple cohesion modules that are connected to other functional molecules. It also may contain CBM, dockerin and Surface layer homology (SLH) molecule which acts as an anchor.
- Cohesion modules: are the major components of the scaffoldin subunit and play an important role in the organization of the cellulolytic subunit in the complex system.
- Dockerin modules: they are important in anchoring the catalytic enzymes to the scaffoldin (Fontes & Gilbert, 2010).

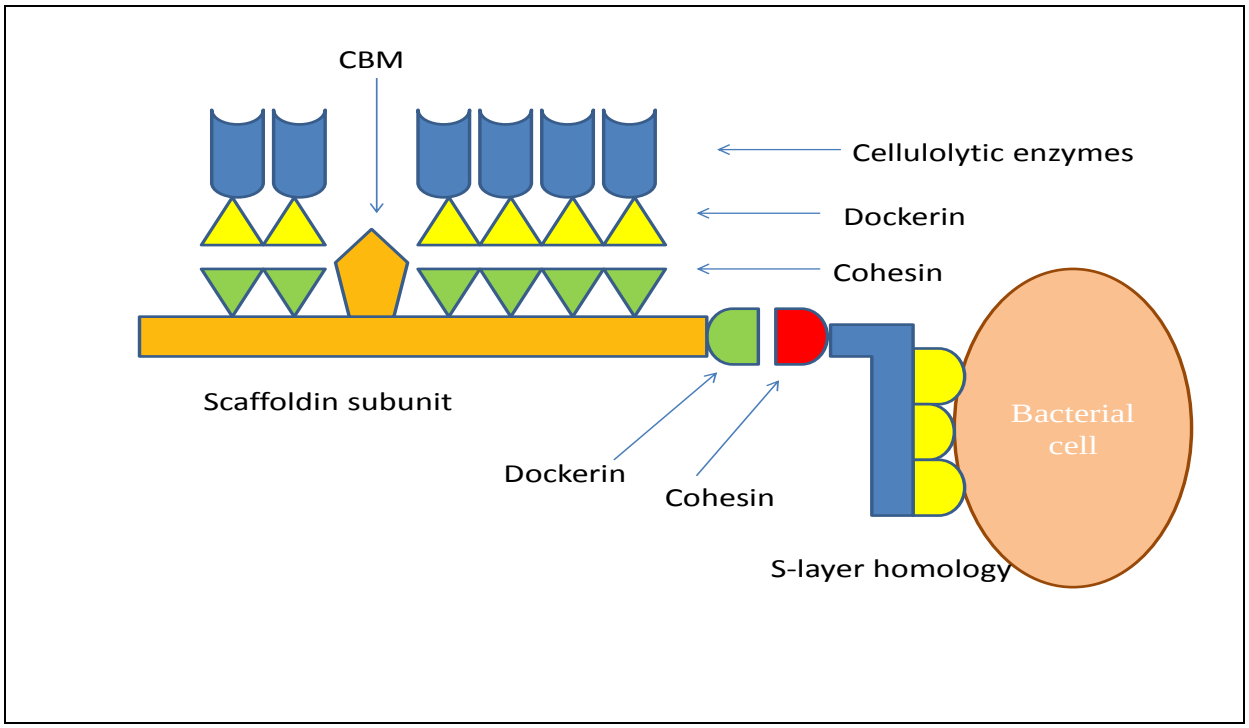

Figure 4. Cellulosome structure. SLH projects out of the cell connecting it to the dockerin through the cohesion. The scaffoldin holds cohesin type I modules that interact with the cellulolytic enzymes through dockerin type I domains.

Preparation of cellulosome from *Clostridium thermocellum* is highly capable of hydrolyzing microcrystalline cellulose. This high efficiency is ascribed to a) the appropriate ratio present between the catalytic domain which enhances the synergy between them b) the appropriate spacing between the components to additionally favor the synergy and c) the presence of different enzymatic activities to allow the hydrolysis and the removal of other polysaccharides or physical obstacles present (Lynd, et al., 2002).

Under the electron microscope, cellulosomes appear as fist like structure that open upon attaching to microcrystalline cellulose surfaces leading to the spread of the catalytic domains. A stagnant area exists between the cellulosome and the cell wall where contact corridors are present preventing the present oligosaccharides from diffusing to the environment and keeping them in close proximity to the cell (Bayer, Chanzy, Lamed, & Shoham, 1998).

Applications of Cellulases

Energy availability and access bring about many of our current challenges including cost, environmental quality, and continuity. In the light of the new technologies, energy conversions become one of the most significant and interesting ideas. Conversion of lignocellulosic biomass into energy is well suited for this kind of energy production, where it has the advantages of being available at large scale, renewable, low cost and environmentally friendly. Cellulases have many applications in the field of industry. Particularly, energy production based on the cellulosic biomass conversion has nearly zero emission of the greenhouse gases (Lynd, van Zyl, McBride, & Laser, 2005).

It is also involved in other industrial applications as detergents and textile industry, where it is involved in the finishing of cotton fabrics and stone washing of jeans. Additionally, it is included in pulp and paper industry where in combination with hemicellulases, it helps in the drainage and the operation of the machines and aids in the deinking of the fibers (Cao & Tan, 2002).

The success in isolating cellulolytic enzymes will be of a great interest, especially from environments of high temperature and pH similar to that of the Atlantis II brine pool. Such enzymes are expected to be adapted to the harsh conditions of the industrial processes and therefore can be used on large scale degradation of plant wall biomass into fermentable sugars which can be consequently used in the production of cheap and renewable biofuel.

Objectives

Prediction of cell wall and biomass degrading enzymes from environments of high temperature and pH similar to that of the Atlantis II brine pool. Such enzymes are expected to be adapted to the harsh conditions of the industrial processes and therefore can be used on large scale

Materials and Methods

Establishing biomass and cell wall degrading enzymes dataset

Pyrosequencing of the LCL environmental DNA was performed in the department of Biology at AUC using Roche 454 GSFLX genome analyzer and GSFLX Titanium pyrosequencing kit. Reads generated were assembled using Newbler® GS assembler version 2.6 with the default options for overlayer consensus minimum overlap identity of 90% running under was biolinux version 6 system. The assembled contigs and searched for potential open reading frames (ORFs) using MetaGene Annotator (Noguchi, Taniguchi, & Itoh, 2008).

For establishing of a dataset containing all the potential cell wall degrading enzymes, we searched first the Pfam database v.26 (Punta, et al., 2011). Using "glycosyl hydrolase" as a keyword, we managed to obtain most of the glycosyl hydrolase families involved in the cell wall and biomass degrading enzymes. These family domains were used in performing HMM scan (Eddy, 1996) over our assembled reads, specifically ORFs. BLASTx (Altschul, Gish, Miller, Myers, & Lipman, 1990) was performed on these resulting ORFs on NCBI against the non redundant database (nr).

ORFs were then subjected to additional analyses to find the upstream sequences consisting of the promoter region (-10 and -35 regions) using Bprom software online (www.softberry.com) and the ribosomal binding site which we searched for manually. Moreover, signalP (Petersen, Brunak, von Heijne, & Nielsen, 2011) software was used to identify the presence of a signal peptide in the beginning of each potential sequence to determine whether the protein is secreted or not. Full length protein coding sequences, that had no signal peptide were submitted to other analyses online softwares; PSORT (Nakai & Kanehisa, 1992) and SecretomeP (Bendtsen, Kiemer, Fausboll, & Brunak, 2005) to determine their localization and whether they are secreted in a non-classical pathway respectively. They were also submitted to TMHMM online (Krogh, et al., 2001), to determine whether they have transmembrane domains.

Additionally, halophilicity ratio was calculated for each ORF by calculating the ratio of aspartic acid and glutamic acid for the aligned regions between the ORF and its reference sequence. The alignment was readily obtained through the BLASTx results.

Afterwards, sequences with potential full length coding sequences were checked for their 3D structure using Swiss Model software online (Arnold, et al., 2006), and then they were visualized using

RasMol (Sayle & Milner-White, 1995) and Raswin freeware. The halophilic residues that were present in the ORFs but not in the reference sequences were highlighted in the 3D models.

A summary of the steps done to establish the database is shown in figure 5.

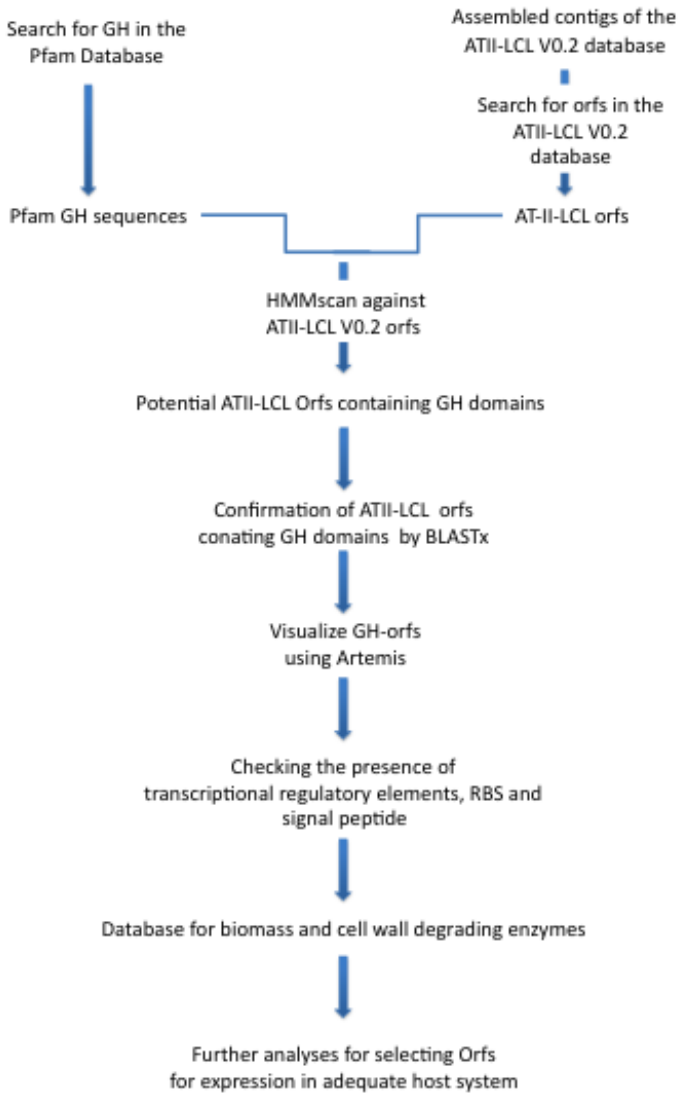

Figure 5. Diagram showing the steps employed in establishing database of the biomass and cell wall degrading enzymes.

Results and Discussion

Establishing biomass and cell wall degrading enzymes database

The pyrosequencing processes were performed in the Department of Biology at AUC. Two runs, each ½ 454 gasket were performed. Table 1 shows the data resulting from these two runs:

Table 1. Assembly of pyrosequencing reads of ATII-LCL samples

	1st assembly(V0.1)*	**2nd assembly (V0.2)****
Total No. of Reads	655,289	1,337,597
Total No. of Assembled Reads	527,359	1,142,385
Total No. of Partial Assembled Reads	53,649	91,561
Singletons	58,351	72,261
Largest Contig Size (bp)	105,705	236,358
Average Contig size	1,781	1,912
Total No. of contigs	20,493	28,547
No. of contigs >10 Kbp	173	221

*Assembly of 655,289 reads

**Assembly of the 655,289 reads (first run) in addition to 682,308 reads (second run)

As it is shown in table 1, the size of the largest contig in the second run has increased (almost doubled) than the first run. This shows that we were able to close more gaps in the assembly by sequencing more samples of ATII-LCL granting us more access to features of the genomes of the microbial community that reside in the LCL.

Upon using ‘glycosyl hydrolase’ as a keyword in the search at the Pfam database, about 70 Pfam accessions (supplementary data) were used to run the HMM scan. Table 2 shows the Pfam domains that were found in our ATII-LCL (V0.2) ORFs and the number of corresponding ORFs.

Table 2. Number of ATII-LCL (V0.2) ORFs that contain domains for biomass and cell wall degrading enzymes.

Family	no. of orfs
GH43	14
GH3	8
GH47	8
GH4	4
GH25	3
Cellulase	3
CBM_5_12	3
Dockerin_1	2
Alpha amylase	2
GH18	1
GH19	1
GH16	1
GH76	1
GH_2C	1
GH31	1

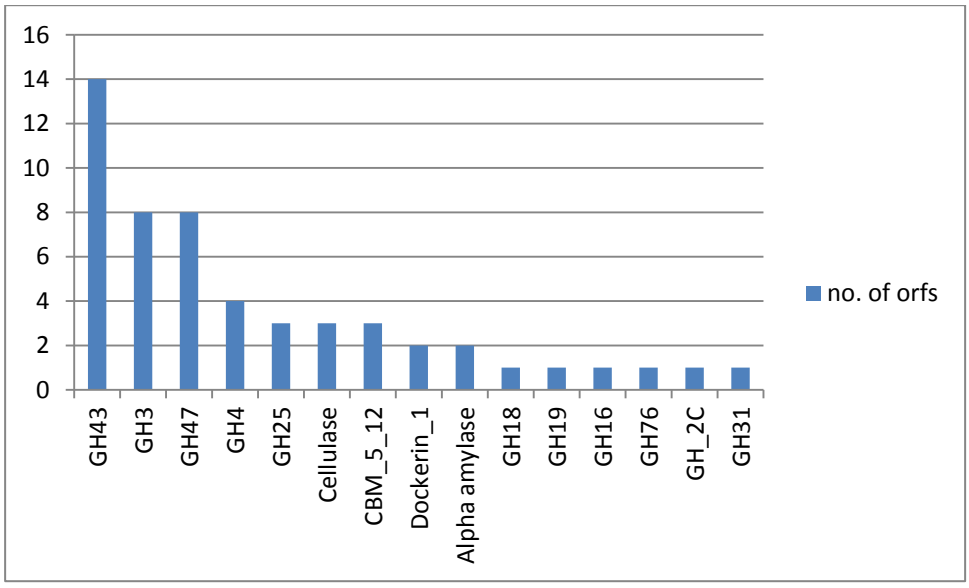

Figure 6. Graphical representation of number of ATII-LCL (V0.2) ORFs that contain domains for biomass and cell wall degrading enzymes.

As shown in Table 2 and Figure 6, not all the Pfam families related glycosyl hydrolases were found in our LCL V0.2 database. The GH43 family had the highest number of related genes identified in our LCL V0.2 database. The annotated activities of each identified pfam family is presented in Table 3.

Table 3. Description of the activities of each Pfam domain found in ATII-LCL (V0.2) database.

Family	Activity
GH42	β-galactosidase,β-fucosidase and α-arabinosidase.Active on lactose and transglycosylation, suggested in cell wall degrading enzymes.
GH2	β-galactosidases, β-glucuronidases, β-mannosidases, exo-β-glucosaminidases
GH3	β-D-glucosidases, α-L-arabinofuranosidases, β-D-xylopyranosidases and *N*-acetyl-β-D-glucosaminidases,some enz.have dual function
GH4	α-glucosidases, α-galactosidases, α-glucuronidases, 6-phospho-α-glucosidases, and 6-phospho-β-glucosidases
GH16	xyloglucosyltransferase, endo-1,3-β-glucanases , lichenases, xyloglucanases
GH18	chitinases and endo-β-N-acetylglucosaminidases and also sub-families of non-hydrolytic proteins that function as carbohydrate binding modules
GH19	chitinases
GH25	lysozymes, involved in the degradation of the carbohydrate backbone of peptidoglycan of bacteria
GH31	α-glucosidase; α-1,3-glucosidase; sucrase-isomaltase; α-xylosidase; α-glucan lyase, isomaltosyltransferase
GH43	α-L-arabinofuranosidases, endo-α-L-arabinanases and β-D-xylosidases
GH47	α-mannosidases
GH76	α-1,6-mannanase
CBM_5	Carbohydrate binding module

Protein similarity search using the BLASTx against NCBI nr database confirmed the identity of the previously identified ORFs. Table 4 shows each Pfam domain, the ORF that has this particular domain and its corresponding BLASTx result.

We further analyzed the identified ORFs for up-stream regulatory elements (-10 and -35), signal peptide for secretion, RBS, and start and termination codons for protein translation. The results are summarized in Table 5.

Table 4. Gene annotation of the identified biomass and cell wall degrading enzymes from ATII-LCL (V0.2)

Pfam	Contig*	Annotation
GH25	contig00001_gene27_29329_30348	cell-wall lytic enzyme [Rhizobium etli CFN 42]
	contig00013_gene83_95811_96638	glycoside hydrolase family 25 [Ochrobactrum intermedium LMG 3301]
	contig00052_gene11_9496_10698	glycoside hydrolase family protein [Mesorhizobium opportunistum WSM2075]
Cellulase	contig00016_gene83_77945_79123	cellulase [Yokenella regensburgei ATCC 43003]
	contig00711_gene2_315_1208	Beta-galactosidase [Agrobacterium tumefaciens F2]
	contig01010_gene1_100_996	putative glycoside hydrolase family protein [Treponema azotonutricium ZAS-9]
GH3	contig00024_gene67_52716_55364	Beta-glucosidase [Burkholderia ubonensis Bu]
	contig00027_gene9_6518_7543	beta-N-acetylhexosaminidase [Rhizobium etli CIAT 652]
	contig00219_gene7_8447_9466	glycoside hydrolase family 3 domain protein [Mesorhizobium australicum WSM2073]
	contig00222_gene10_7298_8347	beta-hexosaminidase [Cupriavidus metallidurans CH34]
	contig01616_gene1_1_1241	glycoside hydrolase family 3 domain-containing protein [Chlorobaculum parvum NCIB 8327]
	contig02816_gene1_1_147	glycoside hydrolase [Spirochaeta smaragdinae DSM 11293], thermostable beta-glucosidase
	contig02816_gene2_152_1711	beta-glucosidase-related glycosidase [Microbacterium testaceum StLB037]
	contig03408_gene2_1094_1663	beta-D-glucoside glucohydrolase [Cupriavidus metallidurans CH34]
GH4	contig00076_gene14_14397_15869	glycoside hydrolase family 4 [Rhizobium leguminosarum bv. trifolii WSM1325]
	contig00076_gene19_21986_23359	alpha-galactosidase [Sinorhizobium meliloti CCNWSX0020]
	contig03199_gene1_1_876	malate dehydrogenase [endosymbiont of Tevnia jerichonana (vent Tica)]
	contig06010_gene2_442_1068	malate dehydrogenase [endosymbiont of Tevnia jerichonana (vent Tica)]
GH47	contig01198_gene3_1740_2939	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497]
	contig01467_gene3_1054_1416	mannosyl-oligosaccharide alpha-1,2-mannosidase [Ajellomyces dermatitidis ATCC
	contig01467_gene2_506_1030	family 47 glycoside hydrolase [Phytophthora sojae]
	contig02102_gene1_1_959	glycosyl hydrolase [Trichomonas vaginalis G3]
	contig08990_gene3_330_788	endoplasmic reticulum mannosyl-oligosaccharide 1,2-alpha-mannosidase-like[Apis mellifera]
	contig14076_gene1_1_438	Mannosyl-oligosaccharide 1,2-alpha-mannosidase [Caulobacter segnis ATCC 21756]
	contig17322_gene1_1_345	mannosyl-oligosaccharide 1,2-alpha-mannosidase [Caulobacter segnis ATCC 21756]
	contig23658_gene1_1_179	mannosyl-oligosaccharide 1,2-alpha-mannosidase MNS1 [Arabidopsis thaliana]
GH16	contig01408_gene4_1735_2751	glycosyl hydrolase family 16 [Prevotella buccae ATCC 33574]

* contig00001_gene27_29329_30348:
Contig # Gene # Start End

Pfam	Contig*	Annotation
GH43	contig00376_gene3_1662_2519	glycosidase [Thermotoga petrophila RKU-1]
	contig00154_gene6_3285_4316	glycosyl hydrolase family 32 protein [Candidatus Solibacter usitatus Ellin6076]
	contig00338_gene4_1876_3174	glycosyl hydrolase family 32 [Bryantella formatexigens DSM 14469]
	contig00359_gene8_3930_5174	Glycosyl hydrolase family 32, N terminal domain protein [Candidatus Solibacter usitatus Ellin6076]
	contig00992_gene4_1283_2299	glycosidase related protein [Thermaerobacter subterraneus DSM 13965]
	contig01343_gene1_1_1207	Glycosyl hydrolases family 32 [Microcoleus chthonoplastes PCC 7420]
	contig01479_gene5_1406_2763	arabinan endo-1,5-alpha-L-arabinosidase [Thermobaculum terrenum ATCC BAA-798]
	contig01487_gene2_207_1508	glycosyl hydrolase family 32 protein [Candidatus Solibacter usitatus Ellin6076]
	contig02380_gene1_2_1366	Glycosyl hydrolase family 32, N terminal domain protein [Candidatus Solibacter usitatus Ellin6076]
	contig03592_gene2_379_1617	laminin G sub domain 2 [Thiorhodovibrio sp. 970]
	contig03733_gene1_1_1215	Glycosyl hydrolase family 32 domain protein [Runella slithyformis DSM 19594]
	contig04460_gene3_309_1391	glycosyl hydrolase family 43 protein [Bacteroides sp. D20]
	contig07434_gene1_1_933	glycosyl hydrolase family 32 protein [Candidatus Solibacter usitatus Ellin6076]
	contig12554_gene2_223_550	glycosyl hydrolase 32 domain protein [Marinithermus hydrothermalis DSM 14884]
GH18	contig00474_gene2_427_2676	glycosyl hydrolase, family 18 [uncultured marine bacterium 159]
GH19	contig00626_gene4_1836_2477	Lytic enzyme [Erwinia sp. Ejp617]
Dockerin	contig02598_gene4_1025_1962	hypothetical protein NatgrDRAFT_2304 [Natronobacterium gregoryi SP2]
	contig02630_gene2_295_1955	hypothetical protein ALOHA_HF4000ANIW141I9ctg2g4 [uncultured marine microorganism HF4000_ANIW141I9]
GH76	contig05280_gene1_1_1221	protein of unknown function DUF255 [Desulfatibacillum alkenivorans AK-01]
GH2_C	contig08149_gene1_1_863	Mannan endo-1,4-beta-mannosidase [Medicago truncatula]
GH31	contig08283_gene1_1_852	glucosidase protein [Ralstonia solanacearum PSI07]
Alpha amylase	contig08922_gene1_1_561	Alpha amylase catalytic domain found in Glycosyltrehalose trehalohydrolase (also called Maltooligosyl trehalose Trehalohydrolase
	contig08922_gene2_456_791	4-alpha-D-((1->4)-alpha-D-glucano)trehalose trehalohydrolase [Ralstonia sp.]
CBM_5_12	contig10990_gene1_57_642	putative fusion protein [Agrobacterium tumefaciens 5A]
	contig16312_gene1_1_381	hypothetical fusion protein [Agrobacterium sp. H13-3]
	contig16327_gene1_1_381	hypothetical protein [Agrobacterium sp. H13-3]

* contig00001_gene27_29329_30348:
Contig # Gene # Start End

Table 5. Transcriptional regulatory elements, secretion peptide, and sequences required for translation process identified in potential genes coding for biomass and cell wall degrading enzymes.

Pfam domain	Orf	Gene annotation	RBS	-10	-35	SP	Start	Stop
GH25	C00001_gene27_29329_30348	glycoside hydrolase family 25 [Ochrobactrum intermedium LMG 3301]	√	√	√	√	√	√
	C00013_gene83_95811_96638	glycoside hydrolase family 25 [Ochrobactrum intermedium LMG 3301]	√	√	√	X	√	TAA
	C00052_gene11_9496_10698	glycoside hydrolase family protein [Mesorhizobium opportunistum WSM2075]	P	√	√	X	TTG*	√
Cellulase	C00016_gene83_77945_79123	cellulase [Yokenella regensburgei ATCC 43003]	√	√	√	√	√	√
	C00711_gene2_315_1208	Beta-galactosidase [Agrobacterium tumefaciens F2]	P	P	P	X	√	TAA
	C01010_gene1_100_996	putative glycoside hydrolase family protein [Treponema azotonutricium ZAS-9]	√	√	√	√	√	√
GH3	C00024_gene67_52716_55364	Beta-glucosidase [Burkholderia ubonensis Bu]	√	√	√	X	TTG*	√
	C02816_gene2_152_1711	beta-glucosidase-related glycosidase [Microbacterium testaceum StLB037]	NA	NA	NA	X	GTG	√
	C00027_gene9_6518_7543	beta-N-acetylhexosaminidase [Rhizobium etli CIAT 652]	√	X	X	√	GTG*	√
	C00219_gene7_8447_9466	glycoside hydrolase family 3 domain protein [Mesorhizobium australicum WSM2073]	√	√	√	√	√	√
	C00222_gene10_7298_8347	beta-hexosaminidase [Cupriavidus metallidurans CH34]	√	√	√	√	√	√
	C01616_gene1_1_1241	glycoside hydrolase family 3 domain-containing protein [Chlorobaculum parvum NCIB 8327]	√	√	√	X	GTG	CT
	C02816_gene1_1_147	glycoside hydrolase [Spirochaeta smaragdinae DSM 11293], thermostable beta-glucosidase	NA	NA	NA	NA	CT	√
	C03408_gene2_1094_1663	beta-D-glucoside glucohydrolase [Cupriavidus metallidurans CH34]	P	√	√	√	√	CT
GH4	C00076_gene14_14397_15869	glycoside hydrolase family 4 [Rhizobium leguminosarum bv. trifolii WSM1325]	√	√	√	√	√	√
	C00076_gene19_21986_23359	alpha-galactosidase [Sinorhizobium meliloti CCNWSX0020]	√	√	√	√	√	√
	C03199_gene1_1_876	malate dehydrogenase [endosymbiont of Tevnia jerichonana (vent Tica)]	X	X	X	NA	CT	TAA
	C06010_gene2_442_1068	malate dehydrogenase [endosymbiont of Tevnia jerichonana (vent Tica)]	P	√	√	X	√	√
GH43	C00376_gene3_1662_2519	glycosidase [Thermotoga petrophila RKU-1]	√	√	√	X	√	TAG
	C00154_gene6_3285_4316	glycosyl hydrolase family 32 protein [Candidatus Solibacter usitatus Ellin6076]	√	√	√	X	√	TAG
	C00338_gene4_1876_3174	glycosyl hydrolase family 32 [Bryantella formatexigens DSM 14469]	P	√	√	X	√	TAG
	C00359_gene8_3930_5174	Glycosyl hydrolase family 32, N terminal domain protein [Candidatus Solibacter usitatus Ellin6076]	X	√	√	X	√	√
	C00992_gene4_1283_2299	glycosidase related protein [Thermaerobacter subterraneus DSM 13965]	X	√	√	X	√	TAA
	C01343_gene1_1_1207	Glycosyl hydrolases family 32 [Microcoleus chthonoplastes PCC 7420]	NA	NA	NA	X	X	X
	C01479_gene5_1406_2763	arabinan endo-1,5-alpha-L-arabinosidase [Thermobaculum terrenum ATCC BAA-798]	NA	NA	NA	X	CT	X
	C01487_gene2_207_1508	glycosyl hydrolase family 32 protein [Candidatus Solibacter usitatus Ellin6076]	P	√	√	X	√	TAA
	C02380_gene1_2_1366	Glycosyl hydrolase family 32, N terminal domain protein [Candidatus Solibacter usitatus Ellin6076]	NA	NA	NA	NA	CT	√
	C03592_gene2_379_1617	laminin G sub domain 2 [Thiorhodovibrio sp. 970]	NA	NA	NA	√	CT	TAA
	C03733_gene1_1_1215	Glycosyl hydrolase family 32 domain protein [Runella slithyformis DSM 19594]	X	X	X	√	X	√
	C04460_gene3_309_1391	glycosyl hydrolase family 43 protein [Bacteroides sp. D20]	√	√	√	√	√	X
	C07434_gene1_1_933	glycosidase PH1107-like protein [Meiothermus silvanus DSM 9946]	NA	NA	NA	√	X	X
	C12554_gene2_223_550	glycosyl hydrolase 32 domain protein [Marinithermus hydrothermalis DSM 14884]	√	√	√	√	√	X
GH18	C00474_gene2_427_2676	glycosyl hydrolase, family 18 [uncultured marine bacterium 159]	√	√	√	√	√	√
GH19	C00626_gene4_1836_2477	Lytic enzyme [Erwinia sp. Ejp617]	√	√	√	√	√	√
GH47	C01198_gene3_1740_2939	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497]	√	√	√	√	GTG*	√
	C01467_gene3_1054_1416	mannosyl-oligosaccharide alpha-1,2-mannosidase [Ajellomyces dermatitidis ATCC	√	√	√	X	√	√
	C01467_gene2_506_1030	family 47 glycoside hydrolase [Phytophthora sojae]	√	√	√	X	√	√
	C02102_gene1_1_959	glycosyl hydrolase [Trichomonas vaginalis G3]	√	√	√	√	√	X
	C08990_gene3_330_788	endoplasmic reticulum mannosyl-oligosaccharide 1,2-alpha-mannosidase-like[Apis mellifera	X	X	X	√	√	X
	C14076_gene1_1_438	Mannosyl-oligosaccharide 1,2-alpha-mannosidase [Caulobacter segnis ATCC 21756] Length=4	NA	NA	NA	√	X	CT
	C17322_gene1_1_345	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497]	NA	NA	NA	√	CT	X
	C23658_gene1_1_179	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497] Length=458	√	NA	NA	√	CT	X
GH16	C01408_gene4_1735_2751	glycosyl hydrolase family 16 [Prevotella buccae ATCC 33574]	√	√	√	√	√	√
Dockerin_1	C02598_gene4_1025_1962	hypothetical protein NatgrDRAFT_2304 [Natronobacterium gregoryi SP2]	NA	NA	NA	X	√	TGA
	C02630_gene2_295_1955	hypothetical protein Calab_3359 [Caldithrix abyssi DSM 13497]	NA	NA	NA	X	X	TAG
GH76	C05280_gene1_1_1221	protein of unknown function DUF255 [Desulfatibacillum alkenivorans AK-01]	NA	NA	NA	√	X	X
GH2_C	C08149_gene1_1_863	Mannan endo-1,4-beta-mannosidase [Medicago truncatula] ┌gb|AES82380.1| Mannan endo-1	NA	NA	NA	√	X	X
GH31	C08283_gene1_1_852	glucosidase protein [Ralstonia solanacearum PSI07] ┌emb|CBJ51675.1| putative glucosidase p	X	X	X	√	X	X
Alpha amylase	C08922_gene1_1_561	Alpha amylase catalytic domain found in Glycosyltrehalose trehalohydrolase (also called Mal	NA	NA	NA	X	X	√
	C08922_gene2_456_791	4-alpha-D-((1->4)-alpha-D-glucano)trehalose trehalohydrolase [Ralstonia sp.	NA	NA	NA	X	√	X
CBM_5_12	C10990_gene1_57_642	putative fusion protein [Agrobacterium tumefaciens 5A]	√	NA	NA	√	√	X
	C16312_gene1_1_381	hypothetical fusion protein [Agrobacterium sp. H13-3] Length=603	NA	NA	NA	√	X	X
	C16327_gene1_1_381	hypothetical protein [Agrobacterium sp. H13-3]	NA	NA	NA	X	X	X

Start ✓: ATG start codon; ✓: elements present; P: Potential; NA: Non Applicable; CT: Contig Terminated.

Table 6. Halophilicity ratio of proteins presented in table 5.

Pfam domain	Orf	Gene annotation	ORF	REF	H.R	H
GH25	C00001_gene27_29329_30348	glycoside hydrolase family 25 [Ochrobactrum intermedium LMG 3301]	3	8	0.37	X
	C00013_gene83_95811_96638	glycoside hydrolase family 25 [Ochrobactrum intermedium LMG 3301]	2	3	0.66	X
	C00052_gene11_9496_10698	glycoside hydrolase family protein [Mesorhizobium opportunistum WSM2075]	14	13	1.07	√
Cellulase	C00016_gene83_77945_79123	cellulase [Yokenella regensburgei ATCC 43003]	16	11	1.45	√
	C00711_gene2_315_1208	Beta-galactosidase [Agrobacterium tumefaciens F2]	3	4	0.75	X
	C01010_gene1_100_996	putative glycoside hydrolase family protein [Treponema azotonutricium ZAS-9]	12	7	1.71	√
GH3	C00024_gene67_52716_55364	Beta-glucosidase [Burkholderia ubonensis Bu]	37	18	2.05	√
	C02816_gene2_152_1711	beta-glucosidase-related glycosidase [Microbacterium testaceum StLB037]	17	29	0.58	X
	C00027_gene9_6518_7543	beta-N-acetylhexosaminidase [Rhizobium etli CIAT 652]	6	4	1.5	√
	C00219_gene7_8447_9466	glycoside hydrolase family 3 domain protein [Mesorhizobium australicum WSM2073]	10	9	1	X
	C00222_gene10_7298_8347	beta-hexosaminidase [Cupriavidus metallidurans CH34]	4	6	0.66	X
	C01616_gene1_1_1241	glycoside hydrolase family 3 domain-containing protein [Chlorobaculum parvum NCIB 8327]	25	21	1.19	√
	C02816_gene1_1_147	glycoside hydrolase [Spirochaeta smaragdinae DSM 11293], thermostable beta-glucosidase	0	1	0	X
	C03408_gene2_1094_1663	beta-D-glucoside glucohydrolase [Cupriavidus metallidurans CH34]	3	3	1	X
GH4	C00076_gene14_14397_15869	glycoside hydrolase family 4 [Rhizobium leguminosarum bv. trifolii WSM1325]	3	2	1.5	X
	C00076_gene19_21986_23359	alpha-galactosidase [Sinorhizobium meliloti CCNWSX0020]	4	9	0.44	X
	C03199_gene1_1_876	malate dehydrogenase [endosymbiont of Tevnia jerichonana (vent Tica)]	14	7	2	√
	C06010_gene2_442_1068	malate dehydrogenase [endosymbiont of Tevnia jerichonana (vent Tica)]	5	3	1.66	X
GH43	C00376_gene3_1662_2519	glycosidase [Thermotoga petrophila RKU-1]	10	7	1.42	X
	C00154_gene6_3285_4316	glycosyl hydrolase family 32 protein [Candidatus Solibacter usitatus Ellin6076]	13	7	1.85	√
	C00338_gene4_1876_3174	glycosyl hydrolase family 32 [Bryantella formatexigens DSM 14469]	17	27	0.59	X
	C00359_gene8_3930_5174	Glycosyl hydrolase family 32, N terminal domain protein [Candidatus Solibacter usitatus Ellin6076]	29	24	1.2	√
	C00992_gene4_1283_2299	glycosidase related protein [Thermaerobacter subterraneus DSM 13965]	12	17	0.7	X
	C01343_gene1_1_1207	Glycosyl hydrolases family 32 [Microcoleus chthonoplastes PCC 7420]	19	23	0.82	X
	C01479_gene5_1406_2763	arabinan endo-1,5-alpha-L-arabinosidase [Thermobaculum terrenum ATCC BAA-798]	18	11	1.63	X
	C01487_gene2_207_1508	glycosyl hydrolase family 32 protein [Candidatus Solibacter usitatus Ellin6076]	56	38	1.47	√
	C02380_gene1_2_1366	Glycosyl hydrolase family 32, N terminal domain protein [Candidatus Solibacter usitatus Ellin6076]	34	22	1.54	√
	C03592_gene2_379_1617	laminin G sub domain 2 [Thiorhodovibrio sp. 970]	17	17	1	√
	C03733_gene1_1_1215	Glycosyl hydrolase family 32 domain protein [Runella slithyformis DSM 19594]	27	21	1.35	√
	C04460_gene3_309_1391	glycosyl hydrolase family 43 protein [Bacteroides sp. D20]	12	14	0.85	X
	C07434_gene1_1_933	glycosidase PH1107-like protein [Meiothermus silvanus DSM 9946]	25	13	1.92	√
	C12554_gene2_223_550	glycosyl hydrolase 32 domain protein [Marinithermus hydrothermalis DSM 14884]	6	5	1.2	√
GH18	C00474_gene2_427_2676	glycosyl hydrolase, family 18 [uncultured marine bacterium 159]	44	38	1.15	√
GH19	C00626_gene4_1836_2477	Lytic enzyme [Erwinia sp. Ejp617]	10	8	1.25	√
GH47	C01198_gene3_1740_2939	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497]	33	24	1.37	√
	C01467_gene3_1054_1416	mannosyl-oligosaccharide alpha-1,2-mannosidase [Ajellomyces dermatitidis ATCC	4	6	0.83	X
	C01467_gene2_506_1030	family 47 glycoside hydrolase [Phytophthora sojae]	11	8	1.37	√
	C02102_gene1_1_959	glycosyl hydrolase [Trichomonas vaginalis G3]	32	22	1.45	√
	C08990_gene3_330_788	endoplasmic reticulum mannosyl-oligosaccharide 1,2-alpha-mannosidase-like[Apis mellifera	10	8	1.25	√
	C14076_gene1_1_438	Mannosyl-oligosaccharide 1,2-alpha-mannosidase [Caulobacter segnis ATCC 21756] Length=4	10	8	1.25	√
	C17322_gene1_1_345	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497]	6	9	0.66	X
	C23658_gene1_1_179	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497] Length=458	0	3	0	X
GH16	C01408_gene4_1735_2751	glycosyl hydrolase family 16 [Prevotella buccae ATCC 33574]	9	8	1.12	X
Dockerin_1	C02598_gene4_1025_1962	hypothetical protein NatgrDRAFT_2304 [Natronobacterium gregoryi SP2]	10	25	0.4	X
	C02630_gene2_295_1955	hypothetical protein Calab_3359 [Caldithrix abyssi DSM 13497]	6	3	2	X
GH76	C05280_gene1_1_1221	protein of unknown function DUF255 [Desulfatibacillum alkenivorans AK-01]	19	8	2.37	X
GH2_C	C08149_gene1_1_863	Mannan endo-1,4-beta-mannosidase [Medicago truncatula] ┌gb|AES82380.1| Mannan endo-1	29	16	1.81	√
GH31	C08283_gene1_1_852	glucosidase protein [Ralstonia solanacearum PSI07] ┌emb|CBJ51675.1| putative glucosidase p	5	1	5	√
Alpha amylase	C08922_gene1_1_561	Alpha amylase catalytic domain found in Glycosyltrehalose trehalohydrolase (also called Malt	1	3	0.33	√
	C08922_gene2_456_791	4-alpha-D-((1->4)-alpha-D-glucano)trehalose trehalohydrolase [Ralstonia sp.	0	0	0	X
CBM_5_12	C10990_gene1_57_642	putative fusion protein [Agrobacterium tumefaciens 5A]	5	10	0.5	X
	C16312_gene1_1_381	hypothetical fusion protein [Agrobacterium sp. H13-3] Length=603	6	7	0.87	X
	C16327_gene1_1_381	hypothetical protein [Agrobacterium sp. H13-3]	5	7	0.71	X

ORF: refer to LCL ORF; REF: best matched (not necessary the first matched);
H.R: halophilicity Ratio = Asp + Glu of ORF/Asp + Glu of REF; H, halophilicity

We selected genes that show high potential for expression in the heterologous system for further analysis (Table 7). All 14 selected genes show ORFs with full length, -10 and -35 transcriptional elements, RBS, and start and stop codons for protein translation. Eleven of the 14 selected proteins have classical signal peptide for secretion, two were found to be secreted by non-classical pathway, and one is not secreted.

Table 7. Full length ORFs

Pfam	Contig	RBS	-10	-35	SP	Start	Stop	D+E (O)	D+E (R)	O/R Ratio	H	Annotation
GH25	00001_gene27_29329_30348	√	√	√	√	√	√	3	8	0.37	X	glycoside hydrolase family 25 [Ochrobactrum intermedium LMG 3301]
Cellulase	00016_gene83_77945_79123	√	√	√	√	√	√	16	11	1.45	√	cellulase [Yokenella regensburgei ATCC 43003]
	01010_gene1_100_996	√	√	√	√	√	√	12	7	1.71	√	putative glycoside hydrolase family protein [Treponema azotonutricium ZAS-9]
GH3	00219_gene7_8447_9466	√	√	√	√	√	√	10	9	1.11	X	glycoside hydrolase family 3 domain protein [Mesorhizobium australicum WSM2073]
	00222_gene10_7298_8347	√	√	√	√	√	√	4	6	0.66	X	beta-hexosaminidase [Cupriavidus metallidurans CH34]
	00024_gene67_52716_55364	√	√	√	X	TTG*	√	37	18	2.05	√	Beta-glucosidase [Burkholderia ubonensis Bu]
GH4	00076_gene14_14397_15869	√	√	√	√	√	√	3	2	1.5	√	glycoside hydrolase family 4 [Rhizobium leguminosarum bv. trifolii WSM1325]
	00076_gene19_21986_23359	√	√	√	√	√	√	4	9	0.44	X	alpha-galactosidase [Sinorhizobium meliloti CCNWSX0020]
GH16	01408_gene4_1735_2751	√	√	√	√	√	√	9	8	1.12	X	glycosyl hydrolase family 16 [Prevotella buccae ATCC 33574]
GH18	00474_gene2_427_2676	√	√	√	√	√	√	44	38	1.15	X	glycosyl hydrolase, family 18 [uncultured marine bacterium 159]
GH19	00626_gene4_1836_2477	√	√	√	√	√	√	10	8	1.25	√	Lytic enzyme [Erwinia sp. Ejp617]
GH47	01467_gene3_1054_1416	√	√	√	X	√	√	4	6	0.66	X	mannosyl-oligosaccharide alpha-1,2-mannosidase [Ajellomyces dermatitidis ATCC
	01467_gene2_506_1030	√	√	√	X	√	√	11	8	1.37	√	family 47 glycoside hydrolase [Phytophthora sojae]
	01198_gene3_1740_2939	√	√	√	√	GTG*	√	33	24	1.37	√	glycoside hydrolase family 47 [Caldithrix abyssi DSM 13497]

H, halophilicity; D + E (O)= Asp + Glu of ORF, D + E (R)= Asp + Glu of reference gene
SP, Signal peptide for secretion; *start codons other than ATG

Detail schematic presentaion of the structure and organization of the selected genes are shown in Figure 7.

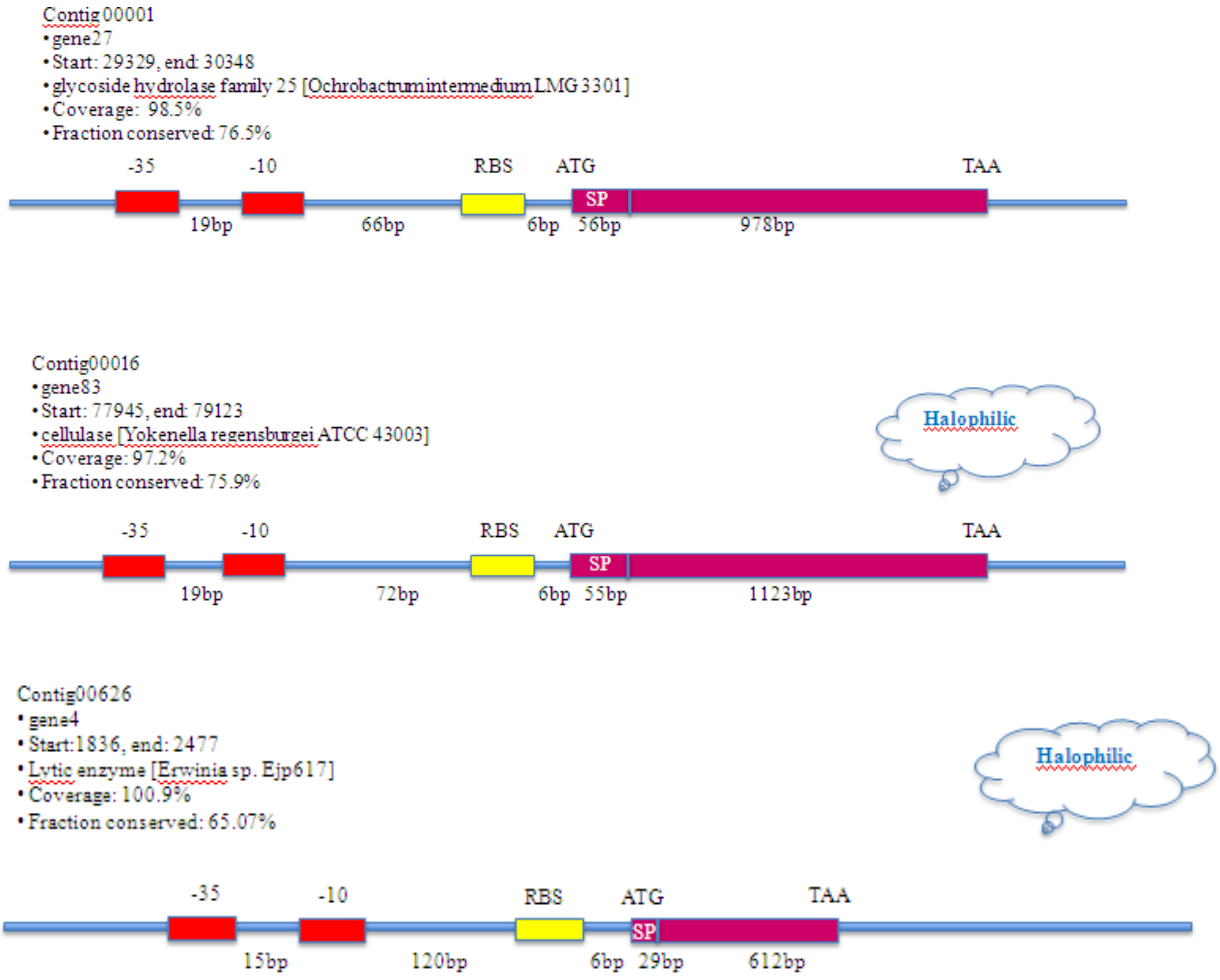

Figure 7. Schematic presentaion of the structure and organization of the selected genes that have signal peptide for secretion (SP).

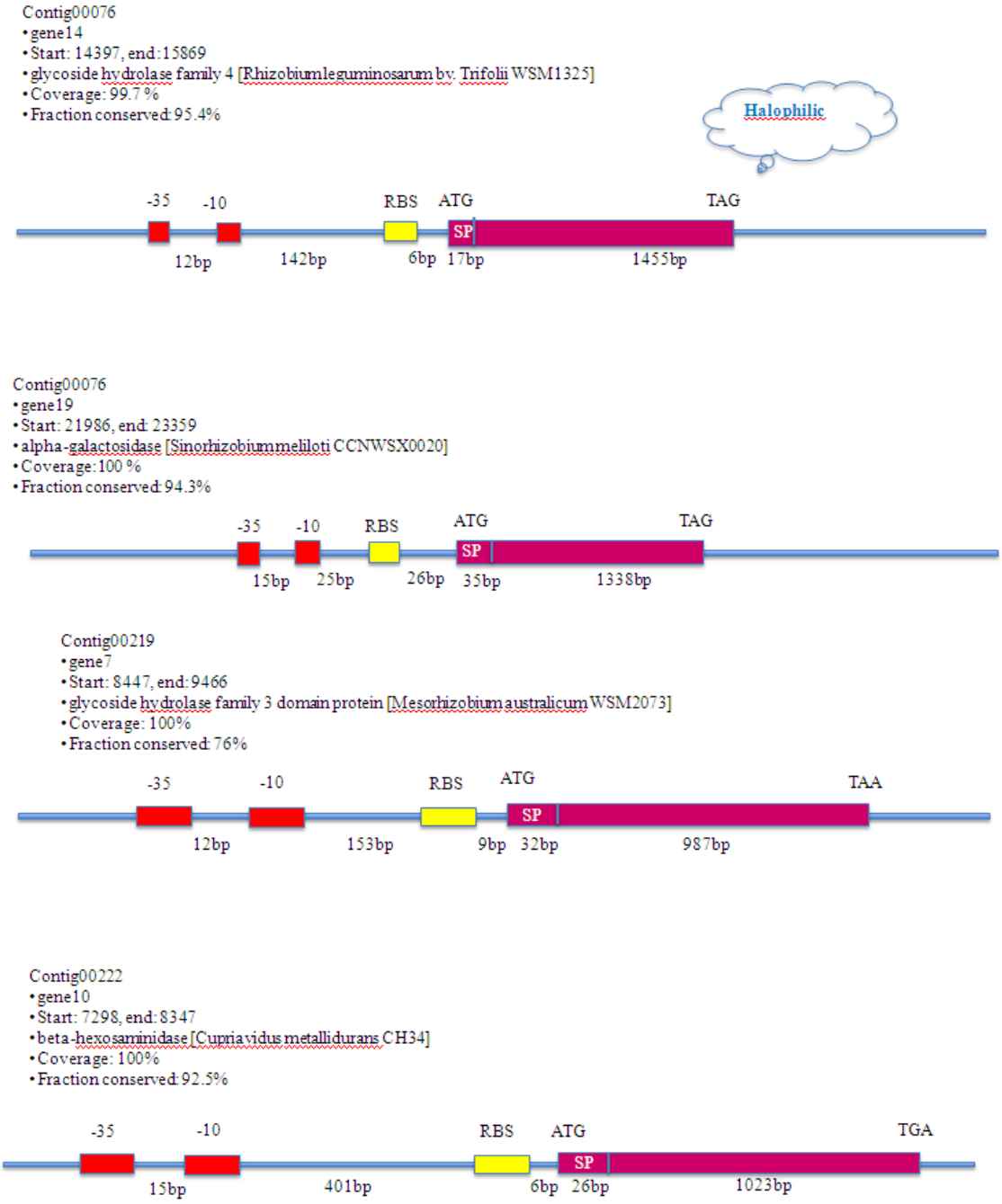

Figure 7. (continue) Schematic presentaion of the structure and organization of the selected genes that have signal peptide for secretion (SP).

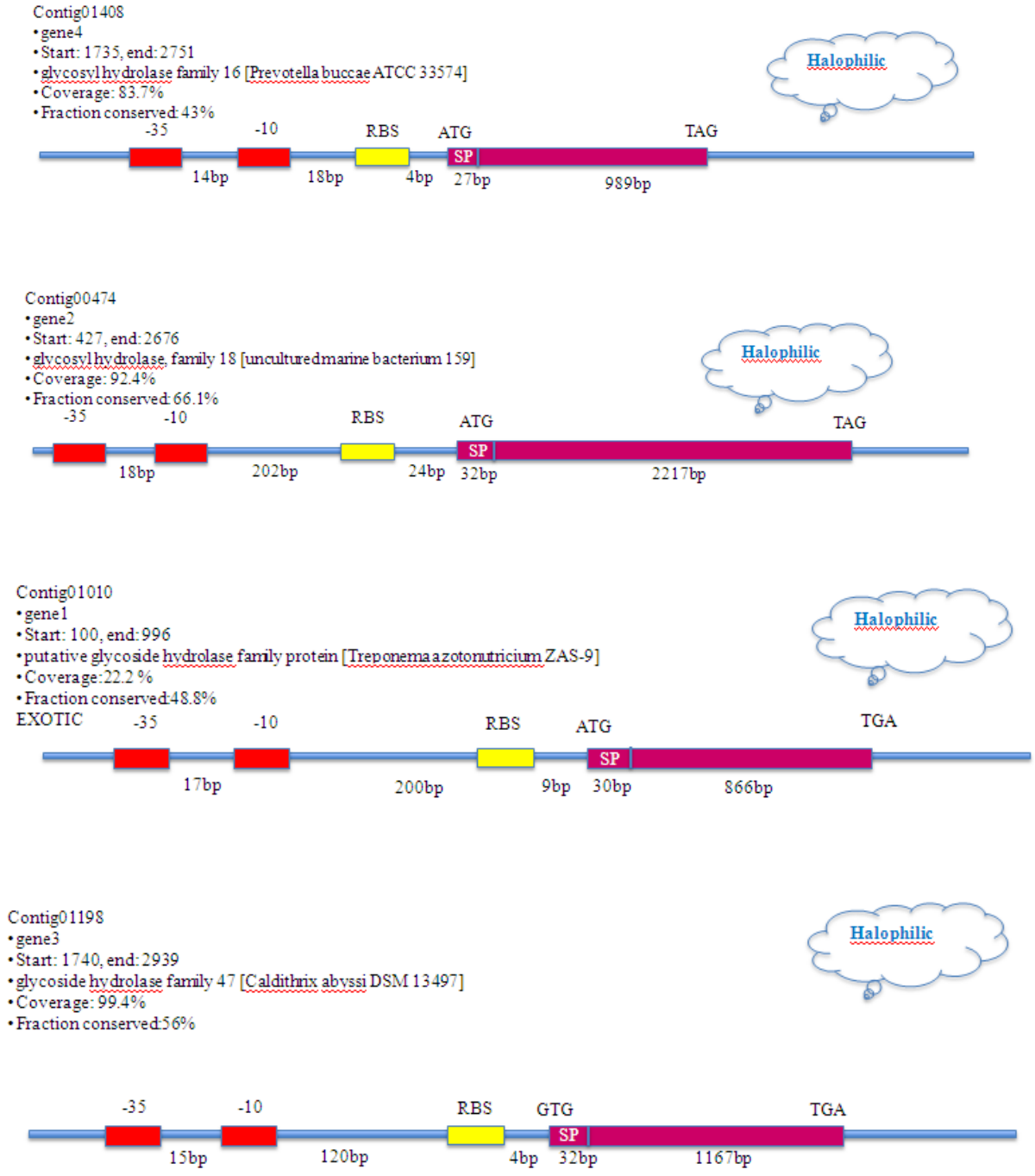

Figure 7. (continue) Schematic presentaion of the structure and organization of the selected genes that have signal peptide for secretion (SP).

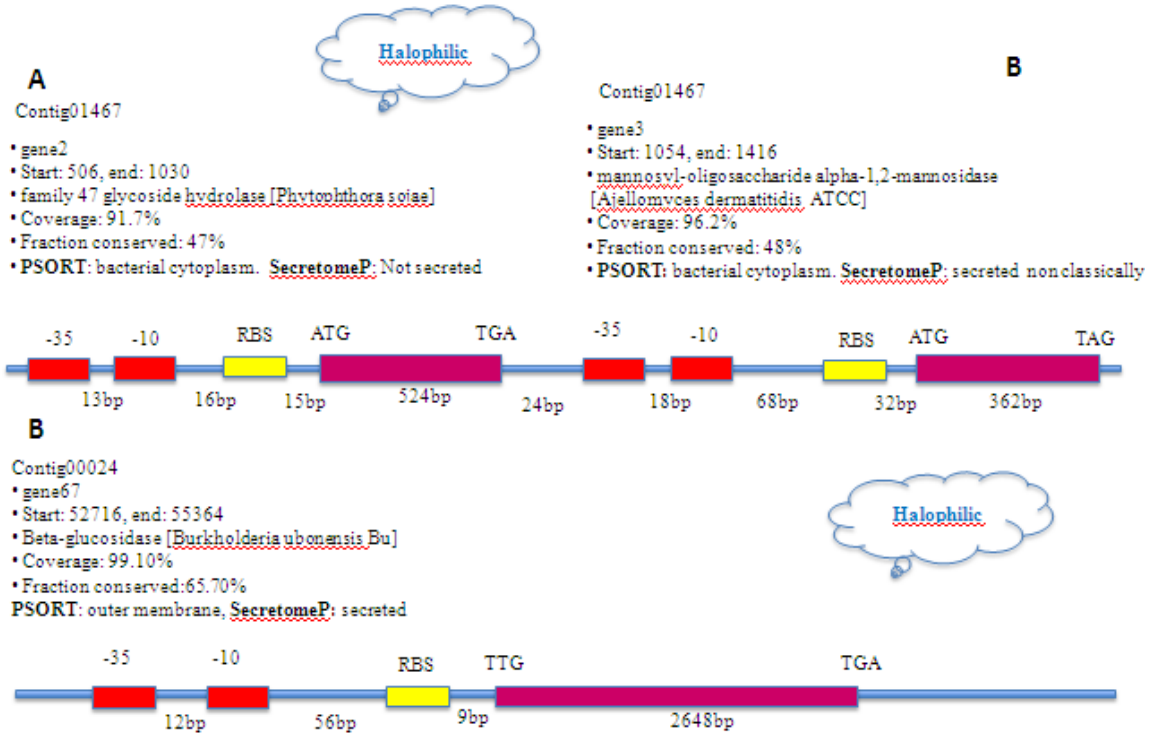

Figure 7. (continue) Schematic presentaion of the structure and organization of gene that not secreted (A) and those that are secretad throgh non-clasical pathway (B).

ORFs having full lengths but with no signal peptides - gene2 (alpha mannosidase) and gene3 (alpha mannosidase) located on contig01467, and gene67 (beta-glucosidase) located on contig00024 - were submitted to PSORT and SecretomeP to check for potential usage of non-classical secretory pathway. For gene2 located on contig01467, PSORT predicted that the protein was cytoplasmic. SecretomeP predicted that the protein is not secreted in agreement with the results given by signalP and PSORT.

Regarding gene3 located on contig01467, PSORT predicted that the protein was cytoplasmic as well, while the result of SecretomeP contradicted that by predicting that the protein will be secreted non-classically.

In the case of gene67 located on contig00024, it was predicted to be secreted to the outer membrane by both PSORT and secretomeP, which indicated that most probably it would be secreted non-classically.

For further analyses, and to indicate whether these proteins contain trans-membrane domains, we used THMM online software. Figure 17 shows the results that demonstrate that no trans-membrane domains existed in any of the three proteins. We used the sequence of Trhxt1; *T. reesi* putative hexose (glucose) transporter protein (Ramos, et al., 2006) which has trans-membrane domains, as a positive control.

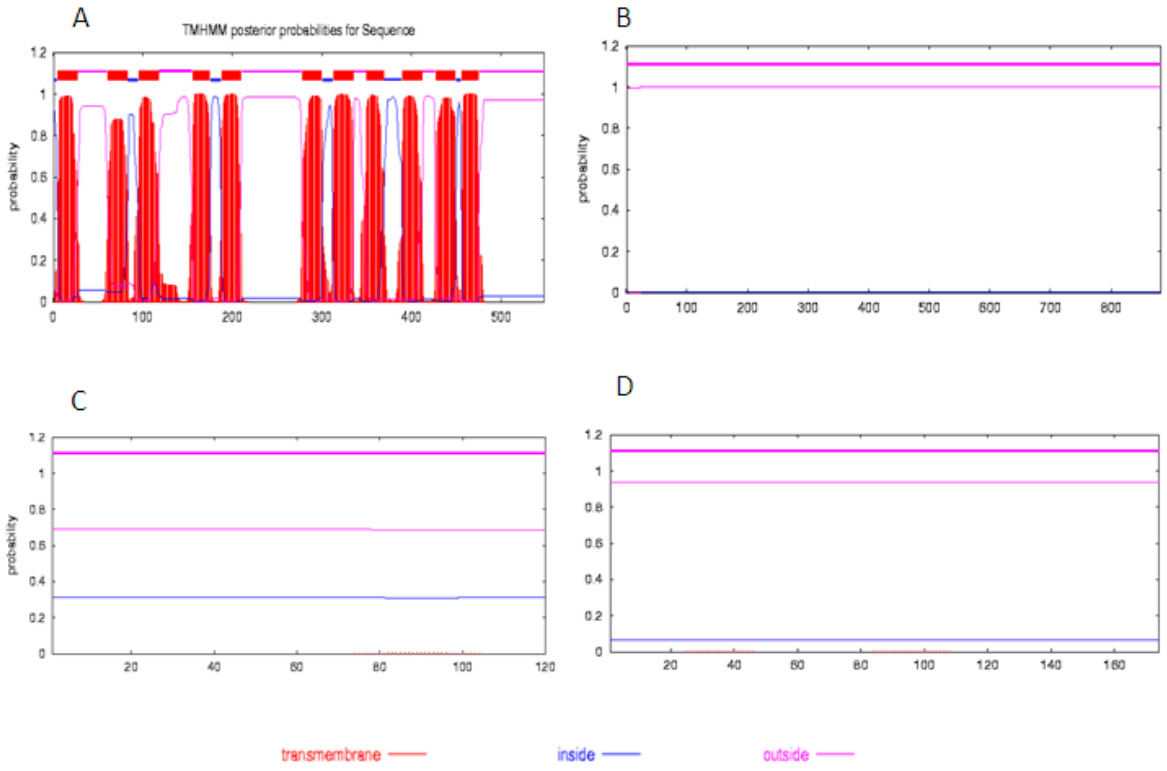

Figure 8. Results of TMHMM of positive control,Trhxt1 protein (A), known to have transmembrane domains, (B) gene67 located on contig00024 (C) gene2 located on contig01467 and (D) gene3 located on contig 01467.

From the 14 examined ORFs we selected four candidates that may have relevant features to understand molecular adaptation of enzymes to extreme halophilic environment. All of the proteins have signal peptides, and therefore they are expected to be exposed to high concentration of salt. Two of them are potential cellulases that may have implication in fuel production from cellulosic biomass. The other two enzymes; a chitinase and an alpha galactosidase.

We describe below the details of the structural features of one of the secreted enzyme located on contig 16 (Contig00016_gene83_77945_79123).

- The ORF on this contig starts from nucleotide number 77945 and ends at 79123 as shown in figure 7.
- RBS: it was predicted manually, where a GA rich region is present. I located 5 bases upstream of the start codon (ATG in this case): GAAGGAG.
- Promoter region (-10 & -35): 300 bases upstream of the translation start codon were identified by Bprom software. This software identified the potential promoter elements located at -10 and -35 from the transcriptional start site.
- The presence of a signal peptide was checked by SignalP (figure 8). The score obtained with SignalP software was found to be 0.856 for this protein, which is higher than the cutoff score of 0.5. The signal peptidase cleavage site was predicted to be between the 19^{th} and the 20^{th} amino acids, therefore the mature protein will start at the 20^{th} amino acid.
- Halophilicity ratio calculated for this protein was 1.45, where aspartic and glutamic acids were 16 for the ORF (substituted by other residues in the reference gene) compared to 11 for the reference sequence (not found in the ORF).

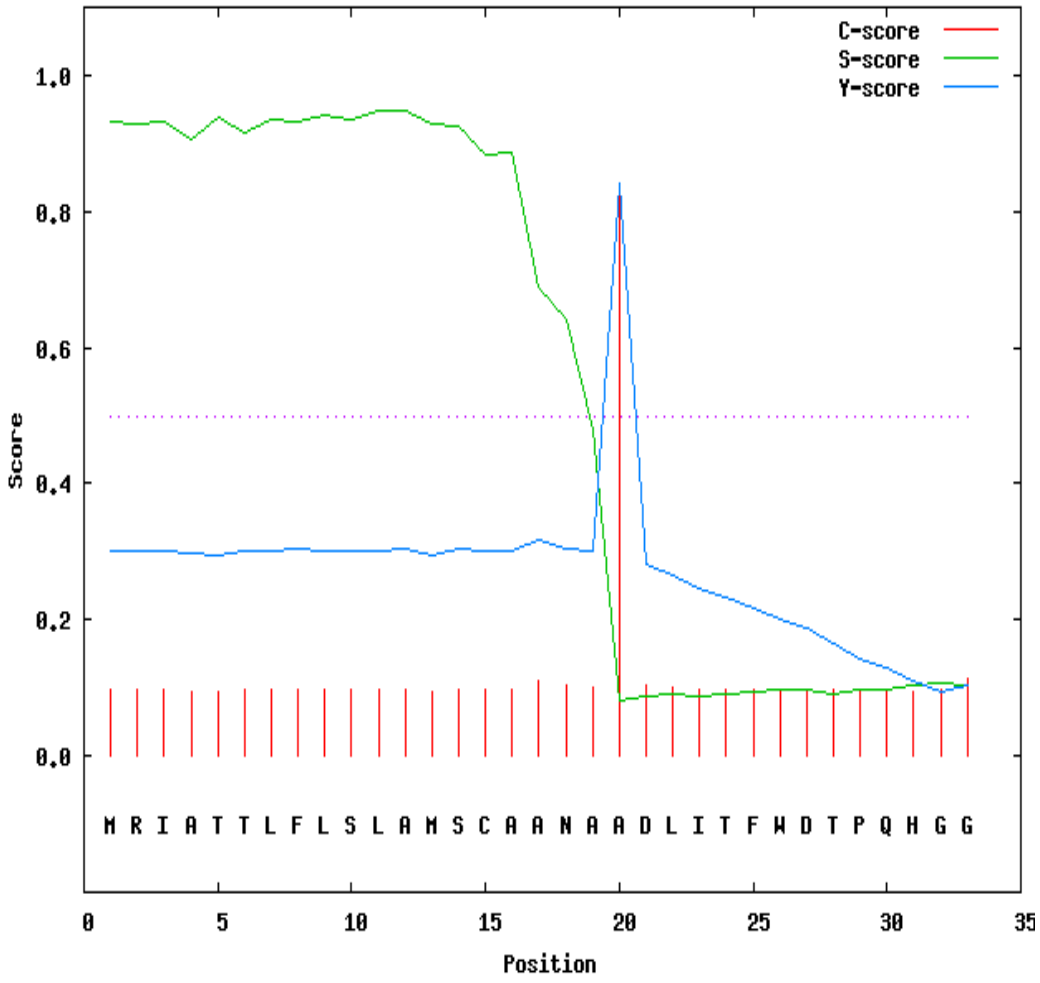

Figure 9. SignalP result of the potential secreted cellulase (gene 83. Contig 16) and signal peptide cleavage site. The C-score indicates the position of the signal peptidase cleavage site (red line). The S-score indicates the prediction for each amino acid to be secreted (green line). The Y-score is a derivative of C-score and S-score to give a better prediction of the cleavage site.

The four proteins have high halophilic ratio when compared with non-halophilic orthologue. To further characterize the four proteins, we modeled their 3D structures using the SWISS-MODEL online server and visualized the predicted structures using RasMol.

The relevant acidic amino acid residues that may contribute to the stability of the proteins in a high salinity environment were located on the surface of the molecules (Figure 10, 11, 12 and 13).

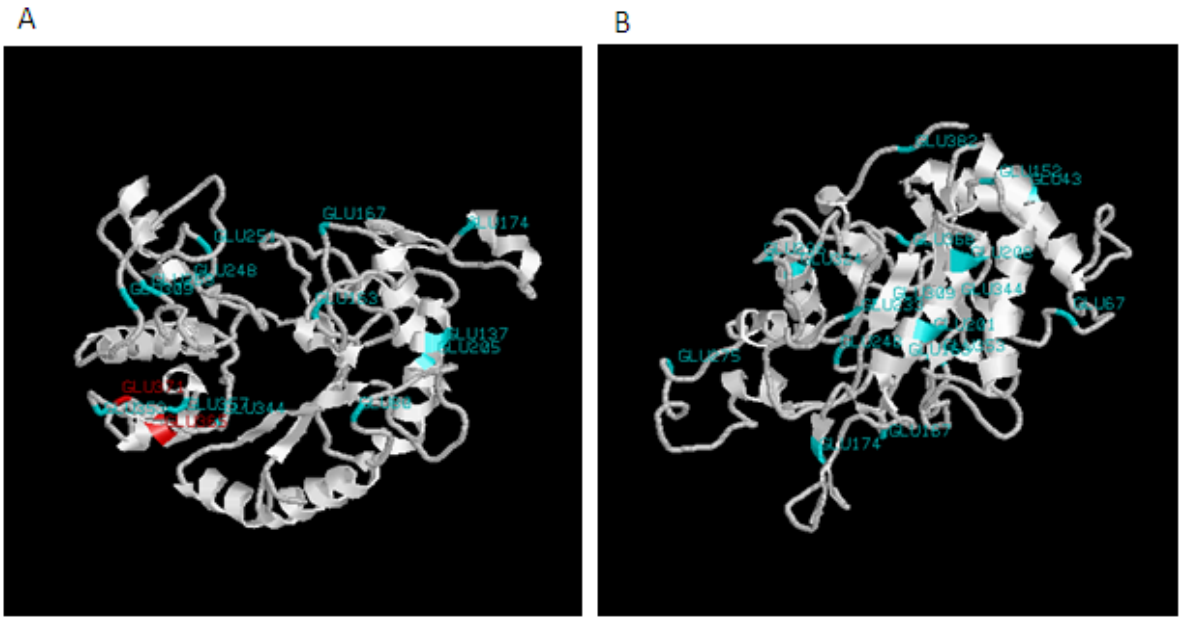

Figure 10. 3D predicted model of gene 83 located on contig 16 (A) and its best reference hit (B). The glutamic acid residue shown in red is not present in the reference protein. The template used to predict the 3D structure for the ORF is *Acidothermus cellulolyticus* Endocellulase E1 catalytic domain, while for the reference the template used is crystal structure of the cellulase enzym from *F. nodosum* RT17-B1

The BLASTx coverage of gene number 83 located on contig 16 was found to be 97.2% while the conserved fraction is 75.9% which suggests that it is well distributed to cover the reference sequence but may have some novel fragments.

Same analyses were followed for all the other ORFs. The second protein that was selected for further characterization was gene14 which was on contig00076 having an annotation: glycoside hydrolase family 4 [Rhizobium leguminosarum bv. trifolii WSM1325]. As mentioned in table 3, GH4 family encloses α-glucosidases, α-galactosidases, α-glucuronidases activities which are

interesting in the process of cell wall degradation. Given that it has a hit coverage of 99.7% and a conserved fraction of 95.4%, this ORF is promising to give us an enzyme that is highly conserved and most probably will be functionally active. In addition, the halophilic residue is shown in red (figure11).

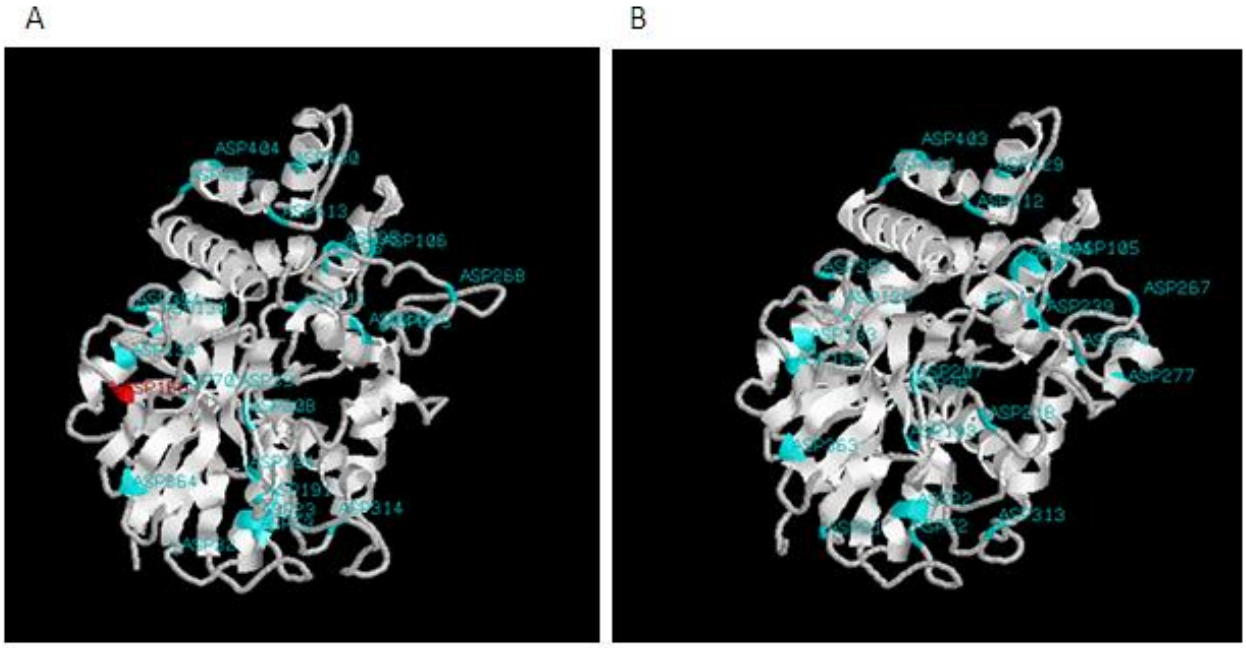

Figure 11. 3D models of contig00076 showing the aspartic acid halophilic residues in the ORF (A) and its reference (B). The aspartic acid colored in red in the ORF is the residue believed to contribute to the halophilicity of the ORF. The template used is crystal structure of phospho-beta-glucosidase

Similarly, gene4 located on contig 626, annotated by BLASTx as lytic enzyme [Erwinia sp. Ejp617], has a hit with 100% coverage, and a conserved fraction of 65%. It had a halophilicity ratio of 1.25. This is shown in figure 11, where the halophilic residues are highlighted in red.

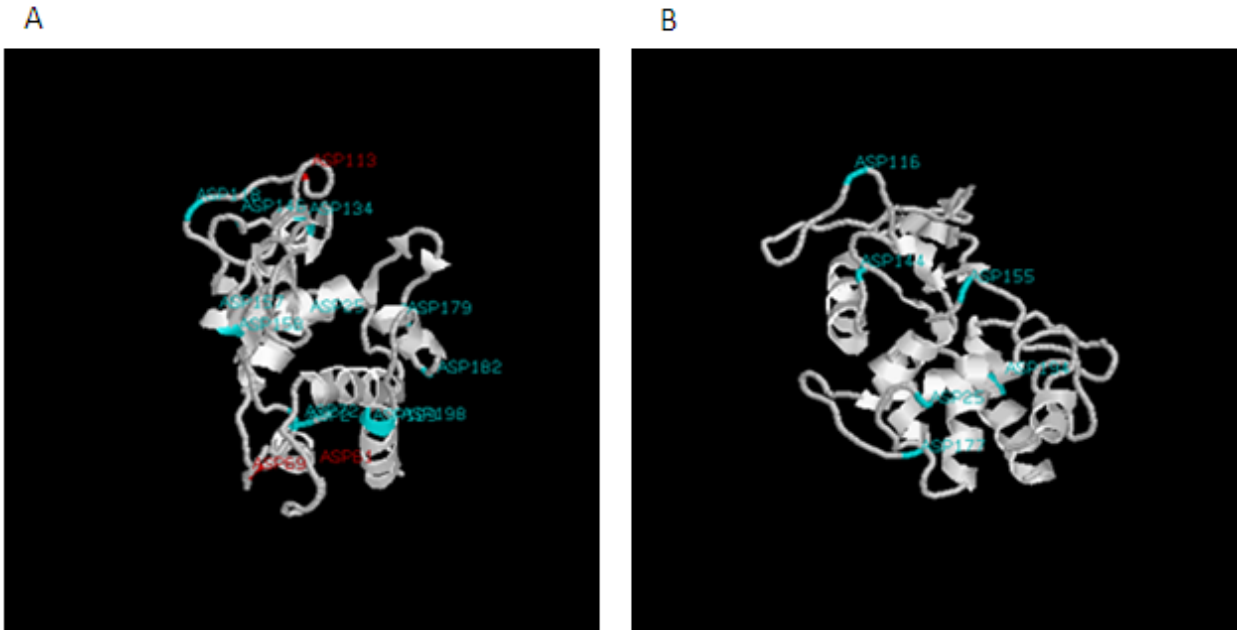

Figure 12. Halophilic residues (aspartic acid in this case) shown on 3D model for contig00626_gene4 in comparison with its reference gene. The different aspartic acid residues in the ORF contributing to its halophilicity are shown in red color. The template was structure of chitinase from Jack bean for the ORF, while the template for the reference is crystal structure of class I chitinase from Oryza sative L Japonicum.

Gene1, located on contig 01010, has a hit coverage of 22.2% while the conserved fraction is 48.8%, which made it far away from being conserved and suggests that it could be a novel cellulase enzyme. It had a signal peptide and a high halophilicity ratio of 1.7. Figure 12 shows the residues that contribute to its halophilicity. All these characteristics together with having a cellulase domain make it a very attractive protein for further characterization.

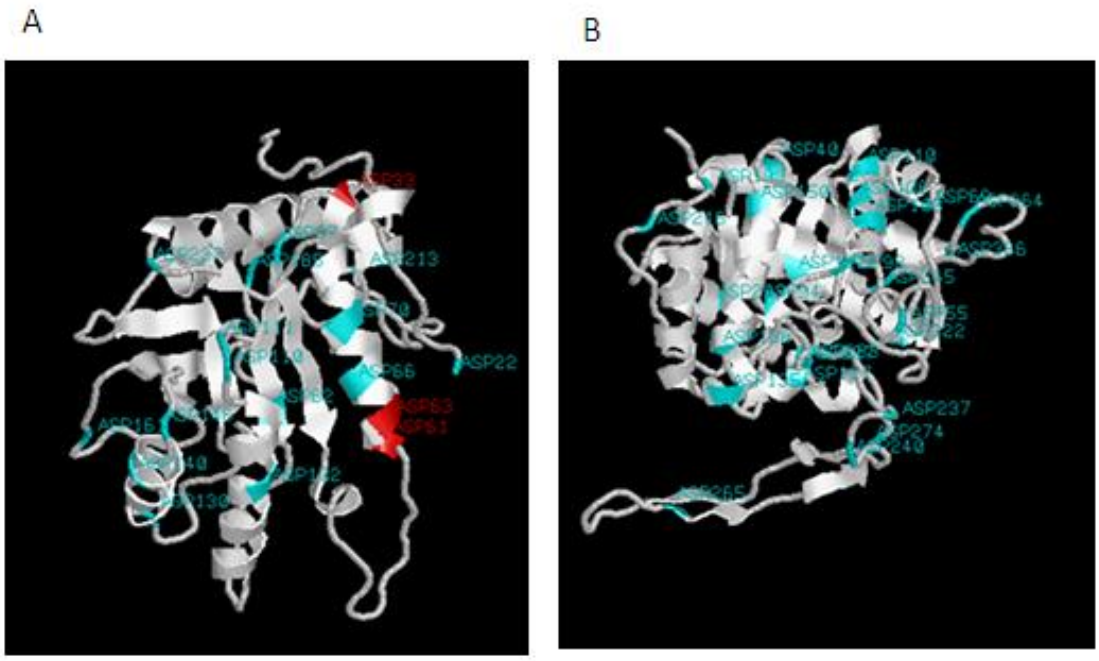

Figure 13. 3D model of contig01010_gene1 and its reference sequence highlighting the different halophilic residue of the ORF in red. The template for the ORF was hyperthermophilic endocellulase from Pyrococcus horikoshii , while for the reference it is crystal structure of *Thermotoga maritima* Cel5A

Based on the structural features of the four proteins, they should have unique properties regarding stability in high saline solution.

In conclusion, this work have established a dataset of genes with potential glycosyl hydrolases, including biomass and cell wall degrading enzymes, that contain activities present in the microbial community of the ATII-LCL environment. In addition, we selected the most promising candidates for further molecular and catalytic characterization.

Future work

Expression of some of the suggested enzymes that could be of industrial importance and characterizing their enzymatic activities. Moreover, we intend to examine the proteins containing only Pfam domains but without any BLASTx similarity result, or with only a low conserved fraction as they could be novel enzymes that have not yet been identified and deposited in the NCBI database.

References

Ahn, Y. B., Rhee, S. K., Fennell, D. E., Kerkhof, L. J., Hentschel, U., & Haggblom, M. M. (2003). Reductive dehalogenation of brominated phenolic compounds by microorganisms associated with the marine sponge Aplysina aerophoba. *Appl Environ Microbiol, 69*(7), 4159-4166.

André Antunes, David Kamanda Ngugi, & Stingl, U. (2011). Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. *Environmental Microbiology Reports 3*(4), 416-433.

Atlantis II Deep. (2011).

Bayer, E. A., Chanzy, H., Lamed, R., & Shoham, Y. (1998). Cellulose, cellulases and cellulosomes. *Curr Opin Struct Biol, 8*(5), 548-557.

Beja, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., et al. (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science, 289*(5486), 1902-1906.

Bonaccorsi, E. D., Ferreira, A. J., Chambergo, F. S., Ramos, A. S., Mantovani, M. C., Farah, J. P., et al. (2006). Transcriptional response of the obligatory aerobe Trichoderma reesei to hypoxia and transient anoxia: implications for energy production and survival in the absence of oxygen. *Biochemistry, 45*(12), 3912-3924.

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

Cao, Y., & Tan, H. (2002). Effects of cellulase on the modification of cellulose. *Carbohydr Res, 337*(14), 1291-1296.

Carbohydrate active enzymes (2012). http://www.cazy.org/GH13.html

Case, R. J., Boucher, Y., Dahllof, I., Holmstrom, C., Doolittle, W. F., & Kjelleberg, S. (2007). Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol, 73*(1), 278-288.

Chu, X., He, H., Guo, C., & Sun, B. (2008). Identification of two novel esterases from a marine metagenomic library derived from South China Sea. *Appl Microbiol Biotechnol, 80*(4), 615-625.

Cochran, J. R. (Ed.) (2008) AccessScience. McGraw-Hill Companies.

Davies, G., & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure, 3*(9), 853-859.

Ding, S. Y., Rincon, M. T., Lamed, R., Martin, J. C., McCrae, S. I., Aurilia, V., et al. (2001). Cellulosomal scaffoldin-like proteins from Ruminococcus flavefaciens. *J Bacteriol, 183*(6), 1945-1953.

Dinsdale, E. A., Edwards, R. A., Hall, D., Angly, F., Breitbart, M., Brulc, J. M., et al. (2008). Functional metagenomic profiling of nine biomes. *Nature, 452*(7187), 629-632.

Divne, C., Stahlberg, J., Reinikainen, T., Ruohonen, L., Pettersson, G., Knowles, J. K., et al. (1994). The three-dimensional crystal structure of the catalytic core of cellobiohydrolase I from Trichoderma reesei. *Science, 265*(5171), 524-528.

Ellegren, H. (2008). Sequencing goes 454 and takes large-scale genomics into the wild. *Mol Ecol, 17*(7), 1629-1631.

Faber, E., Botz, R., Poggenburg, J., Schmidt, M., Stoffers, P., & Hartmann, M. (1998). Methane in Red Sea brines. *Organic Geochemistry, 29*, 363-379.

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science, 281*(5374), 237-240.

Fontes, C. M., & Gilbert, H. J. (2010). Cellulosomes: highly efficient nanomachines designed to deconstruct plant cell wall complex carbohydrates. *Annu Rev Biochem, 79*, 655-681.

Gallardo, G. L., Butler, M., Gallo, M. L., Rodriguez, M. A., Eberlin, M. N., & Cabrera, G. M. (2006). Antimicrobial metabolites produced by an intertidal Acremonium furcatum. *Phytochemistry, 67*(21), 2403-2410.

Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., et al. (2002). Isolation of antibiotics turbomycin a and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol, 68*(9), 4301-4306.

Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., & Field, K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. *Nature, 345*(6270), 60-63.

Henrissat, B., & Bairoch, A. (1996). Updating the sequence-based classification of glycosyl hydrolases. *Biochem J, 316 (Pt 2)*, 695-696.

Iwanicka-Nowicka, R., Zielak, A., Cook, A. M., Thomas, M. S., & Hryniewicz, M. M. (2007). Regulation of sulfur assimilation pathways in Burkholderia cenocepacia: identification of transcription factors CysB and SsuR and their role in control of target genes. *J Bacteriol, 189*(5), 1675-1688.

Jeng, W. Y., Wang, N. C., Lin, M. H., Lin, C. T., Liaw, Y. C., Chang, W. J., et al. (2010). Structural and functional analysis of three beta-glucosidases from bacterium Clostridium cellulovorans, fungus Trichoderma reesei and termite Neotermes koshunensis. *J Struct Biol, 173*(1), 46-56.

Jo Handelsman, Mark Liles, David Mann, Christian Riesenfeld, & Goodman, R. M. (2002). Cloning the metagenome: Culture-independent access to the diversity and functions of the uncultivated microbial world. *Methods in Microbiology, 33*, 241–255.

Kennedy, J., Flemer, B., Jackson, S. A., Lejon, D. P., Morrissey, J. P., O'Gara, F., et al. (2010). Marine metagenomics: new tools for the study and exploitation of marine microbial metabolism. *Mar Drugs, 8*(3), 608-628.

Kennedy, J., Marchesi, J. R., & Dobson, A. D. (2008). Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb Cell Fact, 7*, 27.

Konstantinidis, K. T., Braff, J., Karl, D. M., & DeLong, E. F. (2009). Comparative metagenomic analysis of a microbial community residing at a depth of 4,000 meters at station ALOHA in the North Pacific subtropical gyre. *Appl Environ Microbiol, 75*(16), 5345-5355.

LeCleir, G. R., Buchan, A., Maurer, J., Moran, M. A., & Hollibaugh, J. T. (2007). Comparison of chitinolytic enzymes from an alkaline, hypersaline lake and an estuary. *Environ Microbiol, 9*(1), 197-205.

Li, L. L., McCorkle, S. R., Monchy, S., Taghavi, S., & van der Lelie, D. (2009). Bioprospecting metagenomes: glycosyl hydrolases for converting biomass. *Biotechnol Biofuels, 2*, 10.

Lynd, L. R., van Zyl, W. H., McBride, J. E., & Laser, M. (2005). Consolidated bioprocessing of cellulosic biomass: an update. *Curr Opin Biotechnol, 16*(5), 577-583.

Lynd, L. R., Weimer, P. J., van Zyl, W. H., & Pretorius, I. S. (2002). Microbial cellulose utilization: fundamentals and biotechnology. *Microbiol Mol Biol Rev, 66*(3), 506-577, table of contents.

Martin-Cuadrado, A. B., Lopez-Garcia, P., Alba, J. C., Moreira, D., Monticelli, L., Strittmatter, A., et al. (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One, 2*(9), e914.

Martinez, A., Kolvek, S. J., Yip, C. L., Hopke, J., Brown, K. A., MacNeil, I. A., et al. (2004). Genetically modified bacterial strains and novel bacterial artificial chromosome shuttle vectors for constructing environmental libraries and detecting heterologous natural products in multiple expression hosts. *Appl Environ Microbiol, 70*(4), 2452-2463.

Mba Medie, F., Davies, G. J., Drancourt, M., & Henrissat, B. (2012). Genome analyses highlight the different biological roles of cellulases. *Nat Rev Microbiol, 10*(3), 227-234.

Mishra, R. N., Singla-Pareek, S. L., Nair, S., Sopory, S. K., & Reddy, M. K. (2002). Directional genome walking using PCR. *Biotechniques, 33*(4), 830-832, 834.

Mohnen, D. (2008). Pectin structure and biosynthesis. *Curr Opin Plant Biol, 11*(3), 266-277.

Newman, D. K., & Banfield, J. F. (2002). Geomicrobiology: how molecular-scale interactions underpin biogeochemical systems. *Science, 296*(5570), 1071-1077.

Pace, N., Stahl, D., Lane, D., & Olsen, G. (1985). Analyzing natural microbial populations by rRNA sequences. ***ASM American Society for Microbiology News***
51(1), 4-12.

Quaiser, A., Zivanovic, Y., Moreira, D., & Lopez-Garcia, P. (2011). Comparative metagenomics of bathypelagic plankton and bottom sediment from the Sea of Marmara. *ISME J, 5*(2), 285-304.

Ramos, A. S., Chambergo, F. S., Bonaccorsi, E. D., Ferreira, A. J., Cella, N., Gombert, A. K., et al. (2006). Oxygen- and glucose-dependent expression of Trhxt1, a putative glucose transporter gene of Trichoderma reesei. *Biochemistry, 45*(26), 8184-8192.

Reczey, K., Brumbauer, A., Bollok, M., Szengyel, Z., & Zacchi, G. (1998). Use of hemicellulose hydrolysate for beta-glucosidase fermentation. *Appl Biochem Biotechnol, 70-72*, 225-235.

Reinhold-Hurek, B., Hurek, T., Claeyssens, M., & van Montagu, M. (1993). Cloning, expression in Escherichia coli, and characterization of cellulolytic enzymes of Azoarcus sp., a root-invading diazotroph. *J Bacteriol, 175*(21), 7056-7065.

Riesenfeld CS, Schloss PD, & J, H. (2004). Metagenomics: Genomic Analysis of microbial communities. *Annual Reviews of Genetics 38*, 525-552.

Ross, D. A. (1972). Red sea hot brine area: revisited. *Science, 175*(4029), 1455-1457.

Schloss, P. D., & Handelsman, J. (2003). Biotechnological prospects from metagenomics. *Curr Opin Biotechnol, 14*(3), 303-310.

Schmeisser, C., Steele, H., & Streit, W. R. (2007). Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol, 75*(5), 955-962.

Schwarz, W. H. (2001). The cellulosome and cellulose degradation by anaerobic bacteria. *Appl Microbiol Biotechnol, 56*(5-6), 634-649.

Sebat, J. L., Colwell, F. S., & Crawford, R. L. (2003). Metagenomic profiling: microarray analysis of an environmental genomic library. *Appl Environ Microbiol, 69*(8), 4927-4934.

Sharma, P., Kumari, H., Kumar, M., Verma, M., Kumari, K., Malhotra, S., et al. (2008). From bacterial genomics to metagenomics: concept, tools and
recent advances. *Indian Journal of Microbiology*
48(2), 173-194.

Shashar, N., **Cohen, Y.**, & **Loya, Y.** (1993). Extreme Diel Fluctuations of Oxygen in Diffusive Boundary Layers Surrounding Stony Corals. *The Biological Bulletin, 185*(3), 455-461.

Shoham, Y., Lamed, R., & Bayer, E. A. (1999). The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends Microbiol, 7*(7), 275-281.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., & DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J Bacteriol, 178*(3), 591-599.

Streit, W. R., & Schmitz, R. A. (2004). Metagenomics--the key to the uncultured microbes. *Curr Opin Microbiol, 7*(5), 492-498.

Teather, R. M., & Wood, P. J. (1982). Use of Congo red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from the bovine rumen. *Appl Environ Microbiol, 43*(4), 777-780.

Teeri, T. T., Koivula, A., Linder, M., Wohlfahrt, G., Divne, C., & Jones, T. A. (1998). Trichoderma reesei cellobiohydrolases: why so efficient on crystalline cellulose? *Biochem Soc Trans, 26*(2), 173-178.

Teske, A., & Sorensen, K. B. (2008). Uncultured archaea in deep marine subsurface sediments: have we caught them all? *ISME J, 2*(1), 3-18.

Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science, 304*(5667), 66-74.

von Bubnoff, A. (2008). Next-generation sequencing: the race is on. *Cell, 132*(5), 721-723.

Wang, G. Y., Graziani, E., Waters, B., Pan, W., Li, X., McDermott, J., et al. (2000). Novel natural products from soil DNA libraries in a streptomycete host. *Org Lett, 2*(16), 2401-2404.

Wang, Y., & Qian, P. Y. (2009). Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS One, 4*(10), e7401.

Wang, Z. (2010). Enzymatic saccharification of lignocellulose. *Science Topics*.

William B.F. Ryan , & Schreiber, B. C. (Eds.). (2012) *Encyclopædia Britannica*.

WINCKLER, G., KIPFER, R., AESCHBACH–HERTIG, W., BOTZ, R., SCHMIDT, M., SCHULER, S., et al. (2000). **Sub sea floor boiling of Red Sea Brines: New indication from noble gas data**. *Geochim. Cosmochim. Acta, 64*, 1567-1575.

Wolfgang, R. S., & Ruth, A. S. (2004). Metagenomics - the key to the uncultured microbes. *Current Opinion in Microbiology 7*, 492-498.

Wommack, K. E., Bhavsar, J., & Ravel, J. (2008). Metagenomics: read length matters. *Appl Environ Microbiol, 74*(5), 1453-1463.

Xu, J. (2006). Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. *Mol Ecol, 15*(7), 1713-1731.

Yarbrough, J. M., Himmel, M. E., & Ding, S. Y. (2009). Plant cell wall characterization using scanning probe microscopy techniques. *Biotechnol Biofuels, 2*, 17.