### 4.1.1. A description of the main S&T results/foregrounds (not exceeding 25 pages)

#### 4.1.3.1. Environmental sampling of hot spring biodiversity

More than three hundred samples were obtained from natural thermal environments located worldwide: terrestrial hot springs of Iceland, Italy, China, Yellowstone National Park (USA), Kamchatka, Kuril Islands and Baikal Lake area (Russia), and deep subsurface biosphere (Western Siberia, Norwegian Sea, Troll, Barents Sea, Spitzbergen). The temperature of samples was ranging from 40 to 151$^{o}$C, while pH varied from 2.0 to 10.5. Samples were represented by water, sediments and microbial biofilms and were further used for DNA isolation and for enrichment and isolation of thermophilic microorganisms with hydrolytic activities.

#### 4.1.3.2. In situ enrichment of thermophilic microorganisms with hydrolytic activities

*In situ* enrichments with diverse biopolymeric substrates of interest were set in the hot springs of Kamchatka, Kurils and Island. Substrates were cellulose (MCC, CMC, leaves of corn and bamboo), xylan, starch, alfa- and beta-keratins, xanthan gum, polyester, PVA. The composition of microbial communities developing in primary *in situ* enrichments was studied by PCR-DGGE of 16S rRNA genes. Those showing visible degradation of insoluble substrates and/or containing new phylogenetic groups of thermophilic microorganisms were used for further characterization and isolation work.

#### 4.1.3.3. High-throughput enrichment and isolation of new thermophilic microorganisms with hydrolytic activities.

Hundreds of crude environmental samples and *in situ* enrichments were cultured in the lab with media containing different polymeric substrates of industrial interest. The substrates included cellulose, xylan, lignin, starch, chitin, bamboo leaves, alfa- and beta-keratins, xanthan gum, lichenan, agarose, polyester, and PVA. Enrichment conditions are either aerobic or anaerobic, without acceptor, or with ferric iron, sulfur, sulfate, arsenate as the electron acceptors, and the temperature/pH was adjusted to be close to those found in sampling sites. The combination of different environmental samples, different substrates and different incubation conditions resulted in a huge number of enrichment cultures which involved enormous workforces (Master and PhD students, Post-docs and lab technicians). Although more than one thousand enrichments were set up, less than 10% of these survived three transfers and only these were subjected to DNA isolation, 16S RNA gene sequencing and further isolation of pure strains.

From the enrichments that survived three or more transfers, we isolated pure strains and the information is presented in Table 1.

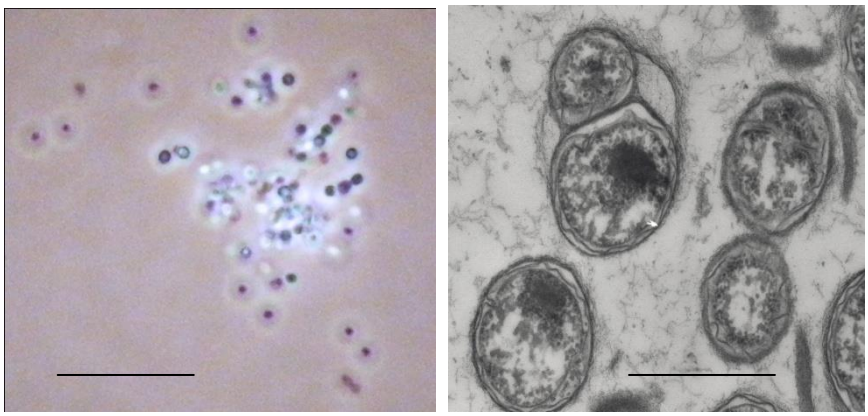
**Table 1. Summary of the isolates growing on polymeric substrates**

| Name | Origin | Substrate | T ($^{o}$C) | pH | Closest relative and 16S rDNA identity |
|---|---|---|---|---|---|
| 7T | Iceland | Xylan | 55 | 7 | *Thermoanaerobacterium aciditolerans 99%* |
| 7Tnr.1 | Iceland | XG | 55 | 7 | Thermomicrobium roseum 91% |

| | | | | | |
|---|---|---|---|---|---|
| 7Tnr1 A | Iceland | XG | | | Geobacillus Vulcani |
| 7Tnr.2 | Iceland | XG | 55 | 7 | Cohnella laevinbosi/thermotolerans 97% |
| 6Tnr.2 | Iceland | XG | 85 | 7(8) | Thermus aquaticus 96% |
| 6Tnr.3 | Iceland | XG | 85 | 7(8) | Thermus aquaticus/thermophilus 95% |
| 8Tnr.3 | Iceland | XG | 85 | 7 | Thermus antranikianii 99% |
| 2T 5 2 | Iceland | XG | 55 | 7 | Meiothermus Silvanus 99% |
| Is2-7*6.2 | Iceland | XG | 55 | 7 | Meiothermus Silvanus 99% |
| 7T nr2,1 | Iceland | Starch | 55 | 7 | Meiothermus Silvanus 98% |
| 7T nr4,1 | Iceland | Starch | 55 | 7 | Meiothermus Silvanus 98% |
| 7T nr4,2 | Iceland | Starch | 55 | 7 | Meiothermus Silvanus 98% |
| 2319x | Russia | Xylan | 85 | 6 | Thermococcus alcaliphilus/aegaeus 99% |
| 2319cl | Russia | Scmc | 75 | 7 | Caldicellulosiruptor owensenis/hydrothermalis 98% |
| 8-7 nr.1 | China | Xylan | 78 | 7 | Dictyoglomus 99% |
| 8-7 nr.2 | China | xylan | 78 | 7 | Fervidobacterium islandicum 98% |
| 2410 | Russia | PVA | 80 | 6 | A:Sulfolobus Islandicus 99% |
| DG#1 3,2 | Denmark | XG | 55 | 7 | Paenibacillus ginsengihumi 92 |
| DG#1 4,1 | Denmark | XG | 55 | 7 | Cohnella laeviribosi 97 % |
| Is3-14,2 | Iceland | XG | 55 | 7 | Thermus igniterrae 98% |
| Is3-24,1 | Iceland | XG | 55 | 5 | Alicyclobacillus sendaiensis 98% |
| Is3-24,6 | Iceland | XG | 55 | 5 | Alicyclobacillus sendaiensis 98% |
| Is3-24,4 | Iceland | XG | 55 | 5 | Alicyclobacillus sendaiensis 98% |
| Is3-23,3 | Iceland | XG | 70 | 5 | Alicyclobacillus acidocaldarius 99% |
| Is3-23,4 | Iceland | XG | 70 | 5 | Alicyclobacillus acidocaldarius 99% |
| DG#1 1,1 | Denmark | Xyl | 55 | 7 | Brevibacillus thermoruber 99% |
| DG#1 2,1 | Denmark | Xyl | 70 | 7 | Geobacillus vulcani 98% |
| DG#1 2,2 | Denmark | Xyl | 70 | 7 | Geobacillus vulcani 98% |
| DG#1 3,2 | Denmark | Xyl | 55 | 7 | Brevibacillus thermoruber 98% |
| DG#1 3,2 | Denmark | Xyl | 70 | 7 | Geobacillus thermoglucosidasius 99% |
| Is3-23,2 | Iceland | xyl | 70 | 5 | Alicyclobacillus sendaiensis 98% |
| Is3-23,4 | Iceland | xyl | 70 | 5 | Alicyclobacillus sendaiensis 99% |
| Is3-24,4 | Iceland | xyl | 55 | 5 | Alicyclobacillus sendaiensis 99% |
| Is3-24,5 | Iceland | xyl | 55 | 5 | Alicyclobacillus sendaiensis 98% |
| Is3-24,7 | Iceland | xyl | 55 | 5 | Alicyclobacillus sendaiensis 99% |
| Is3-21,2 | Iceland | xyl | 55 | 7 | Brevibacillus thermoruber 99 % |
| Is3-21,4.1 | Iceland | xyl | 55 | 7 | Geobacillus 99% |
| Is3-21,4.3 | Iceland | xyl | 55 | 7 | Geobacillus thermoleovorans 99% |
| Is3-23,1 | Iceland | xyl | 70 | 5 | Geobacillus thermoleovorans 99% |
| Is2-8 | Iceland | Xyl | 70 | 6 | Geobacillus kaustophilus 97% |
| Is3-21,3 | Iceland | PVA | 55 | 7 | Thermus brockianus 98 % |
| Is2-7* | Iceland | PVA | 55 | 6 | Thermus brockianus 99% |
| It-5 | Italy | PVA | 78 | 7 | Staphylothermus hellenicus 96% |
| It-5.1 | Iceland | Gelrite | 78 | 5 | Sulfolobus shibataea 99 % |
| 7T | Iceland | polyester | 70 | 7 | Rhizobium leguminosarum 91% |

| | | | | | |
|---|---|---|---|---|---|
| MBl-TLP | TauTona mine | Starch | 52 | 7.1 | *Vulcanibacillus modesticaldus 95%* |
| MRo-LL | TauTona mine | Starch | 60 | 7.1 | *Bellilinea Caldifistulae 99%* |
| MGr-11 | TauTona mine | XG | 40 | 7.5 | *Bacillus jeotgali* |
| MRo-4 | TauTona mine | MCC | 37 | 7.5 | *Sporosalibacterium fauoarense 99%* |
| BPi-2ag | Baikal sediments | Agarose | 54 | 7.3 | *Caloramator australicus 97%* |
| BPi-4-40 | Baikal sediments | Xylan | 40 | 7.3 | *Paenibacillus lautus 99%* |
| 2842 | Gorya-chinsk | XG | 47 | 7.5 | *Phycisphaera Mikurensis 80%* |
| 2918 | Kamchatka, Moutnovski | XG | 54 | 6.0 | *Phycisphaera Mikurensis 80%* |
| Rift-s3 | Guaymas Basin | MCC | 65 | 6.5 | *Thermosipho atlanticus 96%* |

The most significant discovery of WP1 is the isolation of a few Planctomycete strains which represent novel genus, family and order (Fig. 1). Very interesting enzymes have been cloned from these organisms and one was filed for patent application (WP5).



A B

Fig. 1 Cells of *'Tepidisphaera elegans'*, the representative of new genus, family and order in class *Phycishaerae*: cells in light microscope, bar, 10 µm (A); electron micrograph of thin sections bar, 1 µm (B).

**4.1.3.4. DNA sequencing**

We sequenced 15 cellular communities distributed across four continents and a deep sea marine sediments (Table 2), two viral communities, and a number of isolates (Table 3). To identify a potentially novel pathway involved in xanthan gum degradation, we sequenced the transcriptomes of the Planctomycete Thermogutta, in two different media with triplicates.

The major significant result is the creation of a DNA sequence database of high temperature microorganisms and their viruses for screening for targeted novel hydrolase activity.

**Table 2. Summary of metagenomes sequenced by HotZyme**

| Name | Sampling location | T ($^{o}$C), pH | Sequencing | Total seq |
|---|---|---|---|---|

| | | | method | (x $10^b$ bp) |
|---|---|---|---|---|
| **MW-2** | Yellowstone National Park, USA | 84°C, pH8.0 | Illumina (MiSeq) | 1031 |
| **RC-2** | Yellowstone National Park, USA | 85 °C, pH10.2 | Illumina (MiSeq) | 1302 |
| **NL-10 0908** | Yellowstone National Park, USA | 89-92 °C, pH3.0-5.5 | 454 Titanium | 200 |
| **NL-10 0808** | Yellowstone National Park, USA | 89-92 °C, pH3.0-5.5 | 454 Titanium | 200 |
| **CH1102** | Yellowstone National Park, USA | 79.3°C, pH 1.83 | 454 | >1000 |
| **Ch2-EY65S** | Hot spring sediments, Yunnan, China, | 67 °C, pH 7 | Illumina | 1031 |
| **Ch2-EY55S** | Hot spring sediments, Yunnan, China | 55 °C, pH 7 | Illumina | 1036 |
| **Is2-5** | Hot spring sediments, Iceland | 85-90 °C, pH 5 | Illumina | 1033 |
| **Is3-13** | Hot spring sediments, Iceland | 90 °C, pH 3.5-4 | Illumina | 1005 |
| **It-6** | Hot spring sediment, Italy. | 76°C, pH 3-3.5 | Illumina | 1006 |
| **It-2** | Pisciarelli, Italy,<br>Hot soil close to the main pool | 49°C | 454 Titanium | 495 |
| **It-3** | Pisciarelli, Italy<br>Hot spring sediments | 85 °C , pH 3.5 | 454 Titanium | 486 |
| **NGI-7** | Drilling well DH6, 404 m below surface. Adventsdalen, Spitsbergen | 17 °C, experienced temp.before uplift 134 °C, | 454 Titanium | 454 |
| **Par-mat** | Parabel, Tomsk region, Russia | 46 °C, pH 7.3 | 454 Titanium | 303 |
| **Kamch-1** | Sun-spring, Russia | 61-64°C, pH5.8 - 6.0 | 454 Titanium | 356 |

**Table 3. Sequenced isolates and enrichment (enrichments labelled *)**

| Name | Species | T (°C), pH | Substrate | Origin |
|---|---|---|---|---|
| B1-7Tnr1 | Thermomicrobium | 55°C, pH7 | xanthan gum | Iceland |
| B3-6Tnr4 | unknown | 85°C, pH7 | xanthan gum | Iceland |
| B4-12Tnr1 | unknown | 85°C, pH7 | xanthan gum | Iceland |
| A5-7T | Thermoanaerobacterium | 55°C, pH7 | xylan | Iceland |
| A6-2319x | thermococcus | 85°C, pH6 | xylan | Russia |
| A7-Loc1F3* | enrichment mix | 78°C, pH6 | PVA | Iceland |
| R1 | Thermogutta | 60°C, pH6 | XG, xylan etc | Russia |
| DG#1 4,1xg | Cohnella laevinibosi 97% | 55°C, pH7 | xanthan gum | DK |
| It6 | Sulfolobus | 78oC, pH3 | Xylan | Italy |

### 4.1.3.5. Bioinformatics method development

We tested the most widely used assemblers for 454 and Illumina HTS data on our metagenomic and isolate/enrichment sequencing data and established appropriate protocals for the assembly of different types of sequencing data. In this way, assemblies of all metagenomic and enrichment/isolate sequencing data have been completed. For velvet assemblies, we chose *k*-mer sizes between 31 and 71 for individual assemblies, which were then merged using Minimus2 into meta-assemblies. 454-sequenced samples were assembled by Mira, Celera and Newbler and merged into meta-assemblies using the same approach.

We developed *de novo* gene prediction platform ANASTASIA (Automated Nucleotide Aminoacid Sequences Translational plAtform for Systemic Interpretation and Analysis), which integrates the three programs MetaGeneMark, Prodigal and MGA. Using the pipeline, *de novo* gene prediction was performed on all assemblies. Those predicted proteins were further compared to protein databases using BLASTP and HMMER, and subjected to functional classification (WP4). Predicted genes were annotated using the ANASTASIA protein sequence classification tools based on EC numbers and homology to known hydrolase domains. *De novo* gene prediction was done on the meta assemblies only, which however have a lowest contig length of 200 nt. Additionally, singleton reads longer than 200nt were also included as part of the meta-assemblies, in order to not miss protein sequences from those reads.

SignalP and Phobius were used to detect signal peptide sequences in predicted proteins. The hidden Markov models used by SignalP show sufficient sensitivity on archaeal protein sequences and in the case of missed positive predictions, homology-based sequence annotation would complement the assignment of signal peptide status for a given protein sequence.

**4.1.3.6. Global biodiversity of hot environments**

To facilitate storage and analysis of the sequencing data generated during the HotZyme project, partner UCPH established the server "Helios" with 36 CPU cores, 256 GB RAM and 36 TB of storage capacity. The server is accessible by all partners through the ANASTASIA analytical platform, following a user restricted model. This infrastructure has been customized in order to meet the computational needs of the HotZyme consortium through the installation of the appropriate analytical tools and the allocation of adequate storage capacity for each group regarding the storage of raw sequencing data and the results derived from the corresponding analysis.

Taxanomic content of the hot spring metagenomes (Table 2) were analyzed using MEGAN and compared between samples. The results were summarized and published in Menzel et al. 2015 (see publication list in Section 4.2.A). Further, the viral sequences were retrieved from the metagenomes and analyzed in more detail. A manuscript was prepared based on the results and is more or less ready for submission.

**4.1.3.7. *In silico* identification of novel hydrolases**

Prediction of protein function was carried out by utilizing homology-based methodologies and machine learning methodologies. The homology-based analysis was perforemd by exploiting the capabilities of the ANASTASIA platform with the integrated BLAST and HMMER programs. BLAST analysis was performed against two databases; NCBI-nr and a custom database built by partner NTUA from all the annotated hydrolases of UniProt databases. The NCBI-nr database was used in the analysis for comparative purposes regarding the hits that showed low homology to the hydrolase database. HMMER analysis was performed also against two databases; Pfam-A and a customized database built by NTUA containing all the Hidden Markov Models of the representative sequences from all the hydrolases of the UniProt database. The machine learning methodologies were implemented by the help of a hydrolase classifier software that could assign putative EC numbers to unknown sequences built by NTUA as well as the exploitation of the open source software EFICAz. Both tools were integrated in the ANASTASIA platform and became modules of the annotation workflows that were used to analyse the metagenomic sequences.

Priority class 1 enzymes and priority class 2 enzymes were defined by the consortium, and *in silico* searches of the enzyme encoding genes in the new sequences generated by WP2 were performed through ANASTASIA. This generated a long list of potentially interesting genes which was distributed to WP5 and WP6 partners for further selection and cloning.

**4.1.3.8. Development of new high throughput enzyme screening methods**

As efficient and generally applicable screening methods for finding new enzyme activities are scarce, the HotZyme partners invested efforts on the development of new screening methods, including the Microcolony-based screening system and *in vivo* reporter systems.

The Microdish culture chip (MDCC) is a highly compartmentalized porous ceramic based cultivation system for the high density cultivation of microorganisms on a solid surface. The most commonly used MDCC180.10 contains 3300 circular 180 µm diameter wells with a footprint of only 2.88 $cm^2$. The colony density is thus almost two orders of magnitude higher compared to direct inoculation on an agar-based medium. We found that direct functional screening of *E. coli* expression libraries for enzyme activity is challenging due to diffusion problems and the fact that ideal screening conditions (defined buffer system, temperature of 60-80 °C) are not compatible with subsequent recovery of viable clones. Therefore, a replica-plating procedure, termed "microcolony-lift" was developed during the course of the project (Fig. 2).

We also attempted to develop general reporter assays that can be used to rapidly screen expression libraries. In this regard, two different systems were investigated, the transcription regulator based selection system and the riboswitch-based selection/screening system. While the latter is still at the proof of concept stage, significant progress has been made for the former.

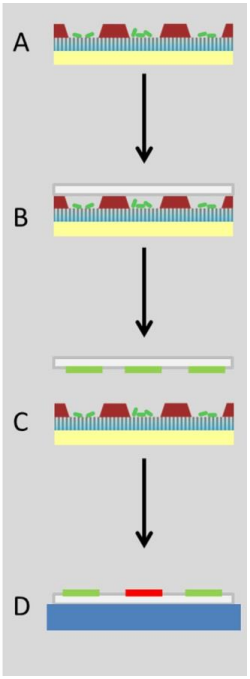



**Figure 2**. Schematic outline of the microcolony transfer procedure. (A) Cross section showing three compartments of an MDCC (blue and red) lying on top of a nutrient source such as LB-agar for *E. coli* (yellow). Growing bacteria are depicted in green. (B) A nitrocellulose membrane is placed on top of the growing microcolonies for transfer. (C) Colonies stick to the membrane (in light green) while a portion of the colonies remains on the MDCC. (D)The transferred cells are used to probe enzymatic activity e.g. by a colorimetric method. The red colour indicates a colony that shows activity.

Different versions of the selection and screening system, varying the selection reporter (kanamycin resistance marker KmR or leucine auxotrophy complementation with LeuB), the copy number (low or medium) and the transcriptional regulator (AraC or LacI), have been developed and characterized. The two best performing system versions in terms of dynamic

range, leakiness and sensitivity were the medium copy version with AraC as regulator and KmR as selection reporter and the low copy version with LacI as regulator and LeuB as selection reporter (Figure 1.5). If changing the inducer specificity is successful the system will be further adapted to make it applicable in finding novel biocatalysts.

### 4.1.3.9. Custom synthesis of new enzyme substrates

The industrial partner SIAL provided continuous support to WP5 and WP6 by providing custom-synthesized substrates for biochemical analyses of novel enzymes (Table 4). Some of these have already been commercialized or are in the process of commercialization and were crucial for a joint publication lead by partner UDE (Kallnik et al., *J. Biotechnol.* 2014; for details see HotZyme dissemination record).

**Table 4.** New compounds synthesized by partner SIAL for screening and enzyme characterization studies

| **Molecule name** | | **Commercialized** |
|---|---|---|
| Epoxides | RS-Vitamin $K_1$ 2,3-epoxide | yes |
| | RS-Vitamin $K_3$ 2,3-epoxide | yes |
| | (3S)-2,3-Squalenoxide | |
| Sugar lactones | L-Fucono-1,4-lactone | yes |
| | L-Xylono-1,4-lactone | |
| | D-Galactono-1,4-lactone | |
| | D-Altrono-1,4-lactone | |
| Other chiral lactones | (R)-γ-Caprolactone | yes |
| | (S)-γ-Caprolactone | yes |
| | (R)-γ-Valerolactone | yes |
| | (S)-γ-Valerolactone | yes |
| | (R)-γ-Valerolactone | yes |
| | (S)-γ-Valerolactone | yes |

### 4.1.3.10. Construction of expression libraries

A total of 8 *E. coli* expression libraries have been constructed for screening purposes (Table 5). The libraries have undergone various levels of quality control including a complementation test with a purine auxotrophic *E. coli* strain. Libraries were distributed to the partners involved in screening activities.

**Table 5**. Small insert expression libraries constructed by Hotzyme consortium members

| Code | Source of insert DNA | Donor temp. optimum (°C) | Donor pH optimum | Primary target | Means of validation |
|---|---|---|---|---|---|
| **CHv 2012-01-14** | hot spring metagenome, viral fraction | 85 | 2.9 | n.a. | restriction enzyme analysis, sequencing of random clones |
| **CHc** | hot spring | 85 | 2.9 | n.a. | restriction enzyme analysis, |

| | | | | | |
|---|---|---|---|---|---|
| **2012-01-14** | metagenome, cellular fraction | | | | sequencing of random clones |
| **Kam2410-PVA** | Kam2410-PVA enrichment | 85 | 3-6 | poly(vinyl) alcohol | restriction enzyme analysis of random clones |
| **2312** | Mixture cotaining Planctomycete | 55 | 7 | various carbohydrates | complementation of E. coli purine auxotroph, sø3274 |
| **DG4.1** | Cohnella sp. | 55 | 7 | Xanthan gum, xylan, Starch | complementation of E. coli purine auxotroph, sø3274 |
| **YNP NG05** | Yellowstone "lifeboat" hot spring metagen. | 59 | 3.3 | n.a. | restriction enzyme analysis, sequencing of random clones |
| **DXG** | Dictyoglomus sp. | 75 | 7 | xanthan gum | restriction enzyme analysis |
| | Thermoanaerobacter | 60 | 7 | epoxide | complementation of E. coli purine auxotroph, sø3274 |

### 4.1.3.11. Enzyme screening

Hundreds of hydrolases of industrial relevance have been identified during the course of the HotZyme project by different and complementary approaches, including over 200 glycosyl hydrolases (GHs), dozens of lipolytic enzymes and a few proteolytic enzymes. Approximately 100 of these have been selected for cloning and expression in a recombinant host, providing a rich resource for the selection of candidates for further analysis by biochemical and structural analyses in WP6.

Of the cloned genes, 36 GHs, 10 lipolytic enzymes and one protease were successfully expressed at various levels in different hosts, *E.coli*, *Aspergillus Orizae* and/or *Sulfolobus acidocaldarius.* Futher biochemical and structural analyses were performed for many of these. Two of the enzymes will make the foundation for patent applications. One GH will be the founding member of a new CAZy glycoside hydrolase family.

### 4.1.3.12. Biochemical and structural analyses of selected hydrolases.

About 15 enzymes were selected for detailed analyses, including esterases, lactonases, epoxide hydrolases, cellulases and proteases. The selected hydrolases were purified in large amounts and characterized in respect to their biochemical and enzymatic properties.
The major focus of the enzymatic characterization was set at substrate specificity, kinetic parameters, (Vmax-, Km-, Ki-values) as well as the stability at different pH-values, temperatures. In addition, the effect of addition of organic solvents as well as activity in organic solvents, ionic liquids, supercritical carbon dioxide, and micro-emulsion systems were determined.

Crystal structures of eight of the enzymes were successfully resolved (Fig. 3), and more are on the way.

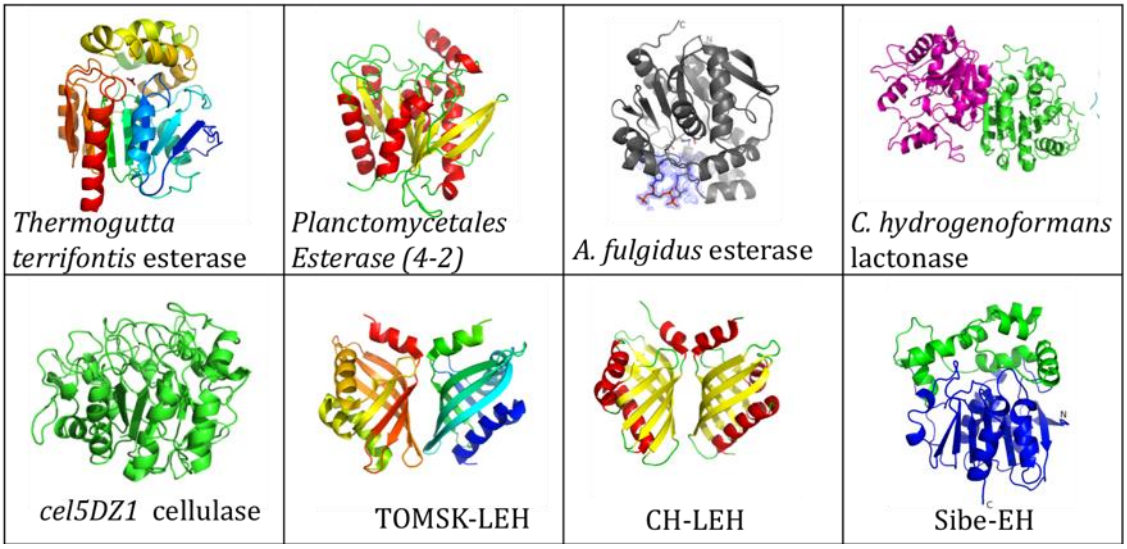



**Fig. 3**. Crystal structures of the enzymes.