Horizon 2020 Work programme
Food Security, Sustainable Agriculture and Forestry, Marine, Maritime and Inland Water Research and the Bioeconomy
Call
H2020-FNR-2020: Food and Natural Resources
Topic name
FNR-16-2020: ENZYMES FOR MORE ENVIRONMENT-FRIENDLY CONSUMER PRODUCTS

FuturEnzyme:
Technologies of the Future for Low-Cost Enzymes for Environment-Friendly Products
Final ID: 101000327

23/12/2022

SET OF 250,000 SEQUENCES PRE-SELECTED

D2.2

MANUEL FERRER
CSIC
Marie Curie n2, 28049, Cantoblanco, Madrid

Document information sheet

Work package:	WP2, Machine learning enzyme bioprospecting integrated into an industrial context
Authors:	CSIC (Manuel Ferrer, Patricia Molina)
Document version:	2
Date:	23/12/2022
Starting date:	01/06/2021
Duration:	48 months
Lead beneficiary:	CSIC
Participant(s):	CSIC, BSC, Bangor, UHAM, UDUS
Dissemination Level:	Confidential, only for consortium's members (including the Commission Services)
Type	Other
Due date (months)	6
Contact details:	Manuel Ferrer, mferrer@icp.csic.es

Summary

SET OF 250,000 SEQUENCES PRE-SELECTED

1. Scope of Deliverable

This deliverable consists in at least 250,000 full-length candidate sequences encoding enzymes relevant to FuturEnzyme (according to the initial proposal). These sequences were selected in the frame of the Task 2.2 by homology and computational search protocols applied to the sequence space in public and consortium repositories. The sequences, to be compiled in fasta file or the Excel tables containing the sequences and information of retrieves and pre-selected sequences, are deposited in the FuturEnzyme internal repository with a report, herein summarized, detailing the tools used and the repositories screened, and outcomes. More in detail, in this deliverable report we present the in silico methods to screen for enzymes relevant to FuturEnzyme, that include: i) DIAMOND BLASTP, PSI-BLAST (EMBL-EBI) and Hidden Markov Model (HMM), high-throughput programs for aligning protein sequences against protein reference databases; ii) MCL algorithm (Markov Cluster Algorithm), an efficient algorithm for large-scale detection of protein families through network analyses; iii) PELE (Protein Energy Landscape Exploration), a protein-ligand Montecarlo simulations software; and iv) the machine learning EP-Pred ensemble classifier. These tools were applied to search enzymes relevant to the project in public sequence repositories and FuturEnzyme genomes and metagenomes sequences. After screening more than 1 billion sequences, about 3.16 million sequences encoding target enzymes were retrieved and pre-selected, which are available in the internal FuturEnzyme repository. The number of pre-selected sequences significantly exceeds that of the initially planned when the project was submitted. This difference is not due to a downward assessment of the initial proposal but to an increase in the computational and bioinformatics capabilities developed and a greater capacity to generate new sequences, which has allowed us to access a greater number of bioinformatics and computational analyses.

2. Reasons for the update

The first version of the Deliverable D2.2 was submitted in November 2021. This update is due to the fact that since the submission, the partners were able to retrieve a new set of sequences. These sequences were identified either by applying the same tools or adapted and improved versions to subsequently retrieve a new set of sequences. In November 2022, the Coordinator (Manuel Ferrer) contacted the Project Officer (Colombe Warin) to explain these circumstances and ask her to re-open the submission of this deliverable (amongst others), at which she agreed.

3. Origin of the deliverable

Along the already 18 months of project, one deliverable has been accomplished from which the present one nourishes. To be mentioned:

Deliverables in the frame of WP2:

- D2.1: Manufacturers' needs and specifications: protocol (August 2021, updated December 2022)
  In this deliverable, information about manufacturers' needs, and enzymes and products specifications (working/storage conditions and stabilities, compositions, etc.) for implementing 3 innovative, real-life, and environment-friendly products (detergents, textiles and consumer care products) are detailed. They include:
  Detergent: Enzymes for removing fatty oil stains
  Cosmetic: Enzymes for degrading hyaluronic acid to size controlled products to be integrated into cosmetics
  Textile: Enzymes for the removal of spinning additives and dyes
- D2.4_Set of 180 enzymes for experimental focus (July 2022; updated December 2022)
  In this deliverable, at least 180 enzymes from the priority sequences retrieved in the frame of WP2 (deliverables D2.2, D2.3) and WP3 (deliverable D3.3), were preliminary selected to proceed with their cloning, synthesis, expression and characterization.

Deliverables in the frame of WP3:

- D3.1: Bio-resources prepared and exchanged (July 2021; updated December 2022)

This deliverable enlists a set bio-resources (enzymes, isolates, enrichment cultures, clone metagenomic libraries, genomes and shotgun metagenome sequences) generated in the framework of previous European and national-funded projects, and that were compiled and exchanged within the consortium, at the beginning of the project, for screening those relevant to FuturEnzyme.

- D3.3: Set of 100 clones, 10 isolates, 10 enzymes shortlisted for sequencing (March 2022; updated December 2022)
  In this deliverable, bio-resources available before the beginning of the project and newly generated during the project were screened by naïve/functional methods to identify those with interest for our project. Bio-resources include previous and new enzymes, environmental samples, isolates, enrichments, and clone libraries that were checked for the purpose of the present project, and best selected ones sequenced and sequences with interest for our project were retrieved.
- D3.4: Sequence, activity, and stability datasets from best positive bio-resources (November 2022)
  This deliverable enlists a set new bio-resources (enzymes, isolates, enrichment cultures, clone metagenomic libraries, genomes and shotgun metagenome sequences) generated during the project for screening, with indication of those found to be positive for the purposes of the project.

Deliverables in the frame of WP4:

- D4.6_The metadata on expression yield, activity and stability available (November 2022)
  This deliverable consists on the datasets informing about the expression yield, activity and stability of all enzymes generated in the project until month 18.

4. Methodology: Source and profiling of enzymes

4.1. Design of reference database

Based on the information provided in Deliverable 2.1, a number of enzymes were selected as study targets. Priority enzymes include lipases-esterases, peroxidases- and laccases-like oxidoreductases, polyester, plastic degrading hydrolases, cutinases, and hyaluronidases. Secondary target enzymes include peptidases, amylases, amidases, and lactonases. One reference database for each of these families was generated using the NCBI repository and FuturEnzyme repository to help BLAST search. The databases included the closest protein homologs of all protein families of interest, and at least one representative sequence from all taxonomic groups (containing such enzymes) was represented (Annex **File 1**). In details, the established and manually curated and customized database contains 37,403 taxonomically diverse protein sequences featuring the key enzyme families, potentially targeting enzymes relevant to the detergent, textile and cosmetic applications that are objectives of FuturEnzyme. The sequences are available in FASTA files, one per each of the target enzymes (Annex **File 1**).

4.2. Source of sequences

A number of public sequence repositories and internal FuturEnzyme sequences, all together accounting more than 1 billion sequences, were targeted for enzyme search, which are detailed below.

Sequences retrieved from 10 public sequence repositories (comprising more than 670 million sequences), included:

- CAZy database (http://www.cazy.org/)
- MarDB - Marine Metagenomics Database (https://mmp2.sfb.uit.no/)
- MarFun - Marine Metagenomics Database (https://mmp2.sfb.uit.no/)
- MarRef - Marine Metagenomics Database (https://mmp2.sfb.uit.no/)
- NCBI non-redundant database (https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/)
- UniProt database (https://www.uniprot.org/)
- Integrated non-redundant gene catalog (IGC database) (http://gigadb.org/dataset/100064)
- Human microbiome database (https://commonfund.nih.gov/hmp/databases)
- Lipase Engineering Database (http://www.led.uni-stuttgart.de/)
- Tara Ocean metagenome (http://ocean-microbiome.embl.de/companion.html)

In addition, the sequences of 64 metagenomes obtained by shotgun sequencing and the sequences of 223 genomes from isolates detailed in Deliverable D3.1 “Bio-resources prepared and exchanged”, were compiled.

They were generated in the framework of previous European and national-funded projects, and were compiled and exchanged within the consortium, at the beginning of the project (see Deliverable 3.1 for details). Finally, the sequences of 54 metagenomes obtained by shotgun sequencing and the sequences of 22 genomes from isolates detailed in Deliverable D3.4 "Sequence, activity, and stability datasets from best positive bio-resources", were also compiled; this last set of sequences were newly generated during the project (see Deliverable 3.4 for details). They all together comprised more than 400 million sequences.

4.3. DIAMOND BLASTP and Network analysis

The sequences encoding enzymes relevant to FuturEnzyme were selected by DIAMOND BLASTP, using default parameters: percent identity >60%; alignment length >70; e-value < $1e^{-5}$. Once enzymes are retrieved by DIAMOND BLASTP, a pre-selection analysis is undertaken. For that, BLASTP (default parameters, percent identity >60%; alignment length >70; e-value < $1e^{-5}$) is performed against all of them, keeping only the alignments with a percentage of identity higher than 50%. With these results, an identity percentage network is built. Then, we clustered the sequences using the MCL algorithm, implemented in the software of the same name (Markov Cluster Algorithm: Enright A.J., Van Dongen S., Ouzounis C.A. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7):1575-1584, 2002; using the parameter Inflation = 1.4). This method is widely used to obtain clusters in sequence networks. With the sequences of each cluster, a multiple alignment using ClustalW (default parameters) is performed, obtaining from it the consensus sequence and a list of reference sequences conforming each of the clusters. **Figure 1** summarizes the DIAMOND BLASTP and MCL pipeline for enzyme pre-selection (Annex **Figure 1)**.

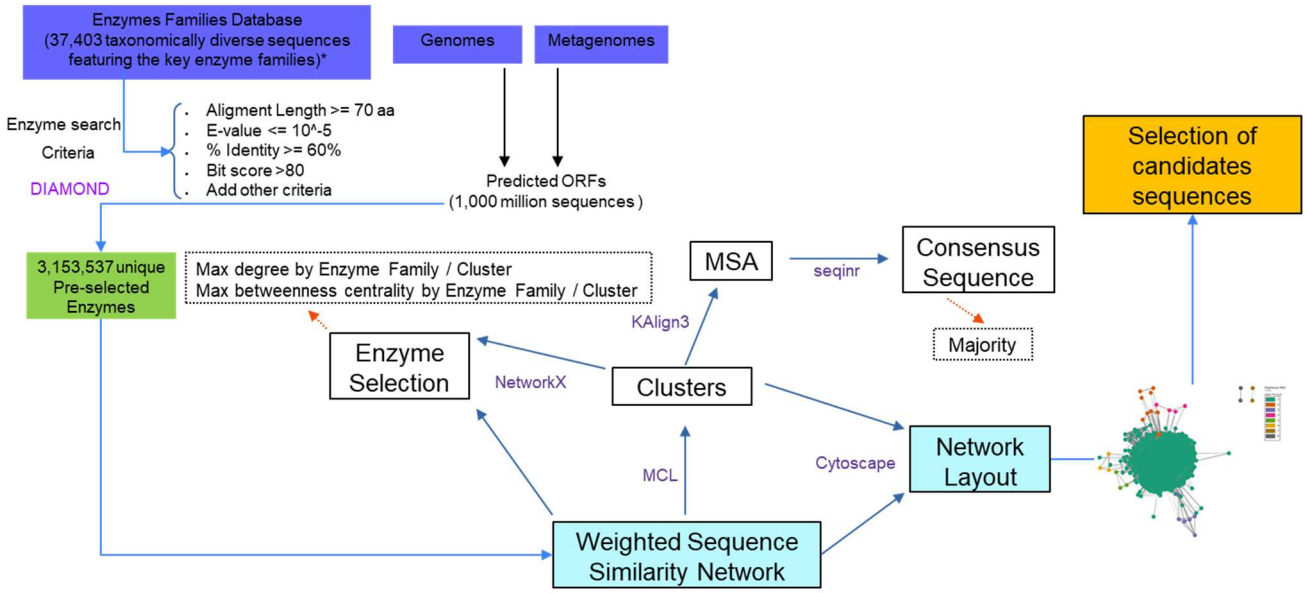


Figure 1. DIAMOND BLASTP and MCL pipeline for enzyme pre-selection.

4.4. Hidden Markov Model (HMM)

The sequences encoding enzymes relevant to FuturEnzyme were also selected by highly sensitive Hidden Markov Models (HMMs) with the AHA-Tool pipeline. This tool automatizes the processes of sequence alignments and HMM construction, in silico database screening and gathering of useful information for candidate selection, such as secretion signals or taxonomical origin of the hits. The constructed models allow detecting active enzymes with a higher success, since all the sequences used to build the models have been tested active previously. Thus, the process of expanding the diversity of active enzymes in the collection is expected to be fast and efficient. In some cases, HMMs was performed to identify homologues to one enzyme that in the frame of the FuturEnzyme project was selected among the prioritary targets, such as the lipase Lip9. In this case, the sequence of Lip9 was compared against NCBI's database (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/nr.gz) and in-house collections of genomes and shotgun sequences of metagenomes obtained by shotgun sequencing. Alignment was performed with diamond 2.0.15.153 (Buchfink et al., 2021 https://doi.org/10.1038/s41592-021-01101-x) and alignments within 25%

range of top alignment score were reported. **Figure 2** summarizes the HMMs pipeline for enzyme pre-selection.

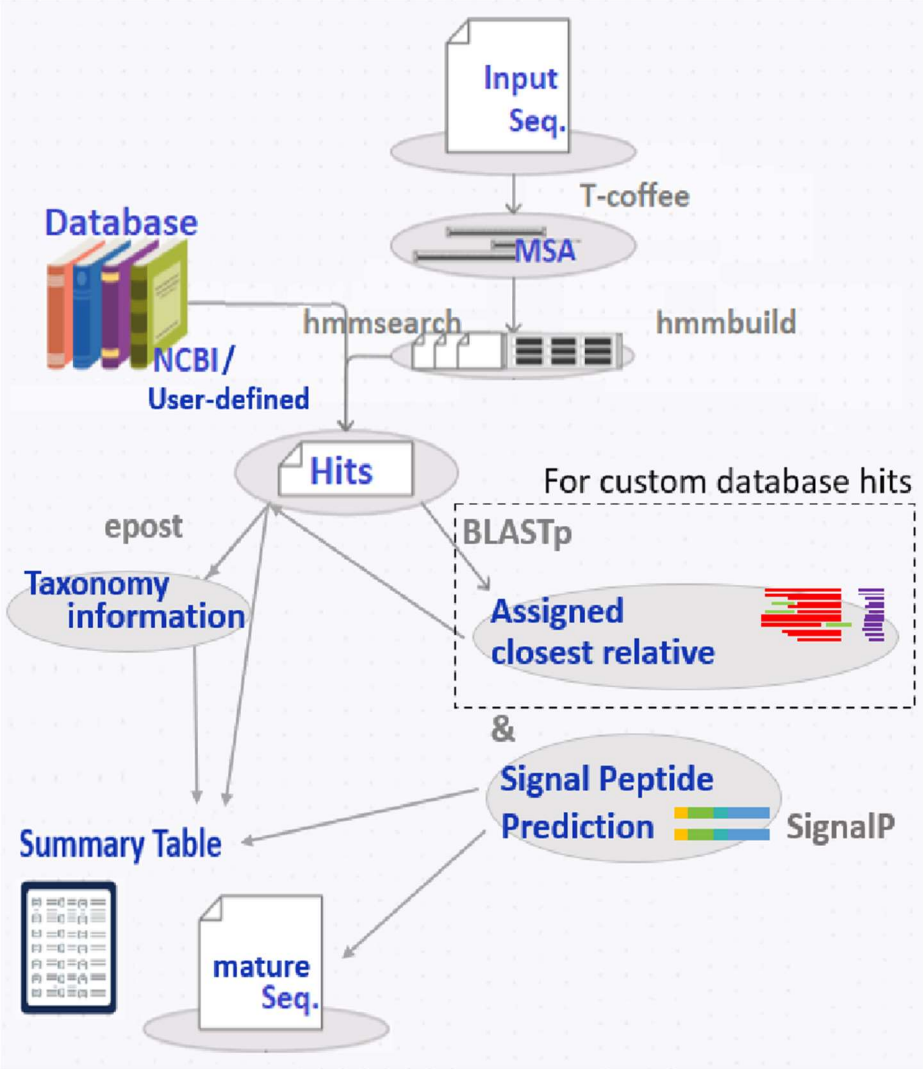


Figure 2. HMMs pipeline for enzyme pre-selection.

4.5. Search of new sequences combining PSI-BLAST and PELE (Protein Energy Landscape Exploration)

In some cases, PSI-BLAST and PELE (Protein Energy Landscape Exploration) were used in combination to identify homologues to one enzyme that in the frame of the FuturEnzyme project was selected among the prioritary targets, such as the lipase Lip9. In brief, we applied PSI-BLAST to find a reasonable amount of similar enzymes in databases and test them using the protein-ligand Montecarlo simulations software, PELE (Protein Energy Landscape Exploration). By doing so, one can ensure that the selected sequences are not overly similar to an original enzyme, but do retain similar or improved characteristics. More in details, for the bioprospecting of Lip9, we searched for hundreds of sequences by means of PSI-BLAST, a tool designed to find distant homologs for a certain protein, using Lip9 as the seed. Then, we performed filtering of the sequences through different parameters, including AlphaFold confidence level, the alignability of the catalytic residues to the Lip9 catalytic triad, the existence of a spatially well designed triad in the AlphaFold models, and the low resemblance to a patented lipase (WP_106066877.1). We docked 15 different ligands, all of them being triglycerides constituting different grease stains. Finally, we ran protein-ligand simulations on the selected new sequences with our in-house all-atom Monte Carlo molecular modelling sampling technique, PELE. **Figure 3** summarizes the PSI-BLAST and PELE pipeline for pre-selecting enzymes homologous to Lip9.

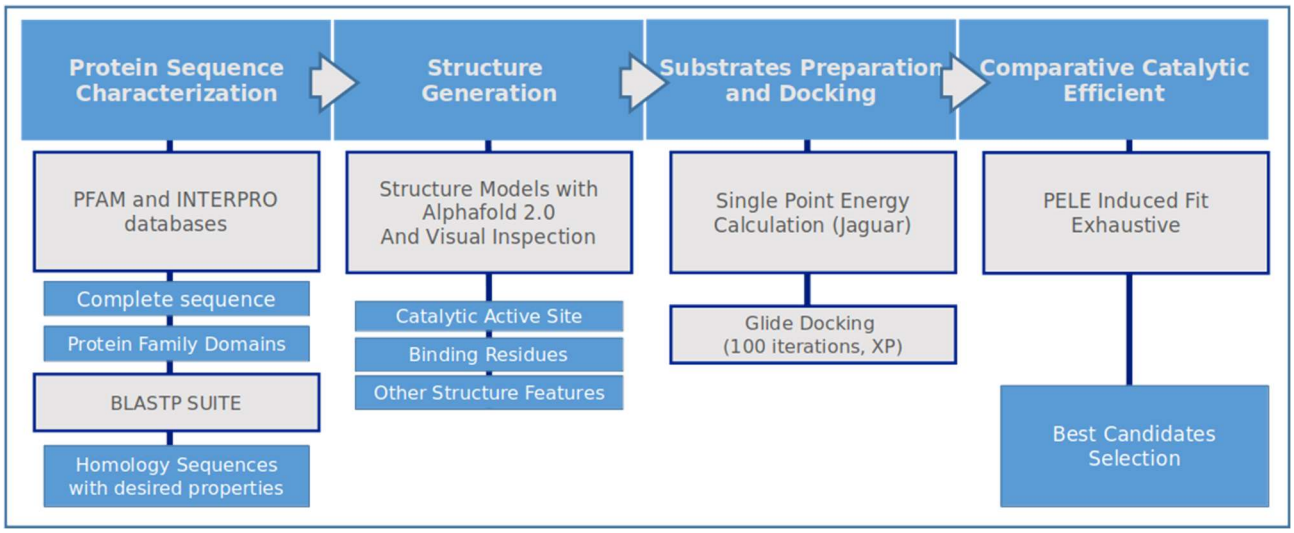


Figure 3. PSI-BLAST and PELE pipeline for enzyme pre-selection.

4.6. Search of new sequences using the machine learning EP-Pred method

A machine learning tool for bioprospecting enzymes relevant to FuturEnzyme was also implemented. Briefly, we implemented a method called EP-Pred, an ensemble binary classifier built to predict the promiscuity of ester hydrolases. It combines 3 different machine learning algorithms: Support vector machines (SVC), K-nearest neighbours (KNN) and a lineal model (the RidgeClassifier implementation on Sckit-Learn. It was trained on a dataset containing 147 phylogenetically diverse esterases and their activity on 96 distinct ester substrates. The labelling of the classes was based on the number of substrates catalyzed, where 20 or more substrates were considered promiscuous and less than 20, non-promiscuous. The program can be downloaded in GitHub etiur/EP-pred: A machine learning program to predict promiscuity of esterases (github.com). To use it is required to install 3 external programs: Ifeature, Possum and Blast+ NCBI. It is also need a protein database, in this case, the Uniref50. The main.py script will then perform the rest of the operations if provided with the input esterases and the appropriate flags. It will transform the uniref50 into a Blast database and use it to extract the PSSM profiles. It will generate the features used by the models using Ifeature and Possum and finally it will predict the promiscuity of the sequences. EP-Pred has been evaluated against the Lipase Engineering Database (http://www.led.uni-stuttgart.de/) together with a HMMs approach leading to select sequences encoding esterases and lipases. For extensive details see our recent reference (https://www.mdpi.com/2218-273X/12/10/1529). **Figure 4** summarizes the EP-Pred pipeline for enzyme pre-selection.

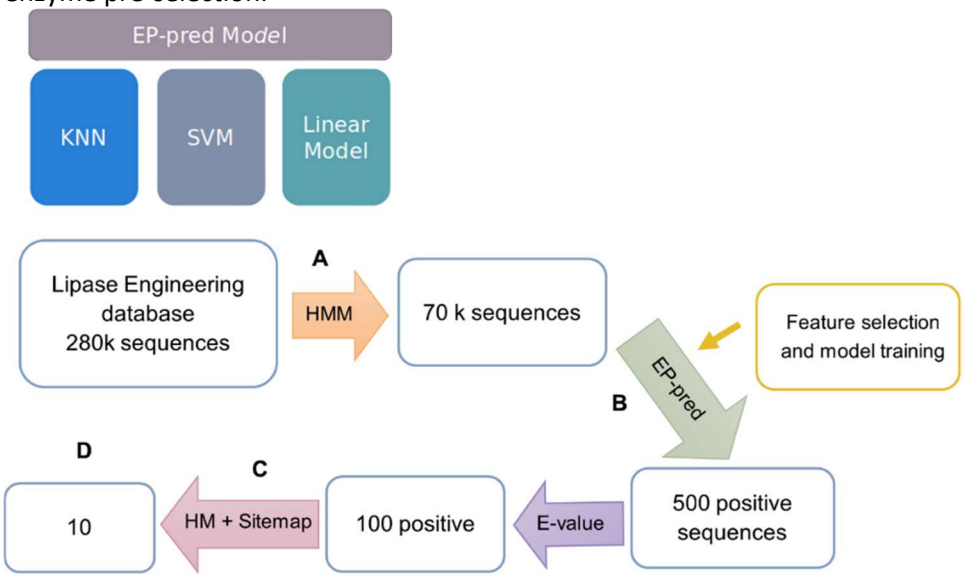


Figure 4.EP-Pred pipeline for enzyme pre-selection.

5. Results

5.1. General overview

Using the different methods described in Section 4, a total of total 3,163,785 sequences were retrieved and pre-selected (**Table 1**).

Table 1. List of selected candidates per each of the reference enzyme classes.

Enzyme class	Enzymes pre-selected in Deliverable D2.1				
	DIAMOND-BLAST[8]	BLAST[8]	HMM	EP-PRED	PSI-BLAST + PELE
Amidases[1]	194	24	0	0	0
Glycosidases[1,2]	1,049,254	0	0	0	0
Hyaluronidases[3,4]	468,020	0	24	0	0
Lactonases (COG1735/EC3.1.1.25)[2]	121,824	0	0	0	0
Lipases-Esterases[1,2]	546	282	409	506	525
Polyester, plastic degrading hydrolases[1,5]	194,375	23	9,115	0	0
Cutinases[1, 2,6]	258,166	0	0	0	0
Peptidases[1,7]	1,044,289	0	0	0	0
Peroxidases- and laccases-like oxidoreductases[2]	16,189	20	0	0	0
SUB-TOTAL	3,152,857	349	9,548	506	525
TOTAL					3,163,785

[1]Relevant to textile sector
[2]Relevant to detergent sector
[3]Relevant to cosmetic sector
[4]Include EC3.2.1.36, 4.2.2.1, 3.2.1.35, pfam01630, cd01083
[5]Include EC 3.1.1.102
[6]Include EC3.1.1.74
[7]Include EC3.4.22.32, EC3.4.22.2, EC3.4.21.14, EC3.4.21.62, M04
[8]Pre-selected sequences retrieved by BLASTP and DIAMOND BLASTP when the manually curated and customized database contain 37,403 taxonomically diverse protein sequences featuring the key enzyme families relevant to FuturEnzyme were subjected to BLASTP and DIAMOND BLASTP against public sequence repositories and internal FuturEnzyme sequences (see Deliverables D3.1 and 3.4 for details).
[9]Pre-selected sequences retrieved when the FuturEnzyme sequences (see detailed in Deliverables D3.1 and 3.4) were subjected to BLASTP against NCBI.

5.2. Sequences pre-selected by BLAST, DIAMOND BLASTP and Network analysis

A total 3,152,857 sequences were pre-selected by applying BLASTP and DIAMOND BLASTP against public sequence repositories and internal FuturEnzyme sequences detailed in Section 4.2, using also the manually curated and customized database contain 37,403 taxonomically diverse protein sequences featuring the key enzyme families relevant to FuturEnzyme (see Section 4.1). The selected sequences (**Table 1;** e-value < $1e^{-5}$) are available in FASTA files, one per each of the target enzymes (Annex **File 2**). Network analysis further revealed that they grouped into 457 clusters, each containing enzymes that most likely do show similar properties (Annex **File 3**). Reference sequences conforming each of the clusters were pre-selected and further checked whether the sequence contained the proper catalytic domains and catalytic residues, and the presence of signal peptides, which are detailed in Deliverable D2.3 "Set of 1,000 enzymes selected using motif screens". In parallel, we searched in the UniProt and the non-redundant GenBank databases UniProt database using PSI-BLAST (EMBL-EBI). The translated protein sequences generated in FuturEnzyme (see Deliverables D3.1 and 3.4 for details) were annotated using BLAST searches of UniProt and the non-redundant GenBank databases using default parameters. A total of an additional set of 349 sequences were pre-selected (**Table 1**; Annex **File 4**).

5.3. Sequences pre-selected by HMMs

The sequences encoding enzymes relevant to FuturEnzyme were also selected by highly sensitive Hidden Markov Models (HMMs) with their AHA-Tool pipeline. A total 9,548 sequences were pre-selected by applying

HMMs, that included sequences encoding hyaluronidases and polyester degrading hydrolases (PTEases) (see Deliverables D3.1 and 3.4 for details) (Annex **File 5**).

5.4. Sequences pre-selected by EP-Pred

The sequences encoding enzymes relevant to FuturEnzyme, namely esterases and lipases, were also selected by the EP-Pred ensemble classifier build to predict the sequences encoding such enzymes with high substrate spectra through the combination of 3 different machine learning algorithms. A total of 506 esterases and lipases were pre-selected by applying EP-Pred from the Lipase Engineering Database (http://www.led.uni-stuttgart.de/) (Annex **File 6**).

5.5. Bioprospecting of Lip9-homologous enzymes with presumptive more activity by PSI-BLAST and PELE

As detailed in the Deliverable 5.1. "The shortlist of at least 18 enzymes nominated for engineering", Lip9 was selected among the 22 enzyme candidates having characteristics of interest for the detergent and textile sectors, being selected for WP5 (genetic and supramolecular engineering), WP6 (large scale production) and WP7 (pre-industrial validations). In brief, this enzyme showed remarkable high activity at 40˚C and pH 9.5, high capacity to degrade all the stained fabrics tested (Pigment with oil on polyester/cotton (PC-09), Mayonnaise on cotton (C-S-05S), Lipstick, pink on polyester/cotton (P-S-16), Fluid make-up on cotton (C-S-17), High discriminative sebum BEY on polyester/cotton (PC-S-132), Beef fat on cotton (C-S-61) and Butterfat on cotton (C-S-10)), showing a preference for Butterfat on cotton (C-S-10), and it is stable in the presence of washing liquor. A BLASTP against the Patented Protein Sequences Database, revealed that Lip9 showed >95.5% identity to patented lipases. This is why, we applied a protocol to find Lip9-homologous sequences, but ensuring that the newly selected sequences are not overly similar to a patented lipase, which has a high percentage of similarity with Lip9. Below, the PSI-BLAST and PELE protocol applied and the outcomes are detailed.

5.5.1. Search and filtering of new sequences by PSI-BLAST

At first, a searched in the UniProt database using PSI-BLAST (EMBL-EBI) was performed. Using this procedure, several iterations of sequence searching can be done, finding more and more different sequences each one. A total of 10 iterations were done, finding 525 different sequences (Annex **File 7**). We used the EMBL-EBI database because the sequence IDs are shared with UniProt, and recently AlphaFold has been used to find the structures of the proteins of this database. This allowed to save a lot of time, because running AlphaFold takes a significant amount of resources and time. At this point, several filters were used to remove non-desired sequences. First, we downloaded the structures for almost all the found sequences. This led to 494 pdb files, leaving 31 sequences out of the set. In the next step, we removed the terminal parts of the proteins with less than 60% of confidence (given by Alphafold itself). This caused the worst elucidate proteins by AlphaFold to be left with very few or even zero amino acids. We used this step to further filter the sequence set, removing 30 more sequences, and thus leaving 464 protein structures. Another filter used was based on the catalytic triad. On the one hand, all the catalytic triads were searched, composed of serine, histidine, and aspartate, in all the structures. On the other hand, each of the sequences were aligned with the Lip9 sequence and the residues matching the catalytic triad of the lipase were extracted. Finally, all the sequences which had the same triad found at the sequence and structure levels were selected. This led us to a set of 348 proteins. Finally, a last filtering based on the similarity to a patented lipase (WP_106066877.1) was run. Using a threshold of 75% identity, a final set of 288 new sequences was selected.

5.5.2. Ligands used

A total of 15 different ligands were used, all triglycerides (**Table 2**). As the goal is to find enzymes capable of cleaning stains of different mixes of lipids, all the interesting triglycerides were tested, with the idea of clusterize the results by stain. In this way, the 288 different proteins with the 15 ligands were prepared, giving a total of 4320 protein-ligand systems (and PELEs to run). Note that the ligands go from small triglycerides, with only 2 carbons per fatty acid chain, to 18 carbons, with 0, 1, 2, or 3 unsaturations. As said before, these ligands can be grouped into lipid stains. For instance, coconut oil contains C8:0, C10:0, C12:0, C14:0, C16:0,

C18:1, C18:2, C18:3, and beef lard contains mostly C16:0, C18:0, and C18:1, but also can contain as minor molecular components C12:0, C14:0, C16:1, C17:0, and C18:2.

Table 2. Ligands (triglycerides) used for the PELE simulations.

Ligand name	Identification
GRP	C2:0
GT3	C3:0
GRB	C4:0
GT6	C6:0
GT8	C8:0
U10	C10:0
U12	C12:0
GMY	C14:0
U16	C16:0
I16	C16:1
U17	C17:0
U18	C18:0
I18/TOL	C18:1
D18	C18:2
T18	C18:3

5.5.3. System Preparation

For each of the 288 enzymes, we prepared the structures using PrepWizard from Maestro (Schrodinger), in order to include the possible remaining hydrogens and setting the correct protonation states of some residues. Then, the docking of the ligands to the new proteins was performed, and the best pose for each protein-ligand combination was selected, in order to have a starting pose for PELE. This selection was based on distance and Glide score: the selected pose was the one with the lowest Glide score which fulfilled the distance from any of the target atoms of the ligand to the oxygen of the catalytic serine. Protein-ligand combinations that did not achieve catalytic positions were removed from the set.

5.5.4. PELE simulations

For each of the systems, two different PELE simulations were run. The first one consisted in an extensive search of the best minima in the active site, using the inversely proportional setting of PELE. In this way, the ligand is forced to explore unexplored parts of the surface of the active site, and find better minimas in the energy landscape. The second PELE round consisted in a non-biased PELE simulation, and this is the one that was used to compare all the results.

5.6. Bioprospecting of Lip9-homologous enzymes with presumptive more activity by HMMs

The HMMs protocol was further applied to find Lip9-homologous sequences. Below, the HMMs protocol applied and the outcomes are detailed.

5.6.1. Search and filtering of new sequences by HMMs

The sequence of Lip9 was compared against public sequence repositories and internal FuturEnzyme sequences detailed in Section 4.2. Alignment was performed with diamond 2.0.15.153, and alignments within 25% range of top alignment score. As a result, a total of 409 Lip9-homologus lipases were pre-selected (Annex **File 8**), with coverage up to 100 % coverage and identities down to 85% for candidates pre-selected from public sequence repositories. We further checked whether the sequence contained the proper catalytic domains and catalytic residues, and the presence of signal peptides, which is indicative of higher lipase character as in case of Lip9. In brief, all 409 homologues were modelled with Swissmodel server (https://swissmodel.expasy.org/), and the following were discarded: incomplete sequences, sequences with

an indentity percentage with their crystal below 30%, sequences that model with a cristal too different from 7r25 (Lip 9 Crystal model PDB code), models without a properly located catalytic triad and sequences without a signal peptide (highlighted in yellow). As a results, two Lip9-homologus lipases were pre-selected.

> k127_15135326_1
MAGWVAGLACAIAVVSAVDAVAAPKQYPVVMDFATAIAKSAQNPAASPAGVNVPCTLTAEHPRPVVLINGTHASMMM
NWAGLGPTLANQGFCVYSTALGASASDQIQTCGPVADSIAQIASFVDDVLNRTGAQKVDLIGHSQGGLIAESYTKFYGRDK
VANVALLSPSTHGSDQSGTSVHPTDLGAQIASIGCPAVLDQLQSSDVVRELNTGPITVPGVNYTVIETRYEFIITPTPSAAFIQ
EPGVRNLIVQDYCPQDLSDHLSLAYSEPAWNLLIDAISARTGEISC
Metagenome: AGWS_m_17
Source: Wilhelmsburg soil_oil contaminated
35.2% identity Lip9
Triad: Ser141 His262 Glu229
Signal peptide: underlined

> k127_129897_3
MRRCVTVSVILFLAFVMWSGVASAAPTYPVPDSFLAGVPLELGNPGGSAPGSNDWSCVPSDAHPEPVVLVHGTGGARQT
NWAVYAPLLANEGYCVYSLTYGNFPELPWPLDAIGGMTPIDTGTAQIATFVDQVLSSTGASKVDLVGHSQGTLQANNYVK
FFGGADKVSKIVSLAPPWHGTYGNDQISVGRSMRALGIDDEVAAGFPVCGACPEMFQGSAFIDRMRADGVYVPGIEYANI
ATRYDELVVPYTSGIEPGPNTTNIVVQDDCEQDYSDHVAVAGSARAAGFVLNALDPAHPRDVPCRFVAPVAG
Metagenome: AGWS_m_58
Source: Elbe river_enrichment
32.1 % identity Lip9
Triad: Ser148 His276 Asp244
Signal peptide: underlined

6. Conclusions

In summary, at least 3,163,785 sequences featuring enzyme families relevant to the project were retrived and pre-selected. They have been selected after in silico screening a total of more than 1 billion sequences from public and FuturEnzyme sequence repositories. The pre-selected sequences were further filtered applyging different methods, some of which are extensively detailed in D2.3 “Set of 1,000 enzymes selected using motif screens”.

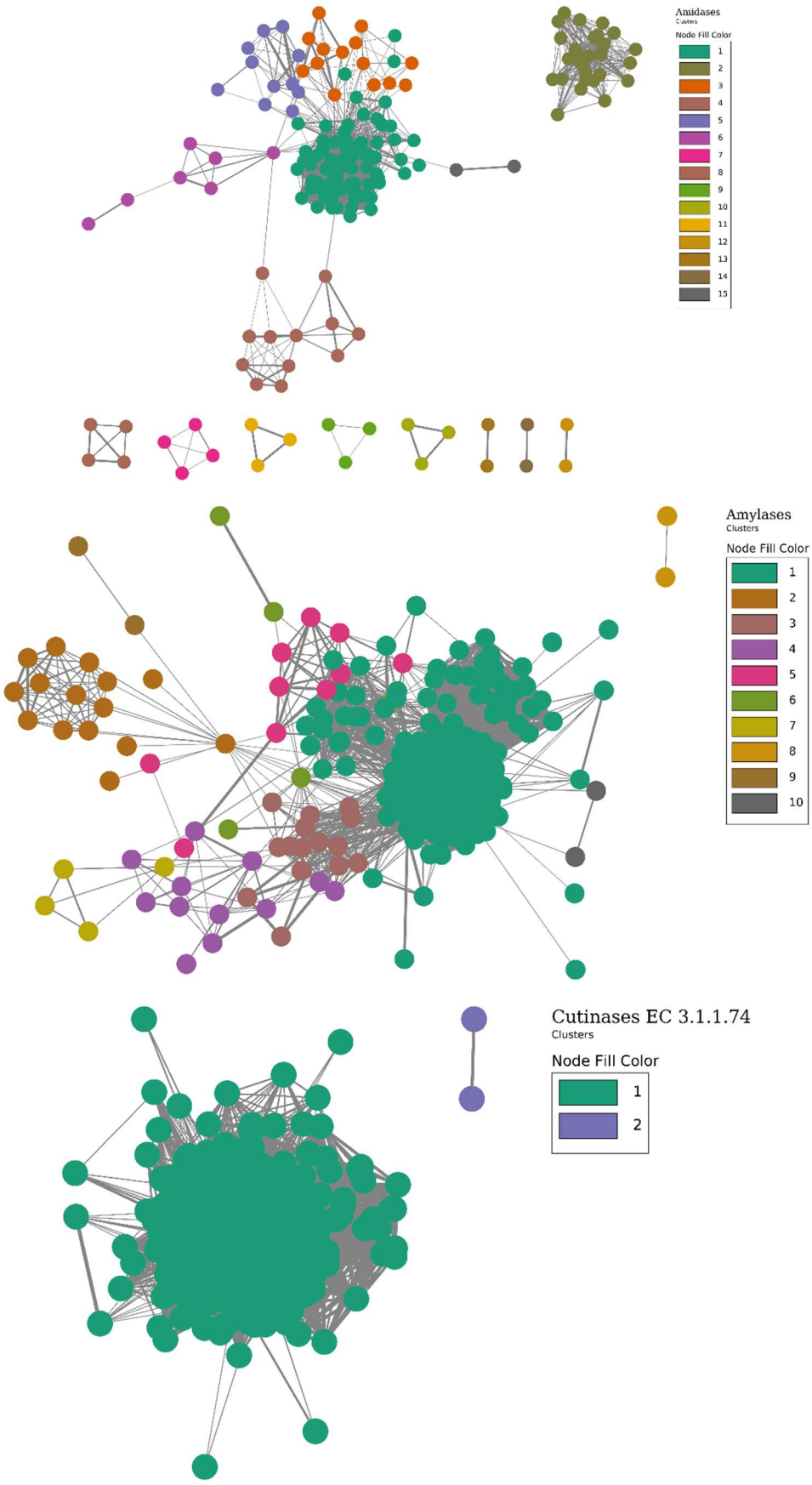
Annex

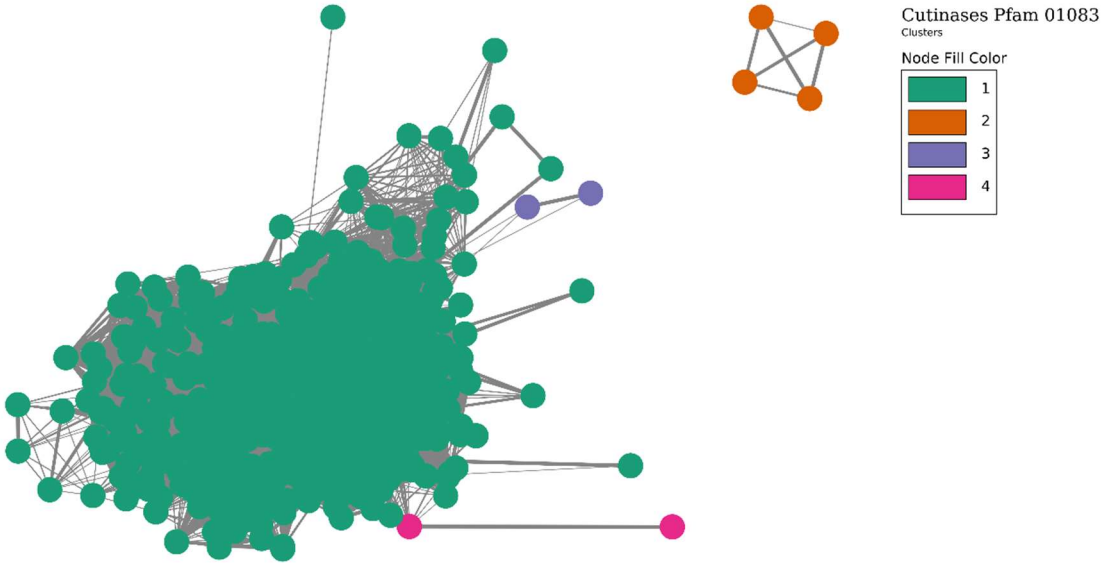
Because of their extensive size, the following Annex files are provided in a separate ZIP file:
See intranet’s project website File 1 (D2_2) in www.futurenzyme.eu -> login -> private-area -> shared-data.

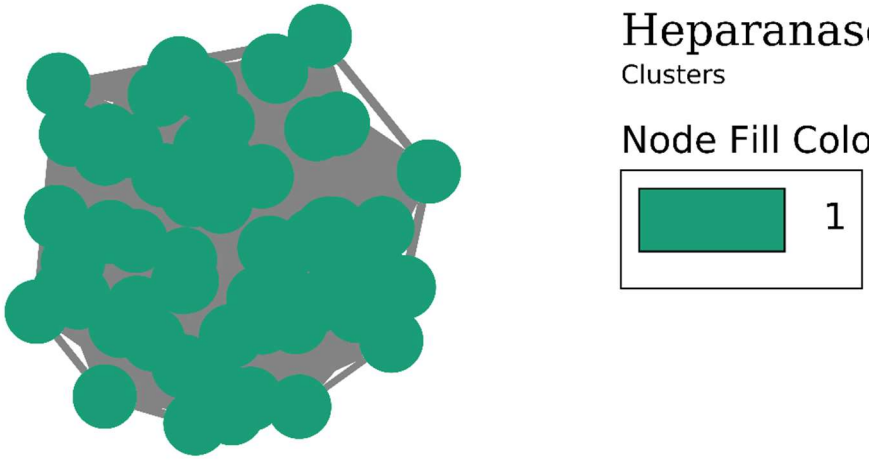
- **Annex File 1_FuturEnzyme Reference Sequences_to_do_BLAST**
  In-house database containing sequences encoding enzymes relevant to detergent, cosmetic and textile sectors. The sequences include those retrieved from bibliographic and patent search as well as one relevant sequence per taxonomic group.
- **Annex File 2_ DIAMOND BLASTP_Results**
  Sequences encoding enzymes potentially relevant to detergent, cosmetic and textile sectors obtained by DIAMOND BLASTP. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin.
- **Annex File 3_Network Analysis Enzymes**
  Sequences encoding enzymes constituting each of the networks identified per enzyme family. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin.
- Annex **File 4_ BLAST_Results**
  Sequences encoding enzymes potentially relevant to detergent, cosmetic and textile sectors obtained by BLAST. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin.

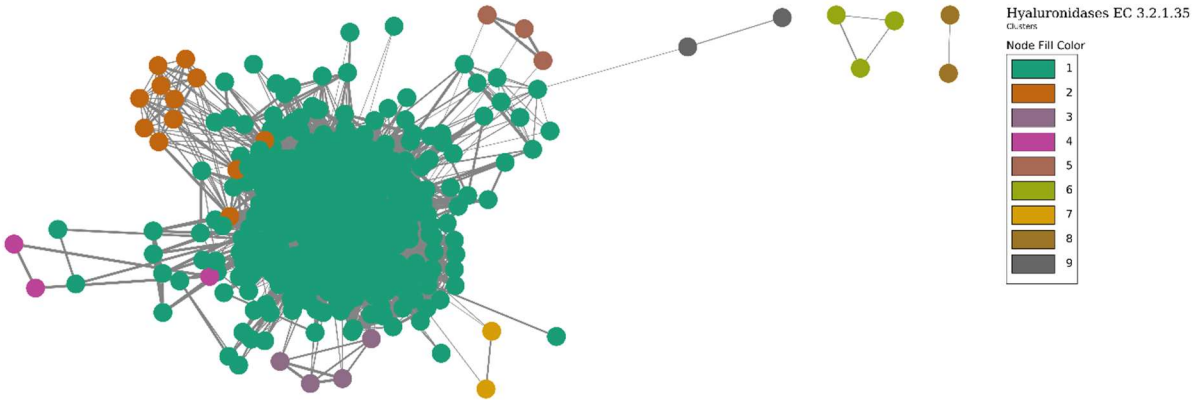
- Annex **File 5_HMMs_Results**
  Sequences encoding enzymes potentially relevant to textile and cosmetic sectors obtained by HMMs. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin.
- Annex **File 6_ EP-PRED_Results**
  Sequences encoding enzymes potentially relevant to detergent and textile sectors obtained by EP-PRED. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin.
- Annex **File 7_ PSI-BLAST and PELE_Results_Lip9**
  Lip9-homologous sequences pre-selected through PSI-BLAST and PELE. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin.
- Annex **File 8_ HMMs_Results_Lip9**
  Lip9-homologous sequences pre-selected through HMMs. The table contain information which include the reference sequence (and ID), the retrieved sequence (and ID), and the origin.

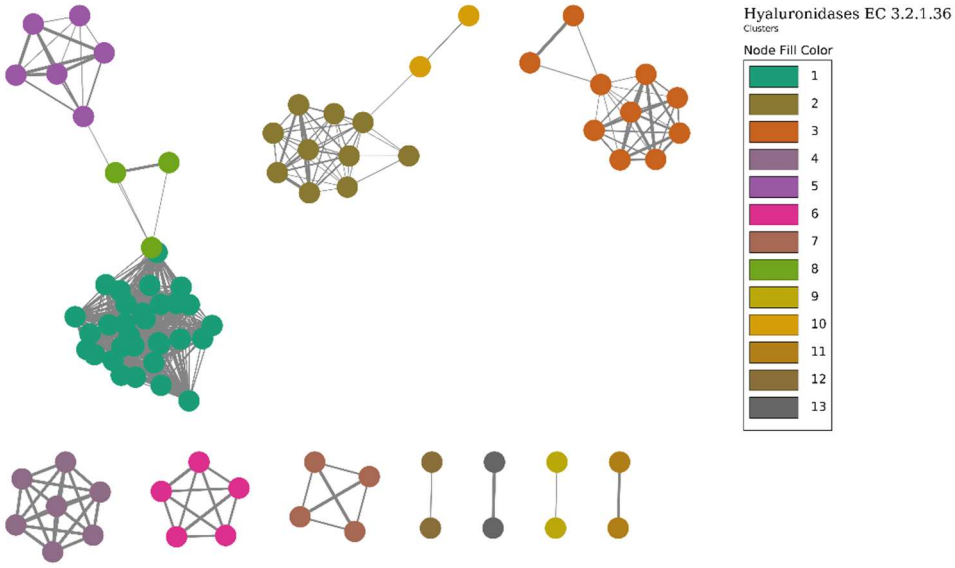
Annex **Figure 1**. Image representing the different clusters identified by MCL algorithm, within pre-selected enzymes retrieved through DIAMOND BLASTP.

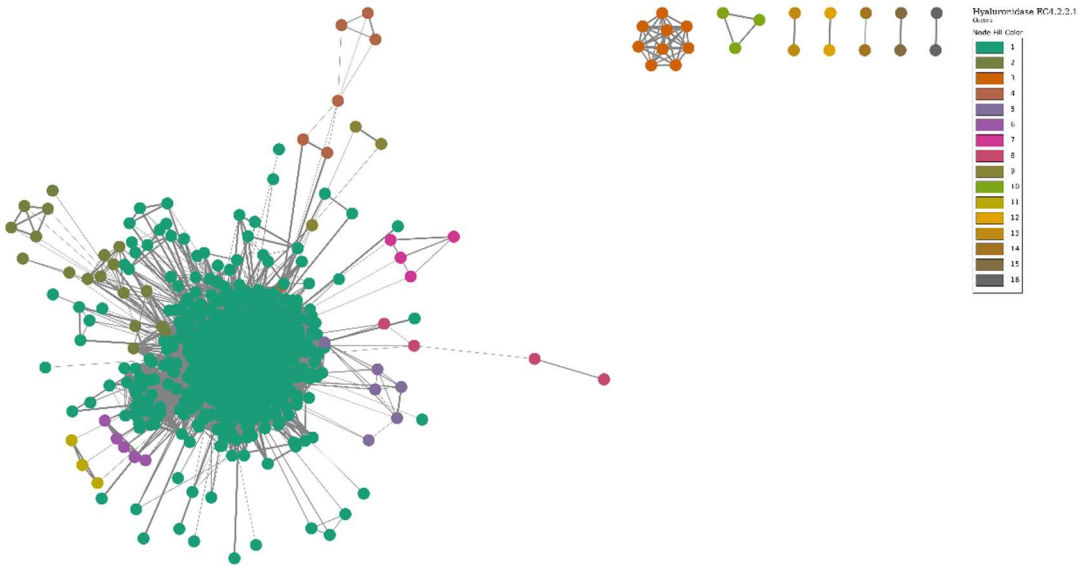


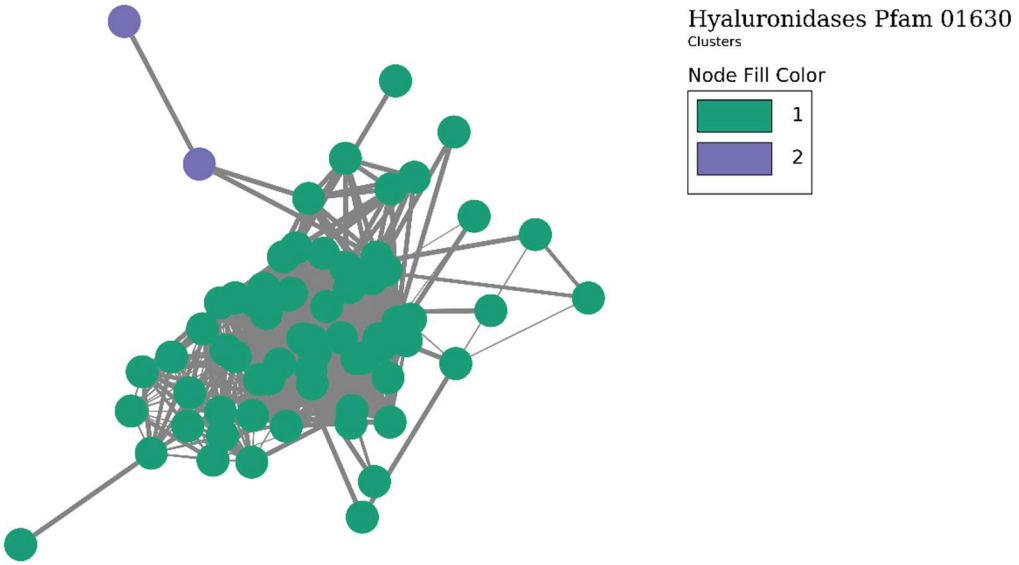

Cutinases Pfam 01083
Clusters
Node Fill Color
1
2
3
4

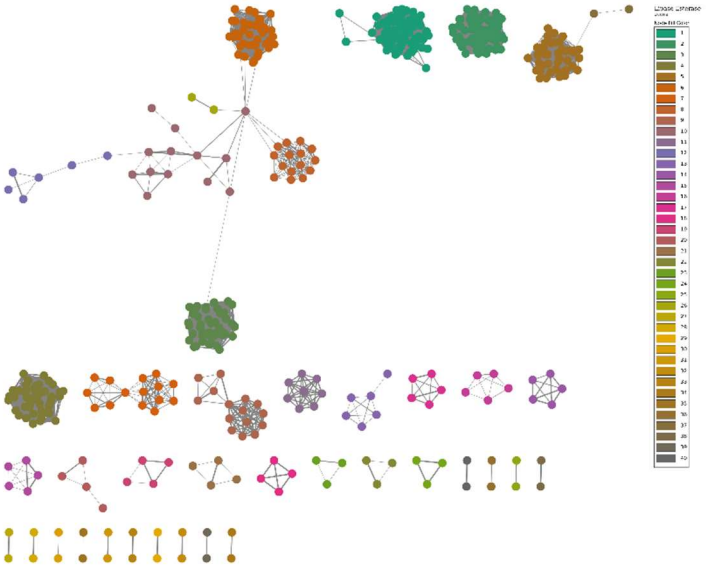

Heparanases
Clusters
Node Fill Color
1

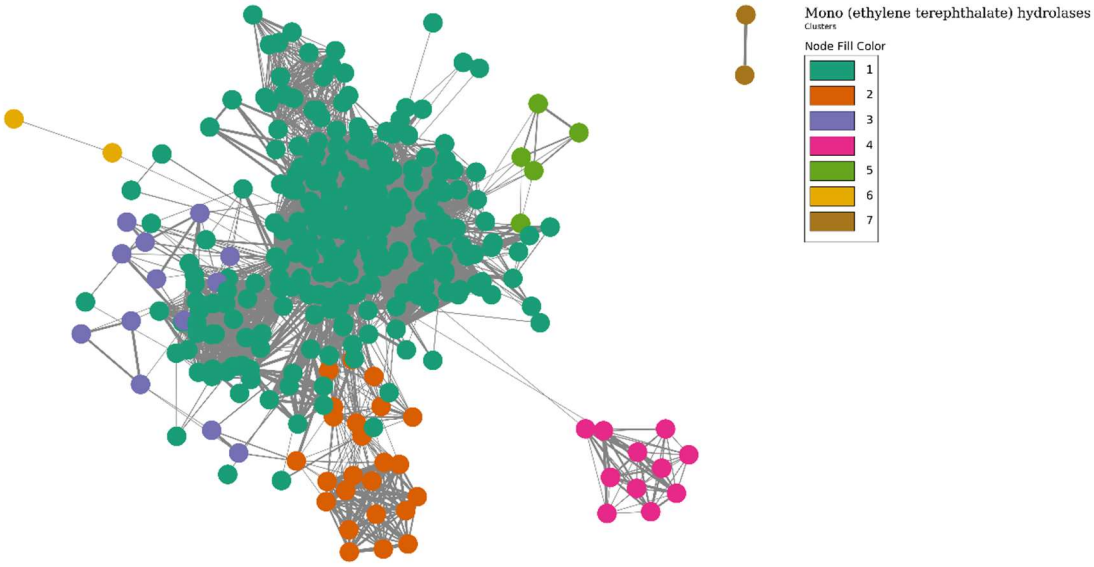

Hyaluronidases EC 3.2.1.35
Clusters
Node Fill Color
1
2
3
4
5
6
7
8
9

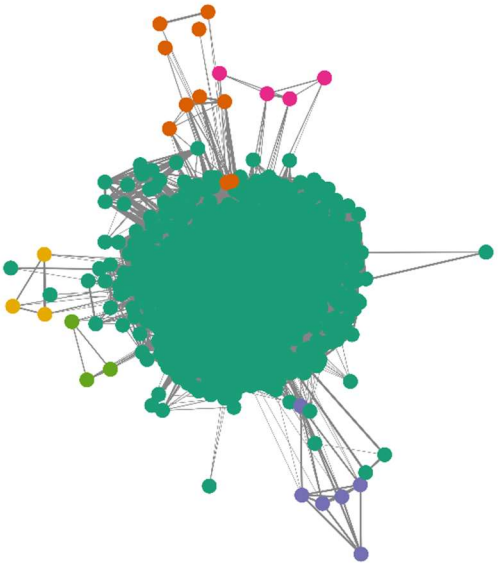

Hyaluronidases EC 3.2.1.36
Clusters
Node Fill Color
1
2
3
4
5
6
7
8
9
10
11
12
13

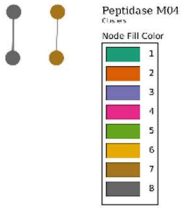

Hyaluronidase EC4.2.2.1
Clusters
Node Fill Color

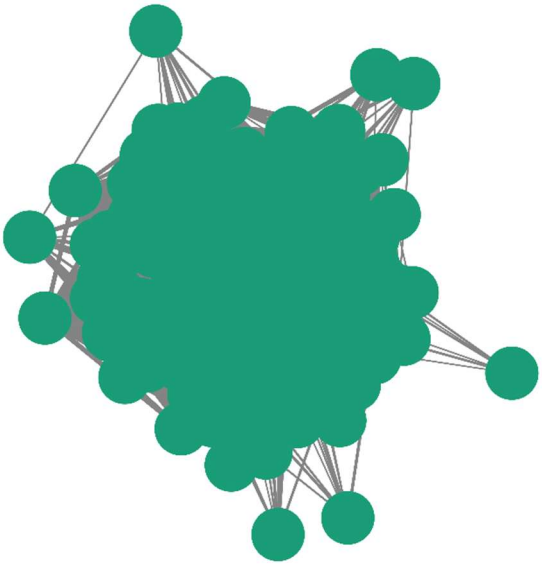

Hyaluronidases Pfam 01630
Clusters
Node Fill Color
1
2




Mono (ethylene terephthalate) hydrolases
Clusters
Node Fill Color
1
2
3
4
5
6
7





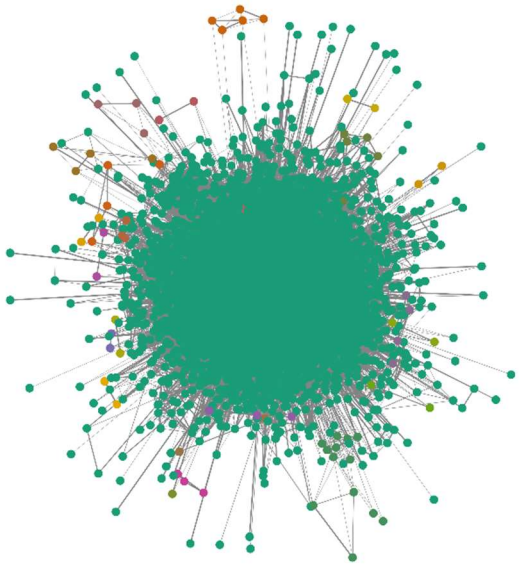


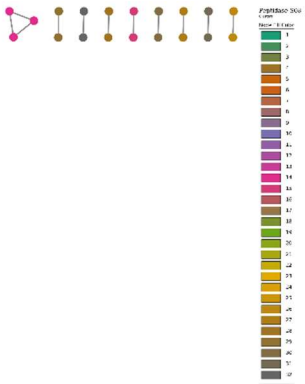
Peptidase type Papain

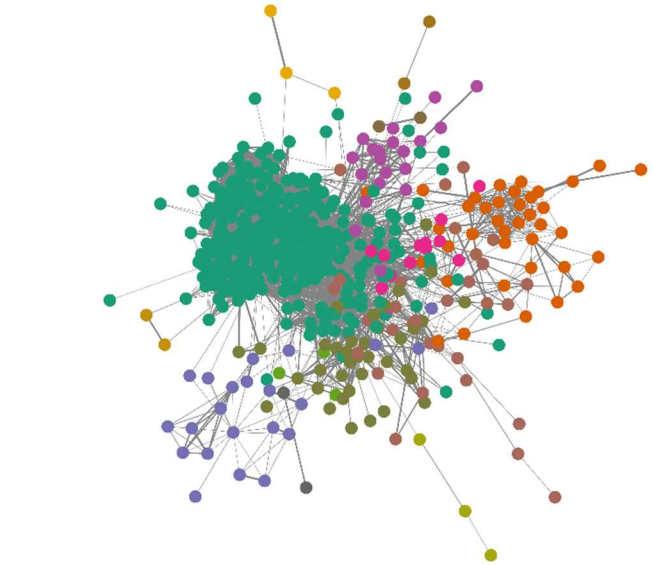
Clusters

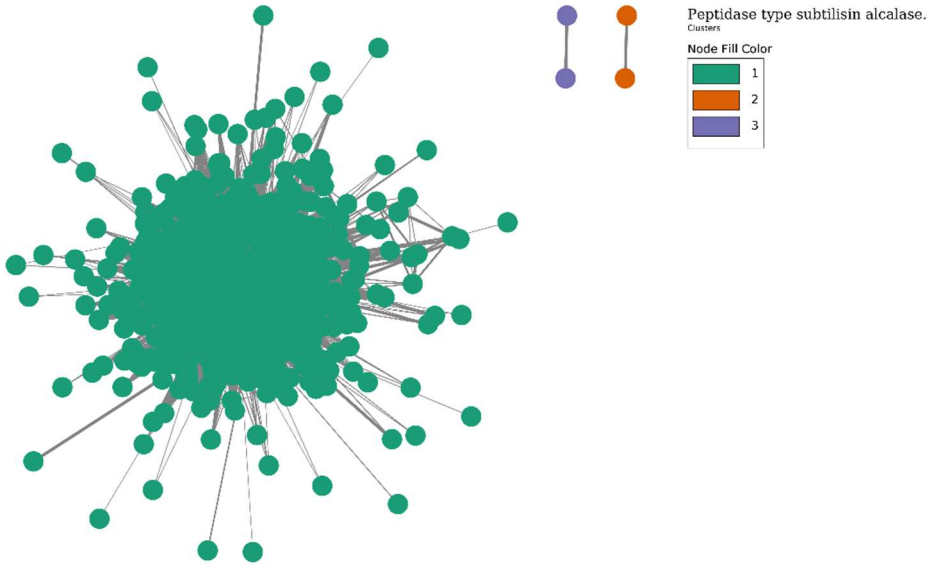
Node Fill Color

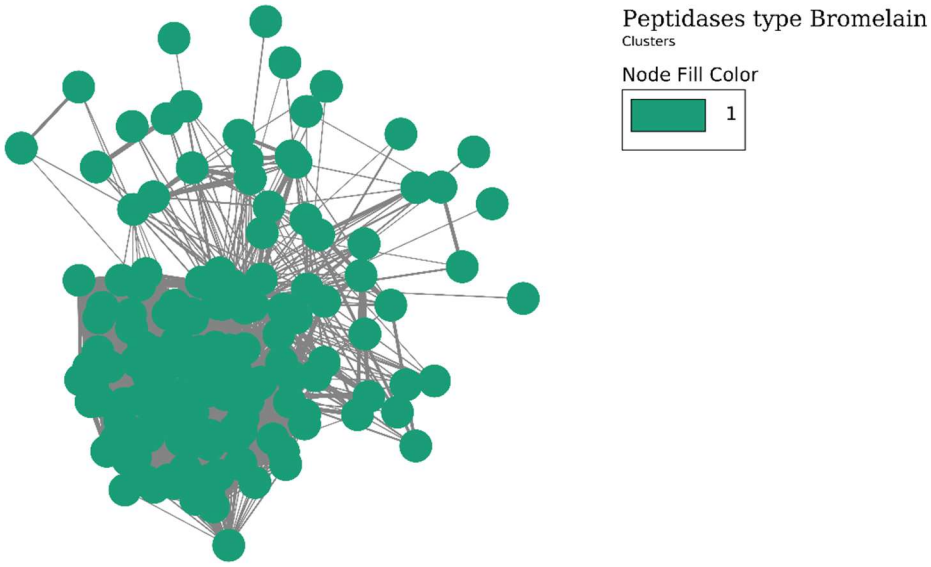
1

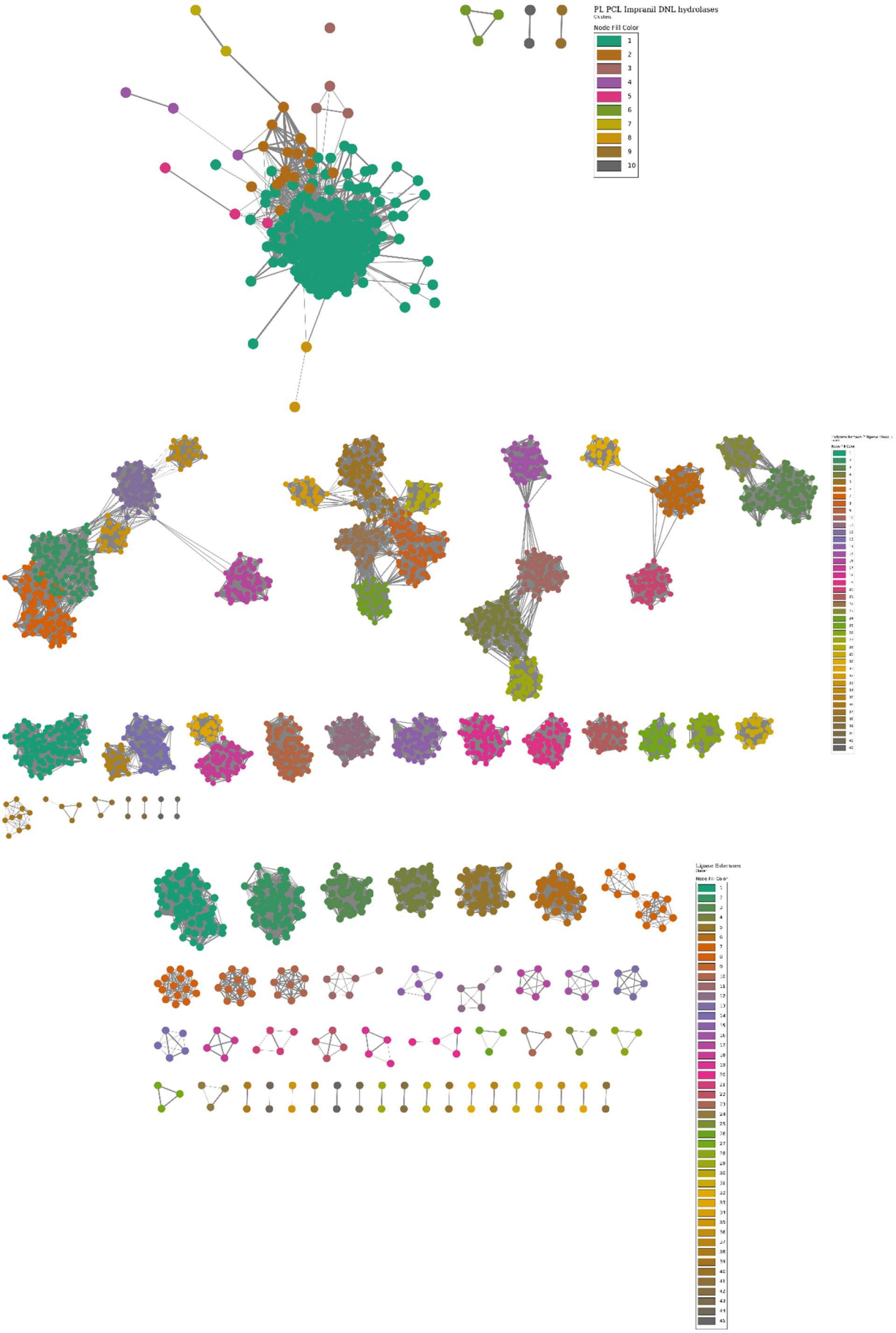





Peptidases type savinase esperase
Node Fill Color
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15


Peptidase type subtilisin alcalase.
Clusters
Node Fill Color
1
2
3


Peptidases type Bromelain
Clusters
Node Fill Color
1


PL PCL Impranil DNL hydrolases
Clusters
Node Fill Color
1
2
3
4
5
6
7
8
9
10