COMPARATIVE STUDY OF BACTERIAL AND FUNGAL ALPHA-AMYLASE INDUSTRIAL PRODUCERS

Maria Torrents Soler

Bioinformatics and Biostatistics Master

Microbiology, biotechnology and molecular biology

Paloma Pizarro Tobías

David Merino Arranz and Carles Ventura Royo

02/01/2019

Título:	Comparative study of bacterial and fungal alpha-amylase industrial producers.
Nombre del autor:	Maria Torrents Soler
Nombre del tutor:	Paloma Pizarro Tobías
PARA name:	Carles Ventura Royo y David Merino Arranz
Fecha de entrega (mm/aaaa):	2/01/2019
Titulación:	Máster en Bioinformática y Bioestadística
Área del Trabajo Final:	Microbiología, biotecnología y biología molecular
Idioma del trabajo:	Inglés
Palabras clave	*Bacillus licheniformis*, Alpha-amylase, *Aspargillus oryzae.*

Abstract

There are several industrial sectors that use enzymes from microorganisms to generate products for a wide range of applications. Among all industrial enzymes, alpha amylase is one of the main enzymes used in industry.

Furthermore, microbial enzymes have gained interest for their widespread uses in industries and medicine owing to their easy availability, stability, catalytic activity, ease of production and optimization than plant and animal enzymes, hence they are the favored sources for industrial enzymes production.

The extensive applications and the broad utility variety of microbial enzymes arises the interest of a comparative study of different alpha-amylase microbial producers along with a structure and sequence based analysis of alpha-amylase protein from *Bacillus licheniformis* and *Aspergillus oryzae*, likewise, they are the most utilized industrial bacterial and fungal alpha-amylase producers.

Phylogenetic analysis pretends to study the evolutive and structural relationship between some *Bacillus spp* alpha-amylase producers and *Aspergillus spp* alpha-amylase producers. Also, its goal is to perform a comparison between natural producers and the industrially ones, in addition to study the evolutionary relationships within the spices. On the basis of the establish evolutionary relationships *Bacillus paralicheniformis* and *Bacillus halotolerans* show potential as new alpha-amylase industrial producers.

Comparative genomic study between the coding alpha-amylase gene obtained from the respective protein and the microbial genome aims to provide a highly detailed view of the protein genetics.

Protein profile analysis results in a structural and functional study of the protein, together with an homology alpha-amylase *B.licheniformis* and *A.oryzae* model.

Resumen

Varios sectores industriales utilizan enzimas producidas por microorganismos para diversos productos con una amplia gama de aplicaciones. De todas las enzimas industrialmente comercializadas, la alpha-amilasa es una de las principales. Así mismo, los microorganismos han ganado interés tanto en la industria cómo en el campo de la medicina debido a sus amplias aplicaciones, siendo la fuente preferida de enzimas industriales. Además, poseen ciertas características ventajosas como la disponibilidad, estabilidad, actividad catalítica, fácil producción y optimización en comparación con enzimas procedentes de plantas y animales.

La infinita cantidad de aplicaciones e utilidades de las enzimas de origen microbiano genera el interés de un estudio comparativo de las diferentes alpha-amilasas industriales de origen microbiano, juntamente con un análisis estructural y en base a la secuencia de la proteína alpha-amilasa proveniente de la bacteria *Bacillus licheniformis* y el hongo *Aspergillus oryzae*. Así pues, son las especies más utilizadas industrialmente.

La finalidad de los analisis filogenéticos es el estudio de la relación evolutiva y estructural entre distintas especies pertenecientes a los productores de alpha amilasa del género *Bacillus* y *Aspergillus*. Mediante los análisis filogenéticos, también, se quiere realizar una comparación entre aquellos productores naturales e industriales de cada uno de los generos. En base a las relaciones evolutivas establecidas los microorganismos *Bacillus paralicheniformis* y *Bacillus halotolerans* presentan potencial cómo nuevos productores industriales.

El estudio genómico comparativo entre el gen codificante para la alpha-amilasa, obtenido de la respectiva proteína, y el genoma del microorganismo del productor proporciona una visión detallada de la genética de la proteína.

Por último, el perfil proteico resulta en un estudio estructural y funcional de la proteína conjuntamente con la construcción de un modelo por homología de la alpha-amilasa de *B.licheniformis e A.oryzae.*

Resum

Diversos sectors industrials utilitzen enzims produïts per microorganismes per a diversos productes amb una àmplia gamma d'aplicacions. De tots els enzims comercialitzats industrialment, l'alfa-amilasa és una de les principals. Tanmateix, els microorganismes han guanyat interès tant a la indústria com en el camp de la medicina, gràcies a la seva diversitat d'aplicacions i a més essent la font preferia dels enzims industrials. A més a més, presenten certes característiques avantatjoses com la disponibilitat, l'estabilitat, l'activitat catalítica, fàcil producció i optimització en comparació als enzims procedents de plantes i animals.

La infinita quantitat d'aplicacions i utilitats dels enzims d'origen microbià promouen l'interès d'un estudi comparatiu de les diferents alfa-amilases industrial d'origen microbià, juntament amb un anàlisis estructural i segons la seqüència de la proteïna alfa-amilasa provinent del bacteri *Bacillus licheniformis* i el fong *Aspergillus oryzae*. Així mateix, són les espècies més utilitzades en la producció industrial.

La finalitat dels anàlisis filogenètics és l'estudi de la relació evolutiva i estructural entre diferents espècies pertanyents als productors d'alfa-amilasa del gènere *Bacillus* i *Aspergillu*s. Mitjançant els anàlisis filogenètics, també, es pretén realitzar una comparació entre aquells productors naturals i industrials de cada un dels gèneres. D'acord amb les relacions evolutives establertes, els microorganismes *Bacillus paralicheniformis* i *Bacillus halotolerans* presenten potencial com a nous productores industrials.

L'estudi genòmic comparatiu entre el gen codificant per a l'alfa-amilasa, obtingut de la proteïna respectiva, i el genoma del microorganisme productor proporciona una visió detallada de la genètica de la proteïna.

Per últim, el perfil proteic resulta en un estudi estructural i funcional de la proteïna conjuntament amb la construcció d'un model per homologia de l'alfa-amilasa de *Bacillus licheniformis* i *Aspergillus oryzae.*

INDEX

FIGURE AND TABLE INDEX

1. INTRODUCTION

The use of enzymes from plants, animals and microorganisms to accomplish certain reactions dates back in ancient times before the function and nature was fully understood. The first reported commercial application of yeast was the production of alcoholic beverages from barley by the Babylonians and Sumerians as early as 6000 BC. However, the roots of modern enzymology are settled in the last century when the first enzyme was used for industrial purposes; the Danish chemist Christian Hansen produced rennet by extracting it from dried claves' stomachs with saline solution (Binod et al., 2013; Leisola, Jokela, Pastinen, Turunen, & Schoemaker, 2009).

Commercial enzyme production, nowadays, has grown in volume and number of products in response to expanding markets and increasing demand for novel biocatalyst. Enzymes are currently used in several different industrial products and processes; besides, new areas of application are constantly being added. They are used in various industrial processes, such as baking, brewing, detergents, fermented products, pharmaceuticals, textiles, leather processing. Indeed, over 500 industrial products are being made using enzymes (Adrio & Demain, 2014).

Enzymes made certain reactions commercially viable for the first time, as they catalyze reactions which are difficult to perform by chemical methods. In fact, the use of enzymes frequently results in many benefits that cannot be obtained with chemical treatment and they are regarded as environmentally friendly technology (Kirk, Borchert, & Fuglsang, 2002; Singh, Kumar, Mittal, & Mehta, 2016) .

Currently, industrial enzymes are commonly manufactured from microorganisms. They are the preferred source of industrial enzymes owing to their economic and technical advantages; reduced processing time, low energy input, cost effectiveness, fast growth rate, nontoxic and eco-friendly characteristics (J.Charnock Simon, 2005; Leisola et al., 2009). In addition, microbial enzymes are more active and stable than plant and animal enzymes (Anbu et al., 2013).

Natural selection, recombinant DNA technologies and classical improvement techniques are used to improve the microorganisms. Genetic modification of production microorganisms is practiced to enhance their productivity and adapt the microorganisms to industrial fermentation conditions. Because all of that, a limited number of microorganisms are considered to be appropriate producers. Such organisms must be recognized as safe (GRAS), well characterized, non-toxigenic and non-pathogenic (Hatti-kaul, 2009) .

The extensive applications and the broad utility variety of microbial enzymes, along with the DNA technology and protein engineering used to improve the production arises the interest for a comparative study between different microbial enzyme producers.

Among the many enzymes that are widely used in the industry, alpha-amylase is one of the most important. Alpha-amylases have potential application in a variety of industrial processes such as food, pharmaceutical, detergent, textile and paper (Saini, Singh Saini, Dahiya, & Harnek Singh Saini, 2017). Moreover, it has been in increasing demand due to its crucial role in the starch hydrolysis and the applications of this hydrolytic action.

Concretely, the aim of this study is to perform a comparative study of different alpha-amylase microbial producers. Resulting in a phylogenetic analysis of the microbial producers and the determination of the studied protein profile. Therefore, alpha-amylase from the bacteria *Bacillus licheniformis* and fungal *Apergillus oryzae* are studied.

1.1. Alpha amylase: General characteristics

Alpha amylases are widespread among living organisms, they are found in all forms of organism regardless of kingdom. They have been derived from several fungi, yeasts, bacteria and actinomycetes. Amylases from plants and microbe sources, have been employed for centuries in brewing. Even though, microbial enzymes have dominated applications in industrial sectors(de Souza & de Oliveira Magalhães, 2010). Moreover, amylases from fungal and bacterial sources meet the industrial demand due to their cost effectiveness, consistency, less time and space requirement for production and ease of process optimization and modification. Nowadays, the most widespread uses of these kind of enzymes are in the starch industry, where they are used for starch hydrolysis in the starch liquefaction process that converts starch into fructose and glucose syrups. (Saini et al., 2017; Sundarram & Murthy, 2014).

Alpha amylase (α-1, 4-glucan-glucanohydrolase, EC 3.2.1.1) is an extracellular enzyme which degrades starch and related products into small units such as glucose and dextrins; catalyzes the hydrolysis of internal alpha 1-4 glyosidic bonds and is also called glycoside hydrolases (Gopinath et al., 2017). Furthermore, the hydrolysate composition obtained after hydrolysis of starch is highly dependent on the effect of temperature, the conditions of hydrolysis and the origin of enzyme (Saini et al., 2017; Sundarram & Murthy, 2014).

In terms of biochemical alpha-amylase's remarkable proprieties are that they are biodegradable and work at milder conditions than chemical catalysts, in addition, they are environmentally safe enzymes. Alpha-amylases are metalloenzymes that require calcium ions to maintain their stability, activity, and structural conformation. The pH optima oscillate from 2 to 12, however, alpha-amylases from fungi and bacteria have their optima in the acidic to neutral range. At last, the temperature optimum for the activity of the enzyme is associated to the growth of the microorganisms (Gupta, Gigras, Mohapatra, Goswami, & Chauhan, 2003; Sundarram & Murthy, 2014).

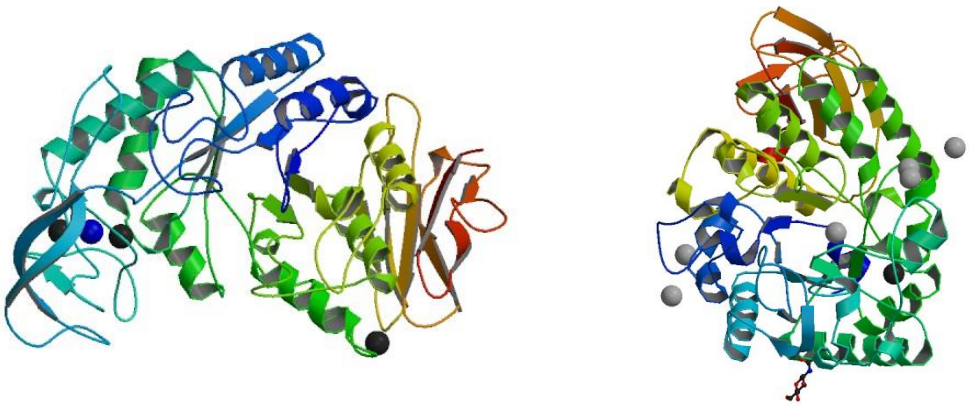


Figure 1. Left: Bacillus Licheniformis alpha-amylase (pdb:1BLI). Right: Aspergillus oryzae alpha-amylase (PDB: 3VX0)

1.2. Alpha amylase industrial producers

There are numerous alpha amylase producers but the most widely source among the bacterial species are the *Bacillus spp*; including *B. amyloliquefaciens* and *B. licheniformis*. In terms of fungal sources, alpha-amylases are mostly obtained from *Aspergillus* species and only from few species of *Penicillium*. *Aspergillus oryzae, A. niger and A. tamarie* are some of the predominant species for commercial production (Vengadaramana, 2014).

1.2.1. *Bacillus licheniformis*

Bacillus licheniformis is a gram-positive, spore-forming and part of the subtilis group. It is also a mesophilic bacterium, however, it can grow at high temperatures (Tmin 11°C, Topt 49°C and Tmax 57°C) with a minimum pH value at 4.6 and an optimal pH at 8.1 (Trunet et al., 2015). *B.licheniformis* is a facultative anaerobe and some isolates are capable of denitrification and it is commonly found in soil and feathers of ground dwelling birds (Ghani et al., 2013).

Figure 2. Bacillus licheniformis.

This microorganism is a causal agent of food poisoning and spoilage, although, it has proved itself as a multipurpose organism. *B.liheniformis* spices are used in the manufacture of industrial enzymes including several proteases, alpha- amylase, penicillinase, pentosanase, cycloglucosyltransferase, β-mannanase and several pectinolytic enzymes (Ghani et al., 2013). In addition, since 1972, it has been safely used for large-scale industrial fermentation to produce amylase (de Boer, Priest, & Diderichsen, 1994); which are ,currently, used for the hydrolysis of starch, desizing of textiles and sizing of paper.

1.2.2. *Aspergillus oryzae*

Aspergillus oryzae is a filamentous fungi with the secretory capacity to produce various and huge amounts of enzymes. This fungus is widely used in traditional Japanese fermentation industries, including soy sauce, sake, bean curd seasoning and vinegar production (Machida, Yamada, & Gomi, 2008). In addition, *Aspergillus oryzae* was used for the first example of commercial production of heterogonous enzyme, the lipase for laundry detergent in 1988.

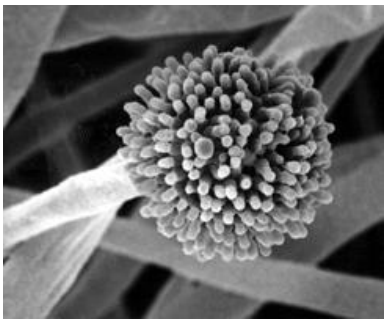


Figure 3. Aspergilus oryzae

2. BACKGROUND AND JUSTIFICATION

Enzyme technology offers a great potential for many industries thanks to their proprieties; the introduction of industrial enzymes results in significant savings in resources, such as energy and water for the benefit of both the industry ,in question, and the environment. Along these lines, enzyme field is constantly developing, therefore, there is always a need for improving as well as a need to find new approaches to enhance enzyme technology.

The project incentive arises from the personal interest on the white biotechnology, which is the branch applied to the industry and its processes. The application of biotechnology to industrial production, also, holds many promises for sustainable development. The investigation and innovation on this field aims to optimize industrial processes.

The whole project approach is more educational than a contribution of new knowledges to the scientific community and It aims to be the base for future complex analysis.

Phylogenetic analysis pretends to study the relationship between some bacterial and fungi enzyme producers such as various *Bacillus* and *Aspergillus* spices. It also arises an interest on the comparison between natural producers and the industrially ones. As well, the purpose of this study is to report the phylogenetic relationships and evolutionary history of the alpha-amylase proteins. In the other hand, the protein profile study intents to stablish the three-dimensional structure of the alpha-amylase proteins studied, to be afterwards compared and used for further studies.

The aim of this project is to study alpha-amylases produced by different bacterial or fungal spices and find new homologs in nature which has not been already exploited.

3. PLANNING

3.1 Tasks

Table 1 shows the several phases, explained at the section above. As we can see, the project will last a total of 19 weeks. The timing is based on the phases level of difficulty and the University established deadlines.

Table 1. Project Chronogram

PHASE NUMBER	PHASE DESCRIPTION	DURATION (WEEKS)
A	**Preparatory actions**	
A.1	Paper contents definition	2
A.2	Project plan	2
A.3	Select the study target	2
B	**Project development 1**	
B.1	Establish the template sequence	1
B.3	Run a BLAST/PBLAST	0.5
B.4	Complete the database	0.5
B.5	Phylogenetic analysis	2
C	**Project development 2**	
C.1	Protein Profile	2
C.2	Comparative genomic study	3 days
C.3	Phylogenetic analyses of alpha amylase family	1
C.4	Tree reliability	2 days
C.5	Results	1.5
D	**Report**	
D.2	Results discussion	1.5
D.3	Conclusions	2 days
D.4	Glossary, bibliography and annex	2 days
E	**Presentation**	
E.1	Presentation preparation	2
E.2	Thesis defense	1

3.2 Calendar

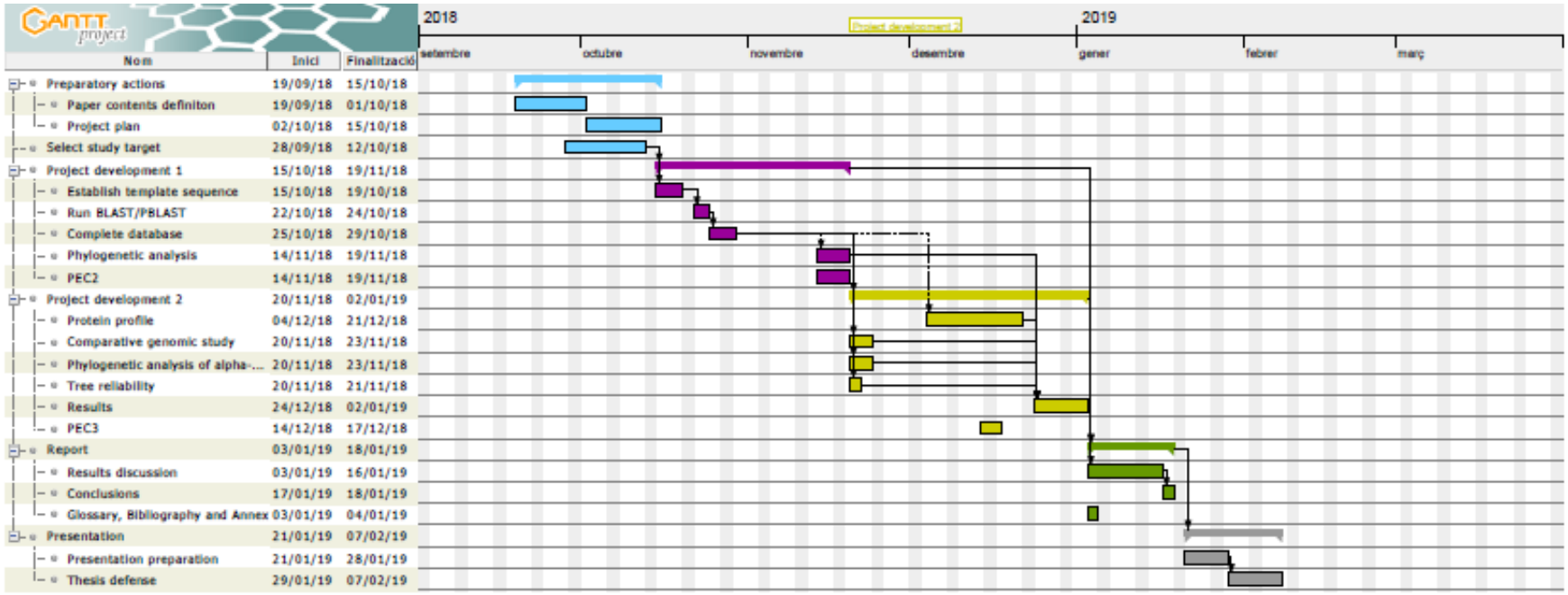


Figure 4 Gantt diagram of the project.

3.3 Highlights

The critical path determines the key points on the project development. It is the longest sequence of activities in a project plan which must be completed on time for the project to complete on due date. The figure 5 shows the PERT chart where the critical route is represented in red.

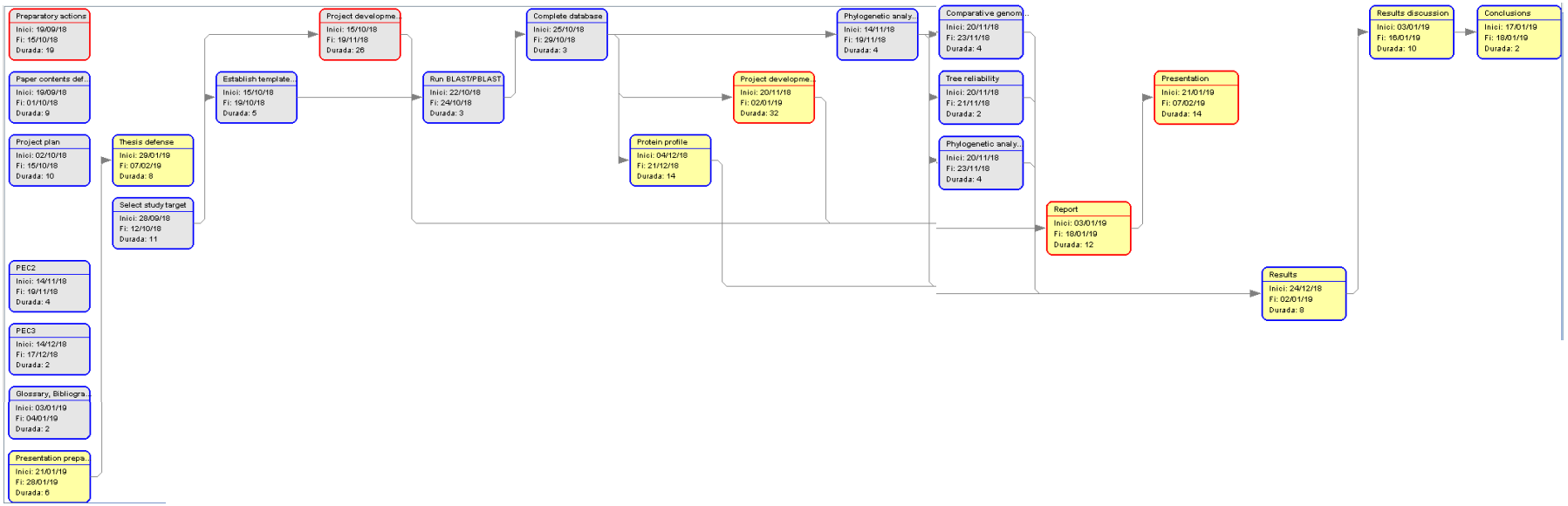


Figure 5. PERT chart for the project

3.4. Brief description of the other report chapters

Objectives

The principal and secondary objectives of the project are exposed. The criteria used to establish them is that they must be accepting goals for the student knowledge in combination with the available resources e.g. public data bases, computational computer capacity, open access programs.

Methodology

It is divided in three main sections (phylogenetic analysis, comparative genomics and protein profile) and contains all the tools and the steps followed to compute the analysis. Its goal is to be reproducible for anyone and to make the readers understand the certain decisions that have been made to perform the different project parts.

Results

It contains the different results in a figure format mostly, as well they are commented and explained. They are divided in sections according to the main organism and the analysis performed. They reflect all the methodology followed to obtain the certain analysis in a unique or multiple figure together with an explanation.

Results discussion

Results from the section below have been discussed and supported by the literature. Meaning that *Bacillus licheniformis* and *Aspergillus oryzae* different analysis results have been compared together arguing the differences and similarities within and between the spices. The discussion a part form describing the results in detailed from a critical point of view, sums up the previously results underlying the most important aspects.

Conclusions

Project conclusions according to the objectives proposed at the very beginning and a critical reflection about their accomplishment. As well as, a critical analysis of the planning accomplishment. Also, some future studies are suggested.

4. OBJECTIVES

This study presents two differentiated main objectives:

- Establish the most important microorganisms used in the production of alpha-amylase.
    - Perform phylogenetic analyses to study evolutionary alpha-amylase's relationships from microbial producers.
    - Compute phylogenetic analyses to study the comparison between natural and industrial producers.
- Determine potential homologous proteins, alpha-amylase, as new study targets.
    - Find new homologues proteins that has not already been used as industrial enzymes. These new potential proteins might have potential to optimize some industrial process.

5. APPROACH AND METHODOLOGY

The study has been divided in three main parts; phylogenetic studies, a comparison at a genomic level of the target sequences and alpha-amylase protein profile analysis.

Phylogenetic analyses examine evolutionary protein's relationships, changes occurred in molecular sequences are evaluated. Phylogenetic reconstruction can ascertain the evolutionary relationships within members of a protein family, which evolved independently after speciation and duplication events. Concretely, this project examines a bacterial and fungal alpha-amylases protein sequences both from important industrial producers. Different approaches, Neighboor Joining and Maximum likelihood, for phylogenetic inference are being used.

Comparative genomics studies the relationship between the genome structure and function through different spices. The major principles of the comparative genomics fields are that a sequence that stays conserved across multiple spices is likely to be preserved due to evolutionary pressures. More precisely, sequences responsible for biological functions are similar from the last common ancestor to the contemporary ones. Likewise, elements responsible for the differences between spices should be divergent. Finally, elements which are not important for organisms' evolutive success will not be conserved.

Finally, the protein profile is performed. On one hand, *B.licheniformis* and *A.oryzae* proteins are characterized from an structural and functional point of view. On the other hand, structure is predicted by homology modelling. The goal of protein modeling is to predict a structure from its sequence with an accuracy that is comparable to the best results achieved experimentally (Krieger, Nabuurs, & Vriend, 2003). Indeed, the resulting structure contains enough information about the spatial arrangement of important residues in the protein and may guide the design of future new experiments.

5.1. Phylogenetic analysis of microbial alpha-amylase industrial producers

5.1.1. Study target selected

Reading several articles and looking up different internet sources, the industrial enzyme alpha-amylase has been established as the study target. As said in the introduction (section 1) it's one of the most important enzymes in the industry.

5.1.2. Study sequences established

Reading and comparing several articles bacterial *Bacillus lincheniformis* and fungal *Aspergillus oryaze* have been selected as the microbial industrial producers. The bibliography used for selecting the microorganisms is as follows:

> Gupta, R., Gigras, P., Mohapatra, H., Goswami, V. K., & Chauhan, B. (2003). Microbial α-amylases: A biotechnological perspective. Process Biochemistry, 38(11), 1599–1616. https://doi.org/10.1016/S0032-9592(03)00053-0
>
> de Souza, P. M., & de Oliveira Magalhães, P. (2010). Application of microbial α-amylase in industry - A review. Brazilian Journal of Microbiology : [Publication of the Brazilian Society for Microbiology], 41(4), 850–861. https://doi.org/10.1590/S1517-83822010000400004
>
> Sundarram, A., & Murthy, T. P. K. (2014). α-Amylase Production and Applications: A Review. Journal of Applied & Environmental Microbiology, 2(4), 166–175. https://doi.org/10.12691/JAEM-2-4-10

Once the microorganisms have been chosen, the protein sequences were search and retrieve from The National Center for Biotechnology Information (NCBI) database, therefore, accession numbers are WP_017474613.1 for *B.lincheniformis* alpha amylase and CAA31220.1 for *A.oryzae* alpha-amylase.

5.1.3. Sequence similarity search: PSI-BLAST (Position-Specific Iterated BLAST)

PSI-BLAST from the NCBI browser is used to search protein sequences and complete the datasets. PSI-BLAST is a protein sequence profile search method that predicts distant evolutionary relationships. It is a sort of machine learning algorithm which uses the results of the first alignment (PSSM) to score the next iteration of alignment (Bhagwat & Aravind, 2007; Jones & Swindells, 2002).

Before the PSI-BLAST is run the parameters are set up. The treshreshold is changed for a value of 0.0001. The threshold is a cutoff for what is considered homologous and what actually goes to the second step of the PSI-BLAST. Proteins sequences are obtained from the Reference Sequence (RefSeq) database, which provides a non-redundant collection of sequences as well as high quality sequences (Pruitt, Tatusova, & Maglott, 2004).

5.1.4. Multiple sequence alignment

The alignment of multiple sequences constitutes a crucial step for the study of the biological relationship because phylogenies will be inferred from the multiple sequence alignment itself. It is assumed that all positions in a column of a multiple sequence alignment derive from common ancestral residue (Pais, Ruy, Oliveira, & Coimbra, 2014; Sutton, 2008).

Sequences have been aligned by MUSCLE. MUSCLE is an iterative, matrix-based algorithm which accuracy and speed are consistently better than currently available programs. Likewise, distance matrices in MUSCLE are clustered using UPGMA because it gives a better benchmark scores and marginally improved results than neighbor-joining. Although neighbor joining is expected to give a more reliable estimate evolutionary tree, in progressive alignment is been assumed that the best accuracy is obtained at each node by aligning the two profiles that have fewest difference, even if they are not evolutionary neighbors (Edgar, 2004).

5.1.5. Phylogenetic trees

There are many methods to reconstruct phylogenetic relationships with molecular data. Nevertheless, they can be classified in two main categories which are distance-based methods and character-based methods.

Distance-based methods infer phylogenetic trees from the degrees of differences between pairs of sequences. Likewise, they assume a molecular clock meaning that sequences evolve at a rate that is relatively constant over time and among different organisms (Kumar, 2005). This category includes the Neighbor Joining (NJ) and the Unweighted Pair Group Method with Arithmetic (UPGMA) clustering algorithms, among others. For this paper, NJ has been chosen over UPGMA, because UPGMA is based on the vary rare condition that the molecular clock is perfect. Meaning that all distances contribute equally to each average that is computed (Bromham & Penny, 2003).

Character-based methods, in the other hand, use directly the individual columns of aligned sequences to infer phylogenetic trees, they are based on discrete characters from molecular sequences from individual taxa (Peng, 2007.). Also, that methods assume that each character substitution is independent of its neighbors. The mostly wide used methods are Maximum Parsimony (MP) and Maximum Likelihood (ML).

Neighbor Joining (NJ)

Neighbor Joining (NJ) is a method for reconstructing phylogenies from a set of distances between each pair of sequences by successive clustering (Howe, Bateman, & Durbin, 2002).The result is a unique final unrooted tree under the principle of minimum evolution. Its characteristics are (Orr, n.d.; Peng, 2007)

- Start with all taxa in a single node and decompose with each iteration. In other words, the first step consists in building a tree with a single internal branch where one node is linked to two neighbors and the other is linked to the rest, all possible neighbor pairs are considered and the tree with the smallest total branch length is selected.
- Pair of nodes pulled out (grouped) at each iteration are chosen so that the total length of the branches on the tree is minimized. The distance matrix is update at each iteration.
- The algorithm does not assume that the divergence of the sequences occurs at the same constant rate at all points in the tree. As well, allowing unequal rates of evolution in different branches of the tree. Moreover, the branch lengths are estimated by least squares.

Maximum Likelihood (ML)

ML is a statistical method that estimates the unknown parameters of a probability method. It aims to find the tree that, given our evolutionary model, results in the highest likelihood of obtaining the data we observe (Paradis, 2012). Its characteristics are (FILOGENIAS

MOLECULARES; “Inferring Phylogeny using Maximum Likelihood in R (phangorn) - AnthroTree - DukeWiki,”):

- ML examines characters at every single site of the multiple sequence alignment to assess the reliability of each position on the basis of all other positions.
- Likelihood provides probabilities of the sequences given a model of their evolution on a particular tree. The more probable the sequences given the tree, the more the tree is preferred.
- Amino-acid replacement matrices are an essential basis of protein phylogenetics. They are used to compute substitution probabilities along phylogeny branches, and thus the likelihood of the data.
- It has the statistical property of consistency; therefore, the resulting tree is informative, and the method gives high confidence scores.

Regarding the evolutionary models of evolution there are several methods to estimate the matrix form the protein alignments, such as Dayhoff, JTT, WAG, LG among others. Besides, common extensions to these models include parameters for a proportion of invariable sites (I) and for gamma-distributed rate heterogeneity among sites (Γ). A simultaneously comparation of the different methods mentioned above have been computed through model testing.

Model testing approach used to select the method are the Akaike information criteria (AIC) and Bayes information criterion (BIC). The two criteria are very similar in form but arise from very different assumptions, nevertheless, they hold the same interpretation in terms of model comparison (Luo et al., 2010). That is, the larger difference in either AIC or BIC indicates stronger evidence for one model over the other so in both criteria the model with the lowest value is picked.

Bootstrap

In phylogenetics is one of the most commonly used tests for the assessment of the reliability of an inferred tree. Bootstrap determines whether the topology of the tree is accurate or whether a better tree can be obtained. Felsenstein (1985) formally proposed bootstrapping as a method for obtaining confidence limits on phylogenies. Moreover, it is based on Efron’s original bootstrap technique of resampling one’s own data to infer the variability of the estimate (Soltis & Soltis, 2003).

Bootstrapping is a computationally performed statistical analysis which relies on random sampling with replacement. In other words, the method consists in resampling analysis that involves taking columns of characters out of your analysis, rebuilding the tree, and testing if the same nodes are recovered (Efron, Halloran, & Holmes, 1996). The method works as follows (Filogen, 2007):

1. Once the tree is built, the sequences are repeatedly resampled until a matrix with the same number of characters than the original is obtained.
2. Consensus tree is performed. It summarizes the topologies information recovered from each bootstrap pseudocopy.
3. Map the bootstrap proportions on the original topology.

As general rule, bootstrap values of 95% or greater are considered statically significant and indicates “support” for a clade, alternative nodes can be rejected if they occur in less than 5% of the bootstrap estimates.

5.3. Comparative genomics

A comparative genome study between the two genomes from microbial alpha-amylase industrial producers and the target sequences is computed. The aim is to establish gene relationships and differences between the bacterial and fungal spices genome, by comparing organizations, functional confirmation and evolutive implications. Therefore, BLAT is used to identify the coding exon locations at the microorganisms' genome. Actually, a TBLASTN is performed due to the query sequences nature, alpha-amylase from *B.licheniformis* and *A.oryzae*. TBLASTN is a mode of operation for BLAST that aligns protein sequences to a nucleotide database translated in all six frames (Bhagwat, Young, & Robison, 2012).

Biosequence analysis using profile hidden Markov models (HMMER) associated tool is also used. HMMER is a software, suit for protein sequence analysis using probabilistic methods and includes four database search; phmmer, hmmscan, hmmsearch and jackhammer (Finn, Clements, & Eddy, 2011) .Concretely, phmmer algorithm is the one incorporated in the ensembl platform and it is analogous to BLASTP; the protein query sequence is searched against the existing genomes at the data base.

5.4. Protein profile

5.4.1. Structural and functional characterization

SOPMA server is used to predict the secondary structure, Scanprostite from Expasy is used to predict known motifs in the sequence and Conserved Domains Database (CDD) is used for domain analysis. CDD is a protein annotation resource that consists of a collection of well-annotated multiple sequence alignment models for ancient domains and full-length proteins. As well, CDD groups proteins that have strong sequence similarity to protein domain fingerprints ("Conserved Domains Database (CDD) and Resources,"; "Using Conserved Domains to Find Protein Homologs | NCBI Insights,")

5.4.2. Homology modeling

Homology modelling is a general tool in which an unkwown structure for a protein is generated using a known structure of a homologous protein as a template. This technique relies on the idea that the tertiary structure of a protein it is better conserved that its amino acid sequence. During evolution, the structure is more stable and changes much slower than the associated sequence, so that similar sequences adopt practically identical structures, and distantly related sequences still fold into similar structures (Krieger et al., 2003). On top, Swiss Model sever will be used to model the proteins.

Swiss Model server will be used to model the protein. In practice, homology modelling is a multistep process that can be summarized in the following steps (Bordoli et al., 2009):

1. Template identification
2. Amino acid sequence alignment
3. Alignment correction
4. Backbone generation
5. Generation of loops
6. Side chain generation and model optimization
7. Model validation

As a general rule the similarity between the model and the template sequence needs to be high. A percentage sequence identity above 50% will mean a relatively strait forward modeling project, while anything below that will require careful planning.

6. RESULTS

6.1. *Bacillus licheniformis* alpha-amylase results

B. licheniformis query sequence used for the analysis is:

>WP_017474613.1 alpha-amylase [*Bacillus licheniformis*]

MKQQKRLYARLLTLLFALIFLLPHSAAAAANLNGTLMQYFEWYMPNDGQHWKRLQNDSAYLAEHGITAVW

IPPAYKGTSQADVGYGAYDLYDLGEFHQKGTVRTKYGTKGELQSAIKSLHSRDINVYGDVVINHKGGADA

TEDVTAVEVDPADRNRVISGEHLIKAWTHFHFPGRGSTYSDFKWHWYHFDGTDWDESRKLNRIYKFQGKA

WDWEVSNENGNYDYLMYADIDYDHPDVAAEIKRWGTWYANELQLDGFRLDAVKHIKFSFLRDWVNHVREK

TGKEMFTVAEYWQNDLGALENYLNKTNFNHSVFDVPLHYQFHAASTQGGGYDMRKLLNGTVVSKHPLKSV

TFVDNHDTQPGQSLESTVQTWFKPLAYAFILTRESGYPQVFYGDMYGTKGDSQREIPALKHKIEPILKAR

KQYAYGAQHDYFDHHDIVGWTREGDSSVANSGLAALITDGPGGAKRMYVGRQNAGETWHDITGNRSEPVV

INSEGWGEFHVNGGSVSIYVQR

The sequence is composed of 512 amino acids and belongs to the region PRK09441, which encodes a cytoplasmatic alpha-amylase and the sequence may be annotated on many different RefSeq genomes from the same, or different, species. Additionally, the domains encountered in the protein are three; active site, catalytic site and Na/Ca binding site (figure 6).

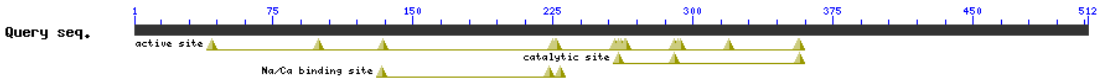


Figure 6. Conserved domains from B. licheniformis alpha-amylase WP_017474613.1.

6.1.1. Phylogenetic analysis of *B. licheniformis* alpha amylase

PSI-BLAST

PSI-BLAST results are shown below (figure 7), they all have high identity percentages and query covers values as well as significative e-value scores. The levels achieve for the three parameters are 69%-100%, 95%-100% and practically 0, respectively. Due to the resulting values, only the query cover parameter has been taken in account to build the dataset, sequence with at least 98% of coverage are retrieved.

Figure 7. PSI-BALST second iteration results for the WP_017474613.1 sequence. At the right top hits are shown.

B.licheniformis alpha-amylase data set is composed of 495 sequences of basically different *Bacillus* species. The whole data set can be found at the annex section.

Multiple sequence alignment

The resulting multiple sequence alignment is graphically represented in figure 8 , where the logos are a graphical representation of the information content stored in a multiple sequence alignment. Thereby, enriched amino acids at each peptide position are represented in the positive y-axis and depleted amino acids are represented in the negative y-axis.

Regarding the consensus sequence obtained from the multiple sequence alignment, it is as follows:

```
---------------M?KRIT----?VGLSVVLFLPSIY?GSKAYA-
DTVNNGTLMQYFEWYAPNDGNHWNRLR?DAENLAQKGITSVWIPPAYKGTTQNDVGYGAYDLYDLGEFNQKGTVRTKY
GTKAQLKSAI?ALHKQNIDVYGDVVMNHKGGADYTETVTAVEVDPNNRNIEVSGDYEISAWTGFNFPGRGD?YSNFKW
KWYHFDGTDWDEGRKLNRIYKFRGIGKAWDWEVSSENGNYDYLMYADLDFDHPDVANEMK?WGTWYANELNLDGFRLD
AVKHIDHEYLRDWVNHVRQQTGKEMFTVAEYWQNDIQTLNNYLAKVNYNQSVFDAPLHYNFHYASTGNGNYDMRNILN
GTVV?NHP?LAVTLVENHDSQPGQSLESVVSPWFKPLAYAFILTRAEGYPSVFYGDYYGTKGNSSYEIPALKDKIDPI
LTARKNFAYGTQRDY?DHPDVIGWTREGDSVHANSGLATLISDGPGGSKWMDVGKNNAGEVWYDITGNQTNTVTINKD
GWGQFHVSGGSVSIYVQQ-
```

Compering the consensus sequence composition and the enriched amino acids from the figure 8, more or less, they are the same. In fact, mostly positions the amino acids characters of the consensus sequences are the same as the enriched amino acids. However, some of the positions in the consensus and the logos differ; in the initial positions of the consensus sequence there are 15 gaps and some no consensus positions along the sequence. In contrast, in the figure 8 there are certain amino acids enriched

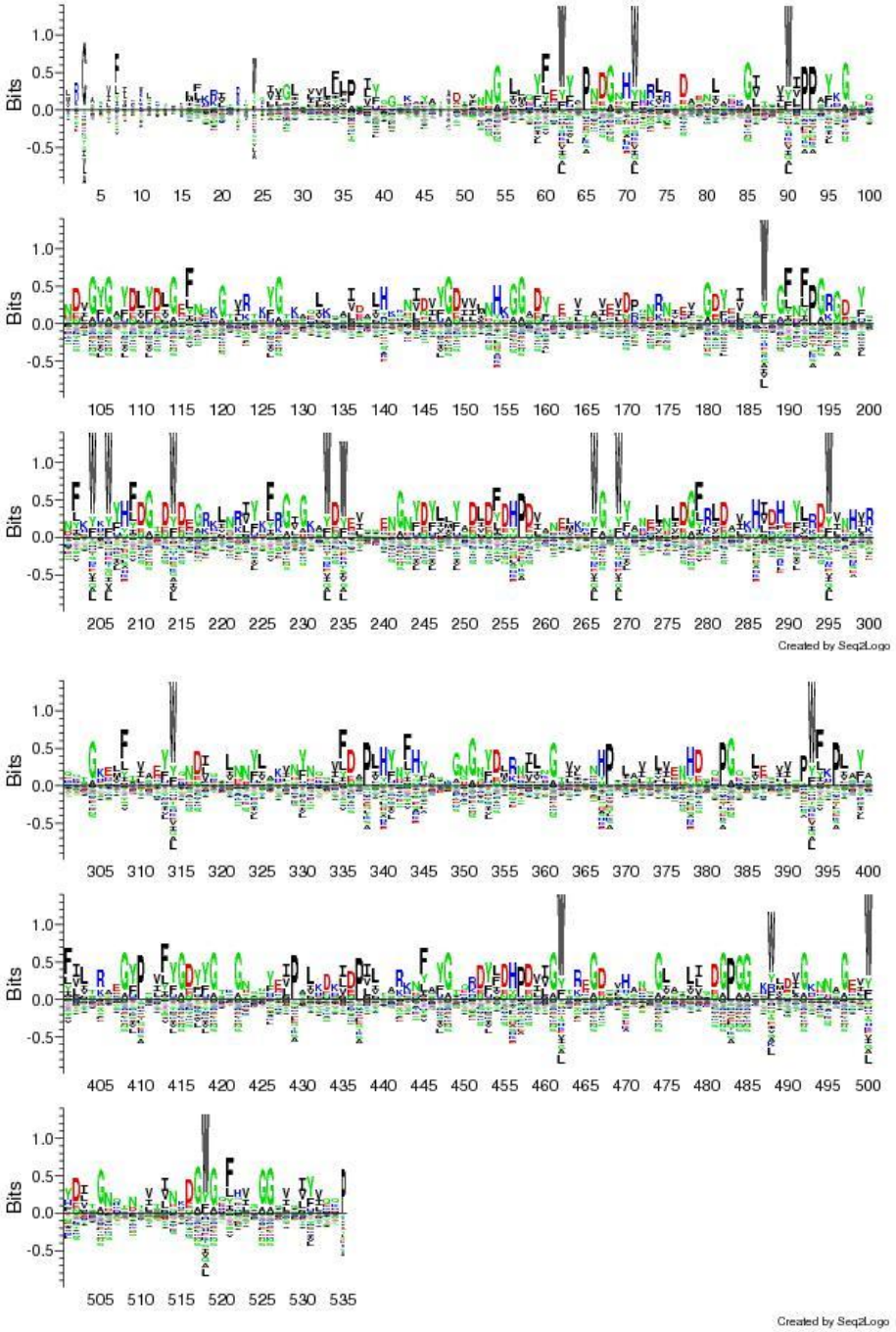


Figure 8. *Multiple sequence alignment B.licheniformis data set output represented by Seq2logo, a web-based sequence logo generation method. Sequence logos are a graphical representation of the information content stored in a multiple sequence alignment (MSA).*

Phylogenetic trees

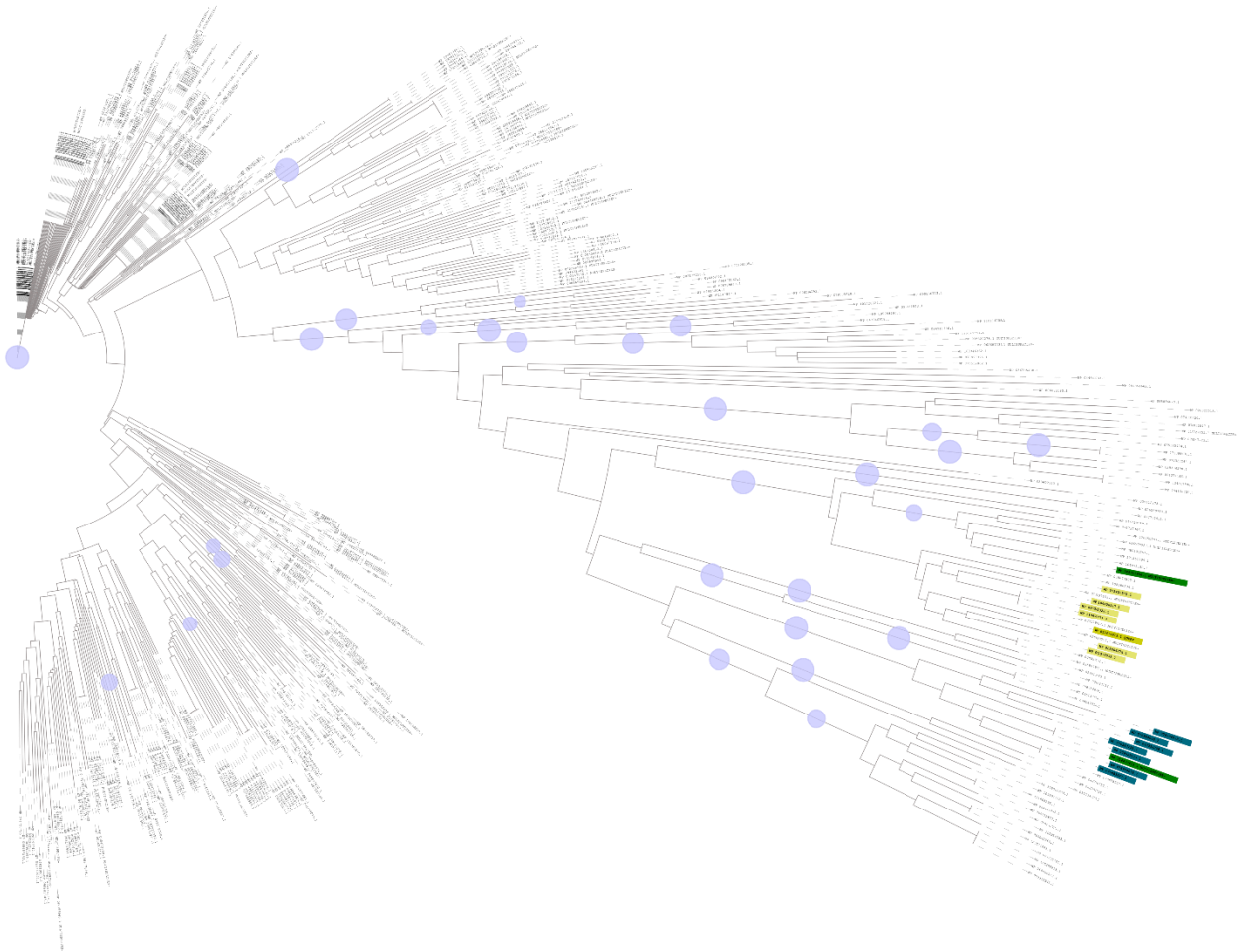


Figure 9.Neighbor joining tree for B.licheniformis. Bootstrap values (95%-100%) are represented by blue circles. Industrial producer sequences are remarked with the following colors: B.licheniformis in yellow (Alpha-amylase query sequence is underlined with a stronger yellow), B. amyloliquefaciens in blue and B.subtilis in green.

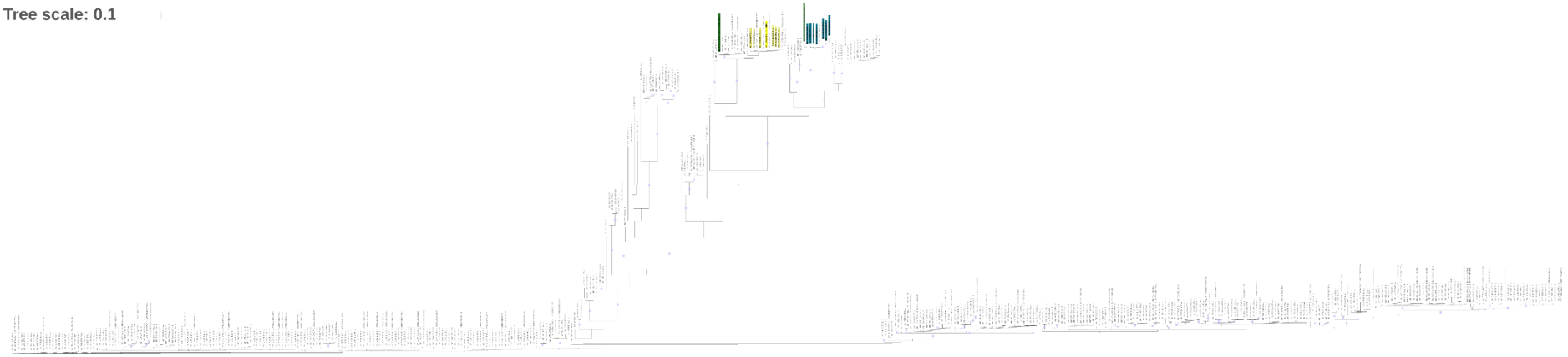


Figure 10.Maximum likelihood.for B.licheniformis. Bootstrap values (95%-100%) are represented by blue circle. Industrial producer sequences are remarked with the following colors: B.licheniformis in yelow (Alpha-amylase query sequence is underlined with a stronger yellow), B. amyloliquefaciens in blue and B.subtilis in green.

6.1.2. B. *licheniformis* comparative genomics

Several hits from the TBLASTN are returned and as we expected all of them are from different *B. licheniformis* genomes. Indeed, the sequence might be annotated in various genomes. The result with the highest parameters in both searches is the one assembled to the *Bacillus licheniformis* WX-02.

TBLASTN resulting parameters have a score of 2692, 99.6% of identity and 0 E-value. The genomic location is the supercontig CP012110 between the positions 695570-697186 and amyL is the gene overlapped (figure 11). AmyL transcript is a cytoplasmatic alpha amylase, besides, all the genes surrounding the region are protein coding.

Regarding HMMER results, the significant hits distribution is all over the bacteria kingdom. The majority of the hits are from *B.licheniformis* species but the one with the highest punctuation corresponds to the *Bacillus licheniformis* WX-02 genome as well (figure 12). In addition, it is also overlapped with the amyL gene.

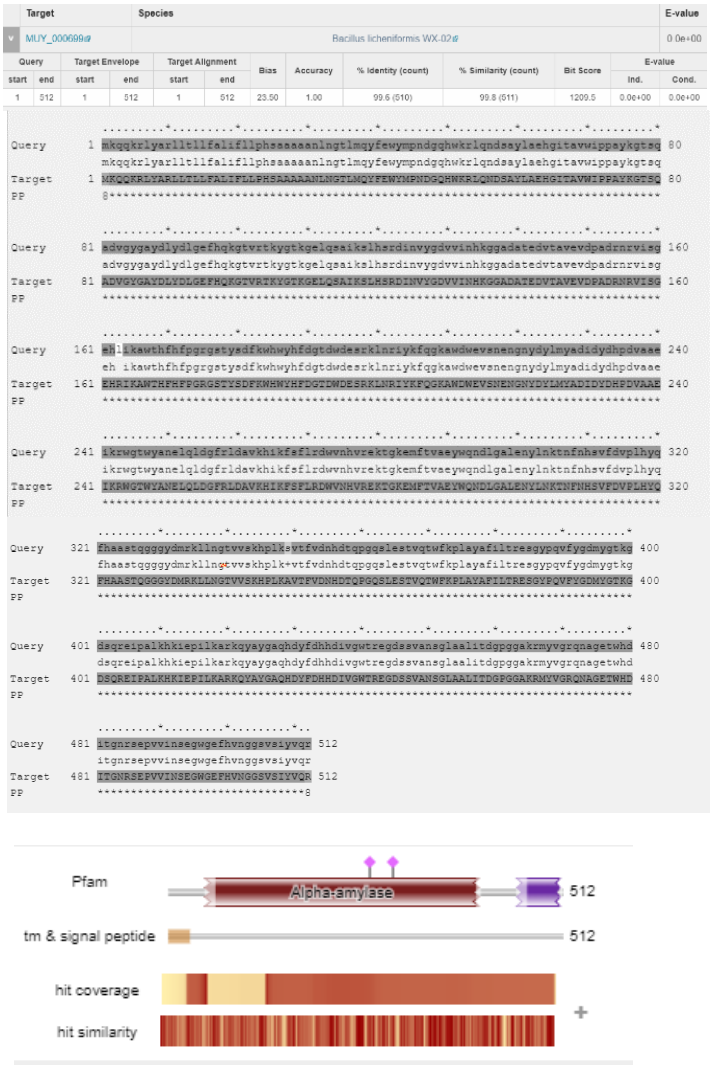


Figure 11. B.licheniformis alpha-amylase search results at the HHMER tool. Right: Sequence matches and features. Left: alignment between query sequence and Bacillus licheniformis WX-02 genome.

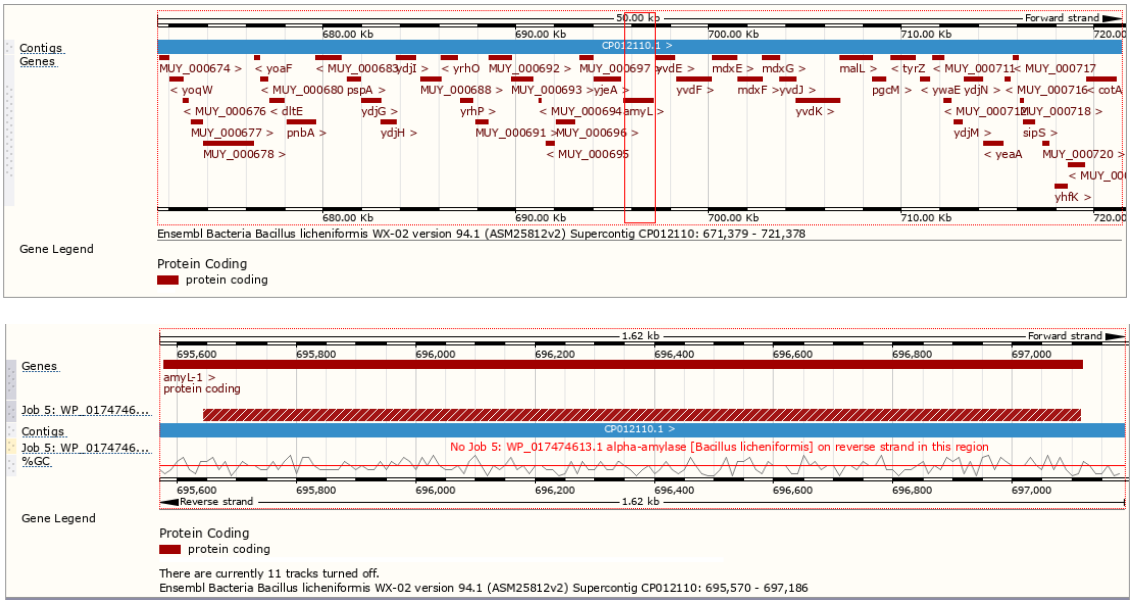


Figure 12. B.licheniformis alpha-amylase sequence (template) aligned at B.licheniformis genome.

6.2. *Aspergillus oryzae* alpha-amylase results

A.oryzae alpha-amylase query sequence used for the analysis is:

>CAA31220.1 alpha-amylase [*Aspergillus oryzae*]

MMVAWWSLFLYGLQVAAPALAATPADWRSQSIYFLLTDRFARTDGSTTATCNTADRKYCGGTWQGIIDKL

DYIQGMGFTAIWITPVTAQLPQTTAYGDAYHGYWQQDIYSLNENYGTADDLKALSSALHERGMYLMVDVV

ANHMGYDGAGSSVDYSVFKPFSSQDYFHPFCLIQNYEDQTQVEDCWLGDNTVSLPDLDTTKDVVKNEWYD

WVGSLVSNYSIDGLRIDTVKHVQKDFWPGYNKAAGVYCIGEVLDGDPAYTCPYQNVMDGVLNYPIYYPLL

NAFKSTSGSMDDLYNMINTVKSDCPDSTLLGTFVENHDNPRFASYTNDIALAKNVAAFIILNDGIPIIYA

GQEQHYAGGNDPANREATWLSGYPTDSELYKLIASANAIRNYAISKDTGFVTYKNWPIYKDDTTIAMRKG

TDGSQIVTILSNKGASGDSYTLSLSGAGYTAGQQLTEVIGCTTVTVGSDGNVPVPMAGGLPRVLYPTEKL

AGSKICSSS

It is a 499 amino acid sequence and present in the region named AmyAc_euk_AmyA. The region corresponds to the catalytic alpha-amylase domain, which is also found in eukaryotic. As well as, *B.licheniformis* alpha-amylase, three different conserved domains have been detected and they are shown in the figure 13.

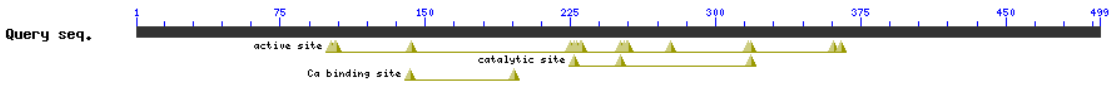


Figure 13. Conserved domains from A.oryzae alpha-amylase CAA31220.1.

6.2.1. Phylogenetic analysis of *A.oryzae* alpha amylase

PSI-BLAST

PSI-BLAST results are shown below (figure 14) and identity and query covers have high percentages, likewise, e-values are 0. The identity percentages range is from 45% to 100% and 84% to 100% for the query cover. Nevertheless, some sequences with a low identity values have query covers of 98%. In this case, only the query cover parameter has been taken into account to built the desired data set and sequences with at least a 98% of coverage have been retrieved.

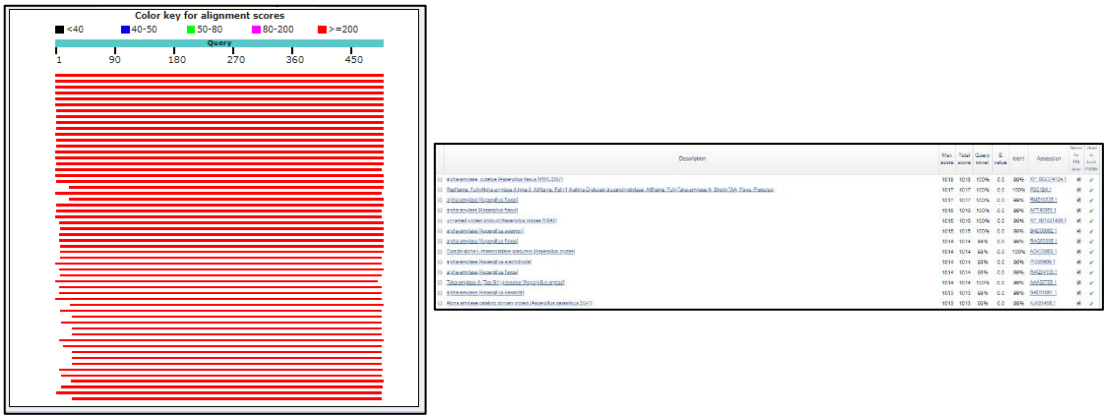


Figure 14.PSI-BALST second iteration results for the CAA31220.1sequence. At the right top hits are shown.

A total of 223 sequences are at *A.oryzae* data set and the majority of the sequences belong to *Aspergillus* spices. The data set can be found at the annex section.

Multiple sequence alignment

Figure 15 shows the graphical representation of the multiple sequence alignment. The Figure provides a compact and highly intuitive representation of the position-specific amino acid composition. In a sequence logo, the height of the bar is equal to the information content at each amino acid position, so as mentioned before enriched amino acids at each peptide position are represented in the positive y-axis and depleted amino acids are represented in the negative y-axis.

The resulting consensus sequence, obtained from the multiple sequence alignment, is full of uncertainties or non consensus positions and gaps. There is a total of 287 variable positions and It is as follows:

```
------------------------------------------------------------------------------
--??????--------????????--?????A--------A????WR?RSIYQ??TDRFA-RTDGST-T----A-
?-------C--?-?????YCGGTW?GII?KLDYIQGMGF?AIWISPV??N--???----?-T?-?G-?AYHGYW-
--??D?--Y?L-N?HFGT??DL??LS?ALH?RGMYLMVDVV?NHMA-----??---------G-????----?D-
-----YS?--F--?-PFN???YFHP-YC?I--?D--?-----?----N---?T????--------CW?G---D--
--??----V?-LPDL?TE???V??????W???LV??YS-----------------IDGLRID?-------AKHV?
??F?P--?F??AA-GVF??GEV???D-----????C?YQ----????G?LNYP?--------------------Y
Y?L??AF--??T-?G-????L???I????-??--C---------?D???--LG?F?ENHD?P-RFAS?---T?D?
?---------------------------LAKNA??F??L?DGIPI---------?Y?----------GQEQH?
?------G???--P?--NREALWL--SGY-?T?---?-?LY??I??LN?IR-??A??--?D?----------?-Y
?-???------??P?Y???---?-??A?RKG?-?G?QV??VLSN?------G??---?--??Y?--?????--?G
?-??G??V?E?L-?C-???T-V???---G?L?V?M-?---?G?PRV??P????--?--GS--GLCG---------
-----------------------------------------------??????????????????????----
-------------------------???????????--------------?????????????----------
------------------------------------------------------
```

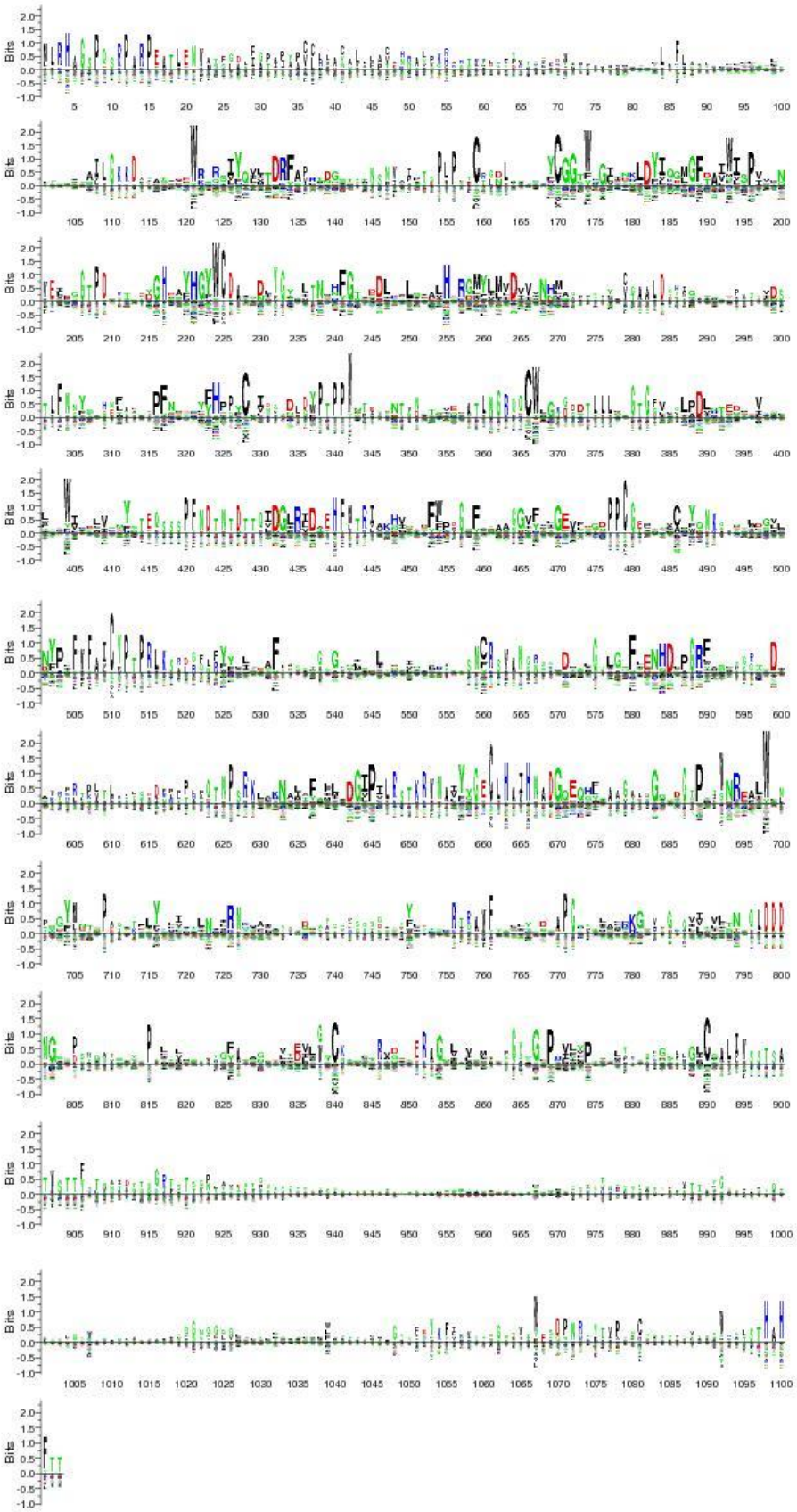


Figure 15.Multiple sequence alignment A.oryzae data set output represented by Seq2logo, a web-based sequence logo generation method. Sequence logos are a graphical representation of the information content stored in a multiple sequence alignment (MSA).

If the consensus sequence composition and the enriched amino acids are compared, the amino acid letters and the enriched amino acids coincide in the consensus positions. Each consensus character match with the enriched amino acid at the same location.

There are enriched amino acids in almost every position of the alignment, however, some of them have a high degree of conservation than others. The most enriched amino acids match with the consensus sequence positions as well.

Phylogenetic trees

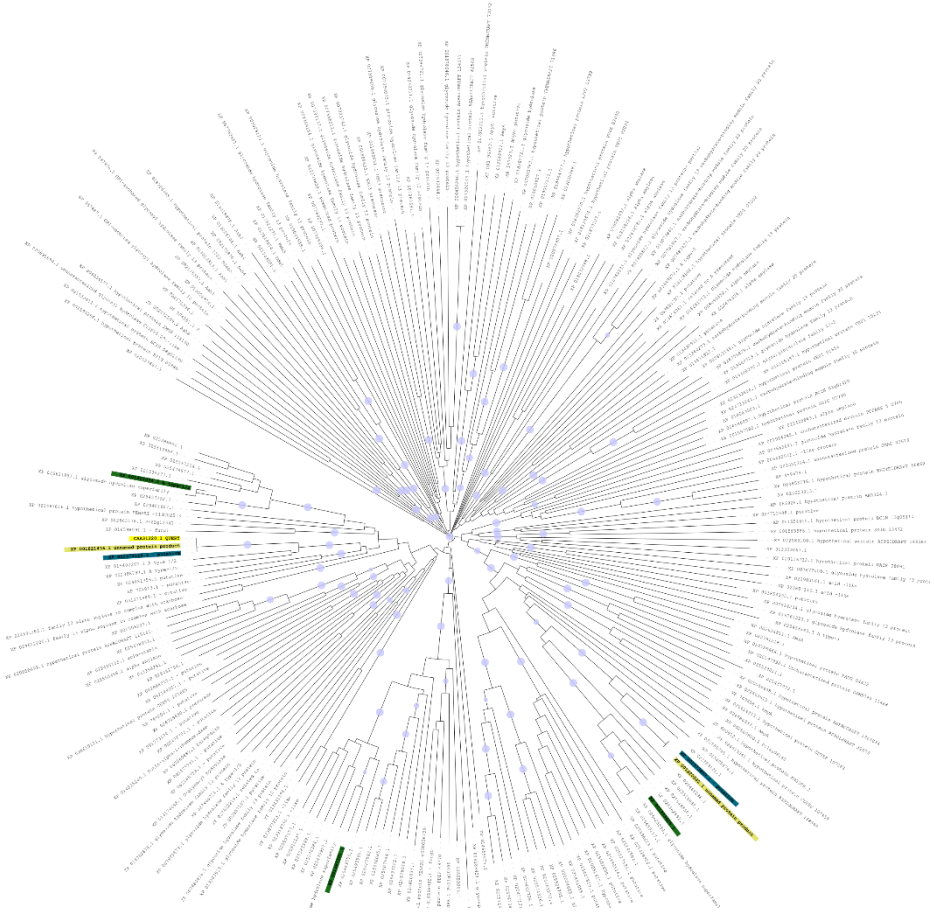


Figure 16.Neighbor joining tree for A.oryzae. Bootstrap values (95%-100%) are represented by blue circles. Alpha-amylase protein sequences from industrial producers are remarked: A.oryzae in yellow (Alpha-amylase query sequence is underlined with a stronger yellow), A.niger in green and A.flavus in blue.

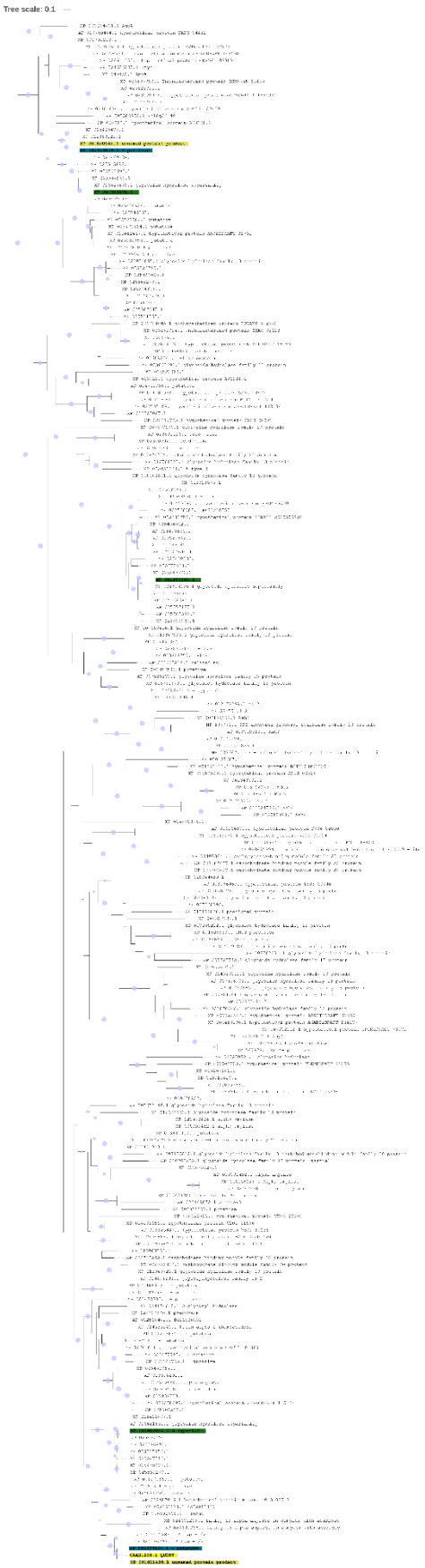


Figure 17.Maximum likelihood for A. oryzae. Bootstrap values (95%-100%) are represented by blue circles. Alpha-amylase protein sequences from industrial producers are remarked: A.oryzae in yellow (Alpha-amylase query sequence is underlined with a stronger yellow), A.niger in green and A.flavus in blue.

6.2.2. *A.oryzae* comparative genomics

TBLASTN results in several hits, as we expected due to the data base contains various *Aspergillus oryzae* annotated genomes; *Aspergillus oryzae* 100-8, *Aspergillus oryzae* RIB40 and *Aspergillus oryzae* 3.042.

HMMER significant hits distribution are all over Eucaryota kingdom. Furthermore, HMMER results contain several hits with high values for *A.oryzae* genomes, but there is only one which coincides with one of the TBLASTN hits, which corresponds to the *A.oryzae* RIB40 genome (considered as a reference genome). On top, the identity percentage for the TBLASTN search is of 100% as well as a 0 e-value.

Because of the previously exposed, the hit which corresponds to the genome *A.oryzae* RIB40 has been chosen over the others (figure 18 and 19). The sequence is placed in the chromosome 5 between the positions 3180379-3180646 and the overlapped gene is AO090120000196, which is a protein coding. It encodes the alpha-amylase A type ½ (P0C1B3) and it is surrounded by other protein coding genes and tRNAs.

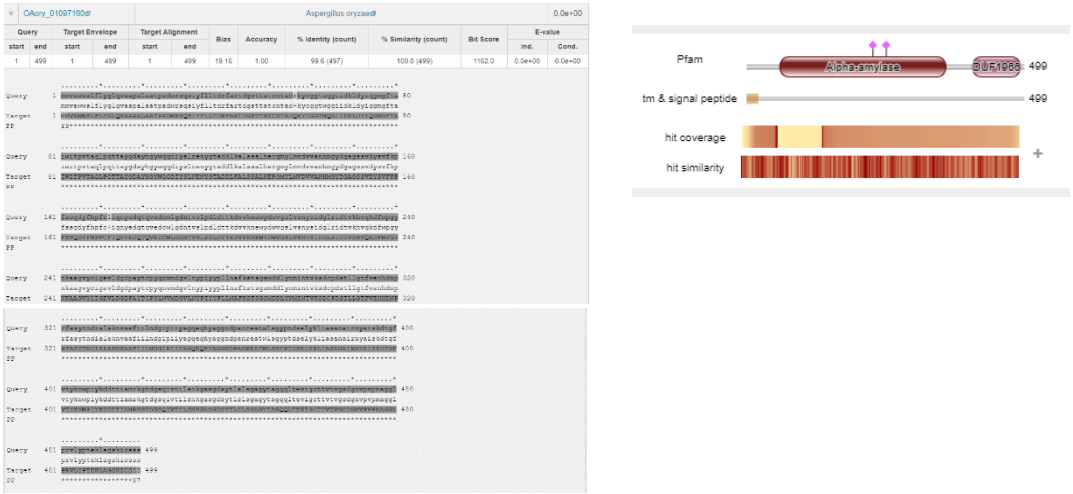


Figure 19. A.oryzae alpha-amylase sequence search results at the HHMER tool. Right: Sequence matches and features. Left: alignment between query sequence and A.oryzae RIB40 genome.

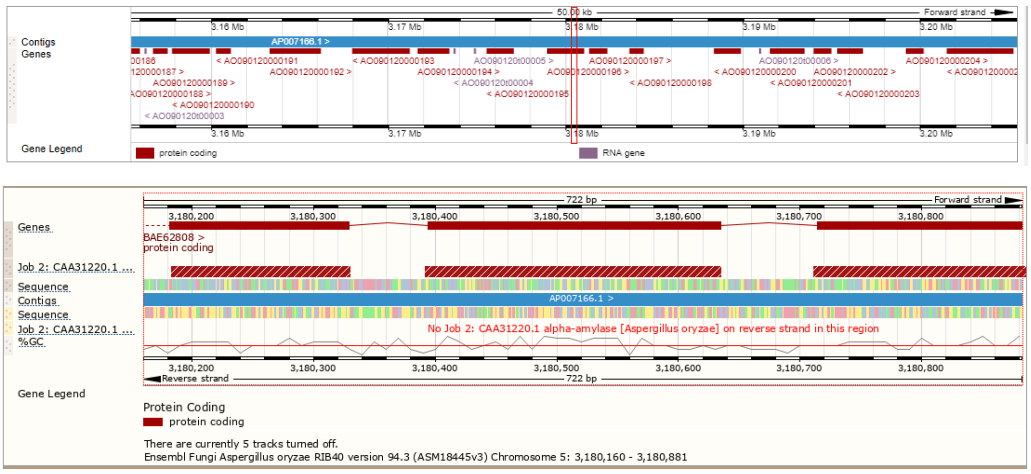


Figure 18. A.oryzae alpha-amylase sequence (template) aligned at A.oryzae genome.

6.3. Protein profile

6.3.1. *B.licheniformis* alpha-amylase

Structural and functional characterization

Secondary structure composition is shown at table 2, where the predominant structure is the random coil followed by the alpha helix.

Table 2. Secondary structure of alpha amylase B.licheniformis protein using SOPMA.

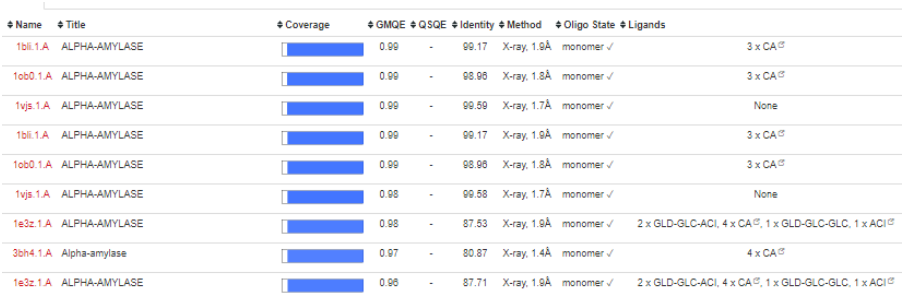
Parameters	Number of amino acids	Amino acids (%)
Alpha helix	171	33.40%
310 helix	0	0%
Pi helix	0	0%
Beta bridge	0	0%
Extended strand	111	21.68%
Beta turn	26	5.08%
Bend region	0	0%
Random coil	204	39.84%
Ambiguous states	0	0%
Other states	0	0%

Several motifs are found in the sequence; 6 N-glycosilation sites, 1 Tyrosine kinase phosphorylation site, 9 N-myristoylation site, 12 Casein kinase II phosphorylation site and 4 Protein kinase C phosphorylation sites.

Domain analysis carried with CCD database, detected the PRK09441 from the cl25947 alpha-amylase superfamily.

Homology modelling

Swiss Model returns a total of 50 possible templates but not all of them correspond to the alpha-amylase protein, for example some of them are alpha-1,4-glucan-4-glucanohydrolase, glucan 1,4-alpha-maltohexaosidase, sucrose isomerase, among others. However, the ten top hits belong to *B.licheniformis* (figure 20) and they do correspond to alpha-amylase protein. Even though some of them correspond to the same PDB entry, they do not pertain to the same biounit, meaning that they correspond to different chains of the same protein. Moreover, the identity for these hits oscillates from 81-99%.



Name	Title	Coverage	GMQE	QSQE	Identity	Method	Oligo State	Ligands
1bli.1.A	ALPHA-AMYLASE		0.99	-	99.17	X-ray, 1.9Å	monomer ✓	3 x CA
1ob0.1.A	ALPHA-AMYLASE		0.99	-	98.96	X-ray, 1.8Å	monomer ✓	3 x CA
1vjs.1.A	ALPHA-AMYLASE		0.99	-	99.59	X-ray, 1.7Å	monomer ✓	None
1bli.1.A	ALPHA-AMYLASE		0.99	-	99.17	X-ray, 1.9Å	monomer ✓	3 x CA
1ob0.1.A	ALPHA-AMYLASE		0.99	-	98.96	X-ray, 1.8Å	monomer ✓	3 x CA
1vjs.1.A	ALPHA-AMYLASE		0.98	-	99.58	X-ray, 1.7Å	monomer ✓	None
1e3z.1.A	ALPHA-AMYLASE		0.98	-	87.53	X-ray, 1.9Å	monomer ✓	2 x GLD-GLC-ACI, 4 x CA, 1 x GLD-GLC-GLC, 1 x ACI
3bh4.1.A	Alpha-amylase		0.97	-	80.87	X-ray, 1.4Å	monomer ✓	4 x CA
1e3z.1.A	ALPHA-AMYLASE		0.96	-	87.71	X-ray, 1.9Å	monomer ✓	2 x GLD-GLC-ACI, 4 x CA, 1 x GLD-GLC-GLC, 1 x ACI

Figure 20. Top ten templetes for B.licheniformis alpha-amylase protein (WP_017474613.1)

The second sequence is selected to build the model (highest identity percentage) and its PDB entry is 1OB0. Proteins aligned are homologs, so the secondary structure regions are highly conserved.

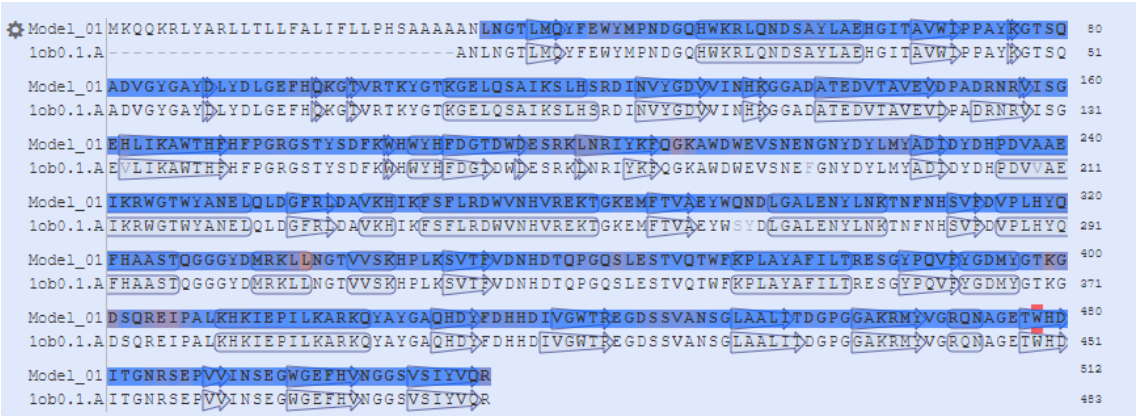


Figure 21. Aligment template 1OB0 and model WP_017474613.1.

Then, resulting parameters for the model are analyzed. The high sequence identity and coverage indicates that the two proteins are homologous, and the secondary structure have been well predicted. Observing figure 21, the model sequence is pretty much blue which is indicative of a good prediction for the structural motifs.

QMEAN is satisfactory. GMQE (Global Model Quality Estimation) value is close to 1 so It indicates that the generated model is reliable. If we observe the local quality estimate graphic (figure 22), most of the residues are between 0.8 and 1 and there are not values below 0.6, therefore, it is an indicative that the resulting model is of quality. Finally, the comparison chart shows that the modeling score is within the range of scores of reference structures of the same size, the absolute Z-score value obtained is -0.60. All in all, the model obtained is very good and of quality.

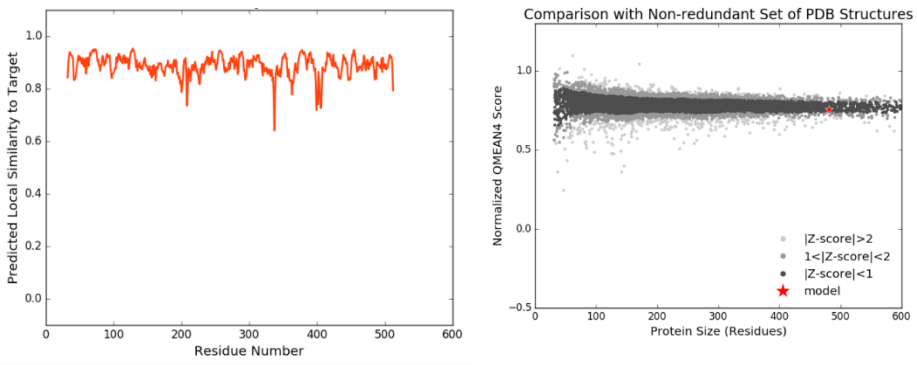


Figure 22. Output graphics. Left: Local quality estimate, for each model residue (x-axis) represents the predicted similarity (y-axis). Right: Comparison with Non-Redundant set of PDB structures, quality punctuations of the individual models are expressed as Z-scores and are compared with each crystalline high-resolution structure punctuation obtained (each point represents a protein structure).

The model obtained is shown at figure 23, as well as the ramachandran plot:

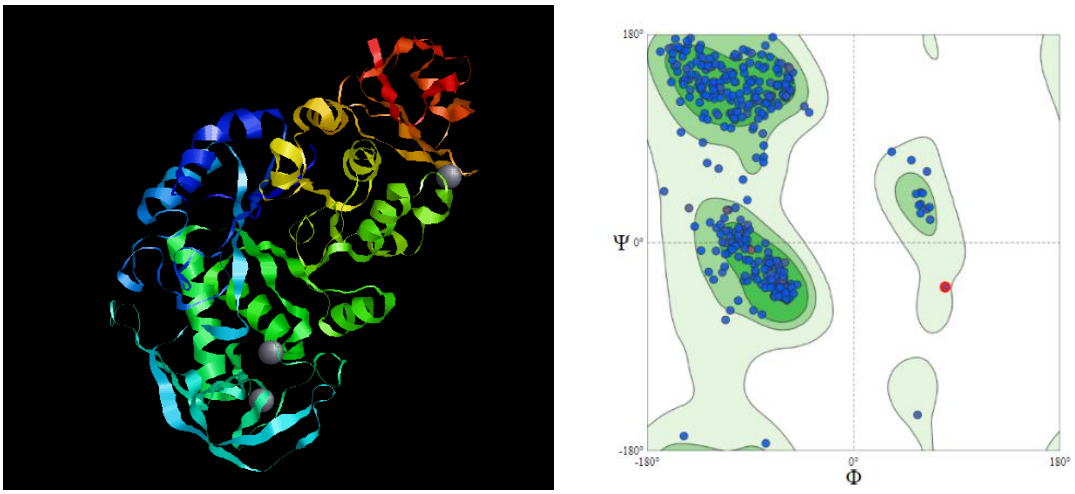


Figure 23. Left: Three-dimensional structure of predicted alpha-amylase B.licheniformis protein by SWISSMODEL.. Right: Ramachandran plot of the B.licheniformis model.

If the ramachandran plot is observed (figure23), most of the residues are in the permitted areas, specially in the β sheets and α helix zones. Nevertheless, β sheets are more present than the other structures as well as in the model shown in figure 23. Regarding α helix, they are mostly right handed.

Ramachandran plot statistics have been calculated with the rampage browser:

```
Residue [A  93 :LEU] ( -92.66,  36.77) in Allowed region
Residue [A 172 :PHE] (  64.02,  61.49) in Allowed region
Residue [A 179 :TYR] (  79.92, -38.33) in Allowed region
Residue [A 225 :LEU] (-111.63, -66.72) in Allowed region
Residue [A 227 :TYR] (  55.80,-148.88) in Allowed region
Residue [A 266 :LYS] ( -38.74, 112.56) in Allowed region
Residue [A 285 :MET] (  50.41,  71.28) in Allowed region
Residue [A 355 :ASN] (-148.02,-167.13) in Allowed region
Residue [A 365 :GLU] ( -42.95, 117.78) in Allowed region
Residue [A 366 :SER] (-165.31,  44.41) in Allowed region
Residue [A 404 :ARG] (-143.55,  29.86) in Allowed region
Residue [A 432 :PHE] (-117.65,  63.96) in Allowed region
Residue [A  87 :ALA] (  33.31,  78.56) in Outlier region
Number of residues in favoured region    (~98.0% expected)  :  466 (
97.3%)
Number of residues in allowed region     ( ~2.0% expected)  :   12 (
2.5%)
Number of residues in outlier region                        :    1 (
0.2%)
```

Predicted statistics and obtained ones are super close, meaning that the resulting model is reliable and can be validated. On top, the amino acids are mostly found in energetic favorable regions.

6.3.1. *A.oryzae* alpha-amylase

Structural and functional characterization

Secondary structure composition is shown at table 3, where the predominant structures are radom coil and alpha helix.

Table 3.Secondary structure of alpha amylase A.oryzae protein using SOPMA

Parameters	Number of amino acids	Amino acids (%)
Alpha helix	163	32.67%
310 helix	0	0 %
Pi helix	0	0 %
Beta bridge	0	0 %
Extended strand	84	16.83%
Beta turn	28	5.61%
Bend region	0	0%
Random coil	224	44.89%
Ambiguous states	0	0%
Other states	0	0%

Several motifs are found in the *A.oryzae* alpha-amylase protein; 12 N-myristoylation site,11 Casein kinase II phosphorylation site, 7Protein kinase C phosphorylation site,1 N-glycosylation site and 1 cAMP- and cGMP-dependent protein kinase phosphorylation site.

Domain analysis detected the cd11319: AmyAc_euk_AmyA and it belongs to the alpha-amylase superfamily.

Homology modeling

Swiss model returns a total of 50 possible templates (figure 24), however, only four of them correspond to the alpha-amylase protein. Moreover, two of the alpha-amylase hits correspond to the same PDB entry being it 3KWX. As well, the other PDB entries are 2D0F and 1UH3. The 3KWX templates source organism is *Aspergillus oryzae*, in contrast, the other two source organism is Thermoactinomyces vulgaris.

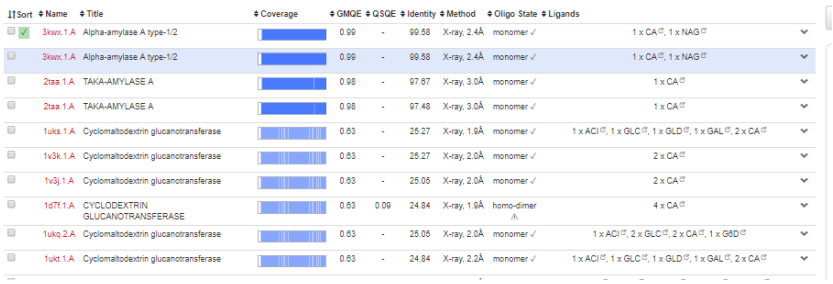


Figure 24. Top ten hits for A. oryzae alpha-amylase protein (CAA31220.1)

In reference to the identity percentages an exponential decrease can be observed after the four top hits, it goes from almost a 100% to a less than 30%. As the percentage identity falls below

30% (in the so-called 'twilight zone'), model quality estimation on the basis of sequence identity becomes unreliable (Bordoli et al., 2009).

All in all, the first template with a 99.58% identity value for the resulting alignment has been chosen to build de desired model. It is assumed that the proteins (model and template) are homologs, thereby, the secondary structure motifs are conserved. Furthermore, the colored blue zones indicate that the model is of quality (figure 25).

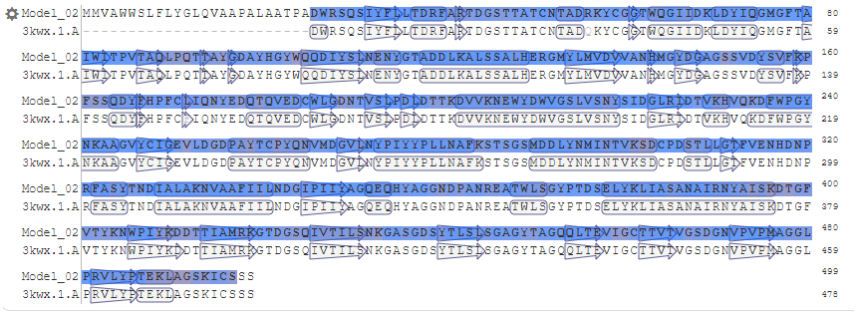


Figure 25. Alignment template 3KWX and model CAA31220.1.

Regarding the model quality, GMQE value is elevated (0.99, being the maximum 1) accordingly the resulting model is reliable. QMEAN value is satisfactory. Local quality estimate chart (figure 26) presents values between 0.6 and 1 but in any case, below 0.6. Finally, the comparison graphic (figure 26) shows that predicted structure falls within the range of scores of reference structures of the same size, the absolute Z-score for the model is 0.75.

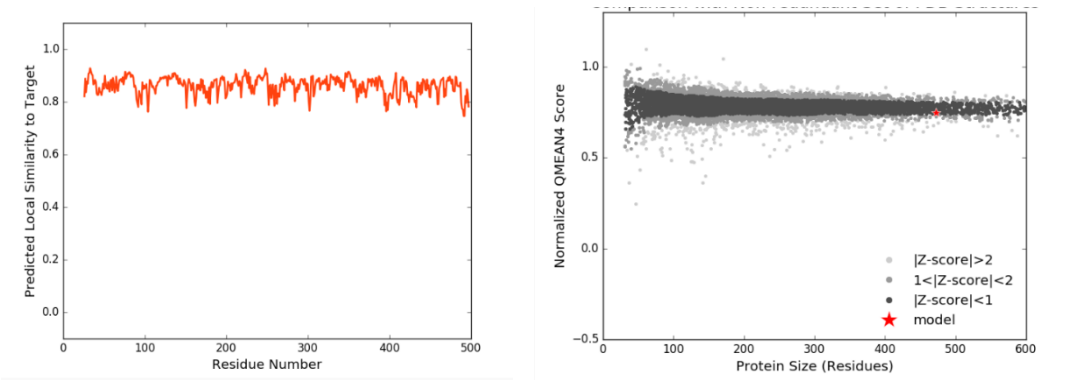


Figure 26. Output graphics. Left: Local quality estimate, for each model residue (x-axis) represents the predicted similarity (y-axis). Right: Comparison with Non-Redundant set of PDB structures, quality punctuations of the individual models are expressed as Z-scores and are compared with each crystalline high-resolution structure punctuation obtained (each point represents a protein structure).

The model obtained is shown at figure 26, as well as the Ramachandran plot.

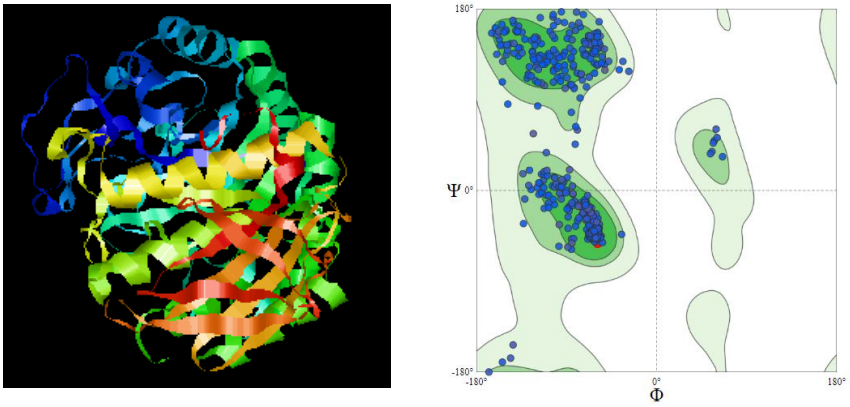


Figure 27.Left: Visualization of the model built with SWISSPROT for the A.oryzae alpha amylase, shaped with the secondary structure. Right: Ramachandran plot of the A.oryzae model.

Ramachandran plot statistics obtained with rampage browser:

```
Residue [A  51 :CYS] (-148.59,  85.57) in Allowed region
Residue [A  99 :ALA] ( -81.28,  45.78) in Allowed region
Residue [A 104 :TRP] (-123.14,  56.30) in Allowed region
Residue [A 106 :GLN] (-132.94, -32.83) in Allowed region
Residue [A 165 :ASP] ( -34.91, -43.52) in Allowed region
Residue [A 189 :ASP] (-143.30,-153.05) in Allowed region
Residue [A 342 :ASN] ( -57.35, 167.17) in Allowed region
Residue [A 350 :ALA] ( -37.29, 119.58) in Allowed region
Residue [A 373 :TYR] (  56.71,  47.05) in Allowed region
Residue [A 405 :ASN] ( -33.65, 127.84) in Allowed region
Residue [A 411 :ASP] (-145.80,-166.13) in Allowed region
Residue [A 473 :PRO] ( -81.26, 102.35) in Allowed region
Residue [A 496 :CYS] (  59.53,  60.78) in Allowed region
Residue [A 361 :ASP] ( -27.73, 118.00) in Outlier region
Number of residues in favoured region    (~98.0% expected)  :  456 (
97.0%)
Number of residues in allowed region     ( ~2.0% expected)  :   13 (
2.8%)
Number of residues in outlier region                        :    1 (
0.2%)
```

The residues in the Ramachandran plot shown in the figure 27, are mostly found in permitted regions, being the β sheet region the most crowded. There is only one residue in an outline region. Expected statistics and obtained ones meet, so the resulting model is reliable and also can be validated due to most of the amino acids are found in energetic favorable regions.

7. RESULTS DISCUSSION

Fungal and bacterial alpha amylases especially *Bacillus spp* alpha amylases are of special concern because of their significant thermo stability (Huma et al., 2014). However, fungal amylases are preferred over other microbial sources because of their accepted GRAS status (Naidu & Saranraj, 2013).

Bacterial alpha- amylase is produced by *Bacillus*, *Pseudomonas* and *Clostridium* species. Among these species, *B. subtilis, B. stearothermophilus*, B. *licheniformis* and *B. amyloliquefaciens* are known to be good producers (Naidu & Saranraj, 2013) and generally preferred because of their productivity (Hussain, Siddique, Mahmood, & Ahmed, 2013). In addition, most thermostable industrial alpha-amylase are produced by *B.licheniformis*.

Fungal sources are confined to terrestrial isolates mostly to *Penicillium* and *Aspergillus* spices. From the genus *Aspergillus; A.oryzae, A.niger, A. flavaus, A. tamarie, A. fumigatus* and *A. kawachii* have been frequently used for the production of alpha-amylase. From *Penicillum spp*; *P. chrysogenum* and *P. camemberti* serve in the production of alpha-amylase (Gowhar, Azra, Ruqeya, Suhaib, & Tauseef, 2014; Kathiresan & Manivannan, 2006) .

As for the two data sets generated, all of the industrial *Bacillus* species, except for *B. stearothermophilus* are present in the *B.licheniformis* dataset. Besides that, A.*oryzae* data set contains three *Aspergillus spp* (*A.oryzae, A.niger* and *A. flavaus*) but not industrial *Penicillium* species.

In the *B.licheniformis* data set there are several entries for the industrial producers being the *B.licheniformis* and B. amyloliquefaciens the majority. As well, other *Bacillus spp* like *B.cereus, B.mycoides, B.paralicheniformis* and *B.hayensii* can be found, thereby, they are natural producers (Hussain et al., 2013). Furthermore, sequences present in the data set are all alpha-amylase proteins. Indeed, PSI-BLAST parameters are an indicative; higher the percent identity is the more significant the match, so the sequences aligned have almost the same residues at the same positions. The resulting sequences have roughly the same length as the query and a 0 e-value indicates that by chance two aligned sequences will never be the same and it's significative. Another thing to take into account is the multiple sequence alignment, where the presence of gaps is minimal and the residues are mostly conserved.

In the other hand, *A.oryzae* data set, also contains several entries for *A.oryzae*, *A.niger* and *A.flavus* as well as natural producers such as *A. fumigatus,* and *P. expansum* (Hussain et al., 2013). In this case not all the sequence present in the data set correspond to alpha-amylase proteins, hence sequences from *Penicillium* spieces *P.rubens, P.arizonense* and *P.digitatum* do not correspond to alpha-amylase protein sequence and some other microorganisms such as *Punctularia strigosozonata, Penicilliopsis zonata, Podospora macrospora, Exaiptasia pallida* and *A. muludensisis* do not either. Additionally, *A.oryzae* entries corresponding to the strain RIB40 are unnamed proteins but If the sequences are BLAST at Uniprot they result to be alpha-amylase proteins. Least similarity has been shown for the *A.oryzae* alpha-amylase data set, as before, can be deduced from the PSI-BLAST parameters results and multiple sequence alignment. Low identity between query sequences and numerous gaps are present, amino acid characters are diverging. Nevertheless, e-values are significative.

Homology can be inferred in both data because of statistically significant similarity. Two proteins are considered homologous when they have the same evolutive origin and a similar structure and function, besides sequence similarity searches identify homologous proteins by detecting

excess similarity along with low statistical estimates (e-value) and more than 30% identity values (Pearson, 2013) .

Regarding the phylogenetic trees, neighbor-joining tree establishes relationships between sequences according to their genetic distance, without taking into account an evolutionary model. Maximum likelihood, in contrast, uses a more complex evolution model so the phylogeny is more robust. Moreover, the concept of homology is central on the computational analysis (Pearson, 2013) which has been previously stablish. Phylogeny of alpha amylase is generally in agreement with their origin, all fungal alpha-amylases are more related to each other than to the alpha-amylases originating from plants or animals. Alpha-amylases from bacteria, instead, are scattered together over several clusters, which group with animal, plant or fungal alpha-amylase (Chen, Xie, Shao, & Chen, 2012; Stam, Danchin, Rancurel, Coutinho, & Henrissat, 2006).

In the present study phylogenetic results shows unrooted trees, which take alpha amylase sequences as a base of analysis. Bacterial alpha-amylase industrial producers are closely related, as they can be grouped together in a cluster (figure 9 and 10), they all share a common node with 54 leaves on it and a distance of 0.014 for the NJ and 0.193 for the ML. Together with other *Bacillus spp* alpha-amylases (*B.paralicheniformis, B.hayenso B.glycinifermatans, B.valenzis, B.nakamurai, B.mojavensis* and *B.halotolerans*) being the spices *B.paralicheniformis* and *B.halotolerans* the majority, like that they are nearby *B.licheniformis* and *B.amyloquefaciens* respectively. Distant relationship from the industrial cluster are represented by numerous alpha-amylases from *B.cereus* and *B.Thurgenesis*.

Steaphens et al., (1984) and Yukki et al.,(1985) reported homology between *B.licheniformis* and *B.amyloquefaciens* alpha-amylase, which is supported by the phylogenetic trees. *B.stearothermophilus* amino acid sequence homology to those from *B.licheniformis* and *B.amyloquefaciens* has been reported by Sudv et al., (2001) , likewise, they are liquefying-type enzymes (Alikhajeh et al., 2010). Even though, there is evidence of the homology between *B.licheniformis* and *B.stearothermophilus*, alpha-amylase sequences of this last spice has not been retrieve. In fact, they exhibit relatively low similarity, 65% identity in Sudv et al., (2001) study and in the actual study all sequences show at least a 98% of identity.

Fungal alpha amylase industrial producers are widely distributed in the phylogenetic trees (figure 16 and 17). Two different groups of industrial alpha-amylase can be observed in the NJ and ML trees and they are distantly related. The clusters are depicted by at least one sequence of each *Aspergillus spp* industrial producer. The smallest one emerges from a 11 leaves node with a distance of 0.025 for the NJ tree and 0.065 for the ML tree, where is closer to the main branch. As well, it contains alpha-amylase sequences of *A. Costaricaensis CBS 11574*, *A.welwitschiae, A.vadensis CBS 113365, A.piperis, A.sclerotioniger, A.bombycis, A.nomius NRRL 13137*. The other group emerges from a 22 leaves node, with a 0.017 distance value for the NJ and 0.094 for the ML and contains alpha-amylase sequences from *A.bombycis, P.expansum, A.novofumigatus IBT 16806, A.costaricaensis, A.vadensis CBS 113365, A.piperis CBS 112811, A.eucalypicola CBS 122712.*

Alpha-amylase primary protein sequence of *A.niger, A.flavus* and *A.oryzae* RIB40 homology has been previously reported by Avwiorolo et al.,(2018), and it is supported by the phylogenetic trees.

A protein family is a group of proteins that share common evolutionary origin, so alpha-amylases belong to a family named glycoside hydrolase family 13 (GH-13). At the same time, most of the

starch hydrolyzing enzymes belong to the alpha-amylase family based on amino acid sequence homology (Davies & Henrissat, 1995; Henrissat et al., 1995; Reddy, Nimmagadda, & Rao, 2002).

Currently fungal alpha-amylases are classified into two subfamilies GH13_1 and GH13_5 (Chen et al., 2012). Extracellular and fungal specific are members of GH13_1, while GH13_5 is formed for the intracellular type and bacterial alpha amylases (Stam et al., 2006). Furthermore, alpha-amylases from GH13_1 display very low similarity with the fungal alpha-amylases GH13_5 (van der Kaaij, Janecek, van der Maarel, & Dijkhuizen, 2007). The existence of two alpha-amylase fungal subfamilies and the low similarity existence between the subfamilies explains the diverse protein presence in the *A.oryzae* data set as well.

At a genomic level, *B.licheniformis* alpha amylase enzyme is codified by amyL gene. This gene is temporally expressed and subject to catabolic repression (Laoide & McConnell, 1989; Rothstein, Devlin, & Cate, 1986) when glucose is present.

A.oryzae strains has two or three alpha-amylase genes; amyA,amyB and amyC. with identical nucleotide sequences (Machida et al., 2008). Nevertheless, amyA has one or two mismatches in the 5'-flanking and coding regions, compared with amyB and amyC which have identical nucleotide sequences in their 5'-falcking and 3'-flacking coding regions. *A.oryzae* alpha-amylase protein sequence used in this paper is overlapped with the amyB gene (searched in AspGD database).

Finally, the structure of alpha-amylase consists of a polypeptide chain folded into three domains called A, B and C besides these sites are generally found in all alpha-amylase. A is a (β/α) 8-barrel; B is a loop between the beta 3 strand and alpha 3 helix of A; C is the C-terminal extension characterized by a Greek key. The catalytic residues are Asp, Glu and Asp. On top, enzymes are believed to have a $(\alpha\beta)_8$ or TIM barrel structure, that contains the catalytic amino acid. All alpha-amylases contain one strongly conserved Ca^{2+} ion for structural integrity and enzymatic activity, however, alpha-amylases from B. *licheniformis* and *B.stearothermophilus* are reported to have two additional calcium binding sites (Saini et al., 2017). In effect, *B.licheniformis* alpha amylase structure have 3 calcium binding sites, whereas, *A.oryzae* alpha amylase has only one binding site.

If figure 6 and 13 are observed three conserved domains are found for the two proteins, they are an active site, catalytic site and Na/Ca binding site. About the second structure in both alpha-amylases the random coil is predominant, approximately 50% of the residues present in the secondary structure, followed by alpha helix (~33%) and extended strand (~20%). Also, beta turn is present (~5%). Random coil structures link the stand and helix, they are short sequences with a non-repetitive conformation. Random coils have important functions in proteins for flexibility and conformational changes such as enzymatic turnover (Buxbaum, 2007).

Protein phosphorylation is the molecular mechanism through which protein function is regulated in response to extracellular stimulation both inside and outside the nervous system and involves protein kinase (Nestler & Greengard, 1999). The appearance of several phosphorylation sites is remarkable and indicates that the protein is frequently regulated.

Myristoylation is one such protein lipid modification, which plays vital roles in cellular signaling, protein–protein interaction, and targeting of proteins to endomembrane and plasma membrane systems and it is observed in plants,animals, fungi and viruses (Udenwobele et al., 2017). As before, the occurrence of several myrstoylation sites is remarkable as it regulates the protein activity.

At last, protein tridimensional structure is crucial to understand the alpha-amylase function and can be used for future analysis to improve the protein production capacity. Even though, both proteins have been already well-characterized and studied.

8. CONCLUSIONS

Nowadays, a large number of microbial amylases are available commercially and they have almost completely replaced chemical hydrolysis of starch in starch processing industry. Even though that alpha amylases are widely use, a few selected strains of fungi and bacteria meet the criteria for commercial.

The most important bacterial and fungal microorganisms used in the production of alpha-amylase have been establish by computing the neighbor joining and maximum likelihood phylogenetic trees and evolutive relationships could have been determinate. In the case of potential homologs proteins, *Bacillus paralicheniformis* and *Bacillus halotolerans* are a potential industrial producers due to the closest phylogenetic relationship with the alpha-amylase industrial producers. Furthermore, they have not been used in the industrial sector, yet. Future studies can be focus on the alpha-amylase protein of *B.paralicheniformis* and *B.halotolerans* in order to achieve the efficient industrial productivity the structural and functional relationships of this spice must be known in detail.

Planning has not been 100% accomplished as the phylogenetic analysis are computationally complex. Additionally, performing all the previous analysis means understanding all the parameters and needs for each individual step. Some changes have been introduced during the Project developing, for example the proprieties of microorganisms that makes some of them more useful and better than others have not been evaluated, because It requires experimental data which is hard to get without laboratory procedures.

Future possible analysis for this project is the alpha-amylase GH13 family analysis, so phylogenetic relationships and evolutionary history of the alpha-amylase family could be reported. Actually, a phylogenetic family analysis draws conclusions of biological functions which might not be apparent and provide information on evolutionary relationship and functional diversity within the family.

Another possible study on the basis of this paper, is the *B.licheniformis* and *A.oryzae* genome and phylogenetic relationship analysis within the two species.

9. BIBLIOGRAPHY

Adrio, J. L., & Demain, A. L. (2014). Microbial enzymes: tools for biotechnological processes. *Biomolecules*, *4*(1), 117–139. https://doi.org/10.3390/biom4010117

Alikhajeh, J., Khajeh, K., Ranjbar, B., Naderi-Manesh, H., Lin, Y.-H., Liu, E., … Chen, C.-J. (2010). Structure of *Bacillus amyloliquefaciens* α-amylase at high resolution: implications for thermal stability. *Acta Crystallographica Section F Structural Biology and Crystallization Communications*, *66*(2), 121–129. https://doi.org/10.1107/S1744309109051938

Anbu, P., Gopinath, S. C. B., Cihan, A. C., & Chaulagain, B. P. (2013). Microbial Enzymes and Their Applications in Industries and Medicine. *BioMed Research International*, *2013*, 1–2. https://doi.org/10.1155/2013/204014

Bhagwat, M., & Aravind, L. (2007). PSI-BLAST tutorial. *Methods in Molecular Biology (Clifton, N.J.)*, *395*, 177–186. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17993673

Bhagwat, M., Young, L., & Robison, R. R. (2012). Using BLAT to find sequence similarity in closely related genomes. *Current Protocols in Bioinformatics*, *Chapter 10*, Unit10.8. https://doi.org/10.1002/0471250953.bi1008s37

Binod, P., Palkhiwala, P., Gaikaiwari, R., Nampoothiri, M., Duggal, A., Dey, K., & Pandey, A. (2013). *Industrial Enzymes-Present status and future perspectives for India*. *Journal of Scientific & Industrial Research* (Vol. 72). Retrieved from https://pdfs.semanticscholar.org/9fb5/cb17da8a2fd265ca55c7ad97e0a63cd6cbf1.pdf

Bordoli, L., Kiefer, F., Arnold, K., Benkert, P., Battey, J., & Schwede, T. (2009). Protein structure homology modeling using SWISS-MODEL workspace. *Nature Protocols*, *4*(1), 1–13. https://doi.org/10.1038/nprot.2008.197

Bromham, L., & Penny, D. (2003). The modern molecular clock. *Nature Reviews Genetics*, *4*(3), 216–224. https://doi.org/10.1038/nrg1020

Chen, W., Xie, T., Shao, Y., & Chen, F. (2012). Phylogenomic Relationships between Amylolytic Enzymes from 85 Strains of Fungi. *PLoS ONE*, *7*(11), e49679. https://doi.org/10.1371/journal.pone.0049679

Conserved Domains Database (CDD) and Resources. (n.d.). Retrieved December 17, 2018, from https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml

Davies, G., & Henrissat, B. (1995). Structures and mechanisms of glycosyl hydrolases. *Structure (London, England : 1993)*, *3*(9), 853–859. https://doi.org/10.1016/S0969-2126(01)00220-9

de Boer, A. S., Priest, F., & Diderichsen, B. (1994). On the industrial use of Bacillus licheniformis: a review. *Applied Microbiology and Biotechnology*, *40*(5), 595–598. https://doi.org/10.1007/BF00173313

de Souza, P. M., & de Oliveira Magalhães, P. (2010). Application of microbial α-amylase in industry - A review. *Brazilian Journal of Microbiology : [Publication of the Brazilian Society for Microbiology]*, *41*(4), 850–861. https://doi.org/10.1590/S1517-83822010000400004

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, *5*(1), 113. https://doi.org/10.1186/1471-2105-5-113

Efron, B., Halloran, E., & Holmes, S. (1996). Bootstrap confidence levels for phylogenetic trees. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(23), 13429–13434. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8917608

Filogen, I. (2007). Tema 5: Métodos de distancia y prueba de bootstrap BioInfo aplicada a estudios de ecología y sistemática molecular de bacterias, UFLA, Lavras, MG, Brasil, Nov.2007, 1–7.

Filogenias moleculares. (n.d.). Retrieved from http://www.fcnym.unlp.edu.ar/catedras/taxonomia/Teoricos2014/filogenias moleculares-2014.pdf

Finn, R. D., Clements, J., & Eddy, S. R. (2011). HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, *39*(Web Server issue), W29-37. https://doi.org/10.1093/nar/gkr367

Ghani, M., Ansari, A., Aman, A., Zohra, R. R., Siddiqui, N. N., & Qader, S. A. U. (2013). Isolation and characterization of different strains of Bacillus licheniformis for the production of commercially significant enzymes. *Pakistan Journal of Pharmaceutical Sciences*, *26*(4), 691–697.

Gopinath, S. C. B., Anbu, P., Arshad, M. K. M., Lakshmipriya, T., Voon, C. H., Hashim, U., & Chinni, S. V. (2017). Biotechnological Processes in Microbial Amylase Production. *BioMed Research International*, *2017*, 1272193. https://doi.org/10.1155/2017/1272193

Gowhar, H. D., Azra, N. K., Ruqeya, N., Suhaib, A. B., & Tauseef, A. M. (2014). Biotechnological production of -amylases for industrial purposes: Do fungi have potential to produce -amylases? *International Journal of Biotechnology and Molecular Biology Research*, *5*(4), 35–40. https://doi.org/10.5897/IJBMBR2014.0196

Gupta, R., Gigras, P., Mohapatra, H., Goswami, V. K., & Chauhan, B. (2003). Microbial α-amylases: A biotechnological perspective. *Process Biochemistry*, *38*(11), 1599–1616. https://doi.org/10.1016/S0032-9592(03)00053-0

Hatti-kaul, R. (2009). ENZYME PRODUCTION, *V*.

Henrissat, B., Callebaut, I., Fabrega, S., Lehn, P., Mornon, J. P., & Davies, G. (1995). Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proceedings of the National Academy of Sciences of the United States of America*, *92*(15), 7090–7094. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7624375

Howe, K., Bateman, A., & Durbin, R. (2002). QuickTree: Building huge neighbour-joining trees of protein sequences. *Bioinformatics*, *18*(11), 1546–1547. https://doi.org/10.1093/bioinformatics/18.11.1546

Huma, T., Maryam, A., Rehman, S. U., Qamar, M. T. U., Shaheen, T., Haque, A., & Shaheen, B. (2014). Phylogenetic and Comparative Sequence Analysis of Thermostable Alpha Amylases of kingdom Archea, Prokaryotes and Eukaryotes. *Bioinformation*, *10*(7), 443–448. https://doi.org/10.6026/97320630010443

Hussain, I., Siddique, F., Mahmood, M. S., & Ahmed, S. I. (2013). A review of the microbiological aspect of α-amylase production. *International Journal of Agriculture and Biology*, *15*(5), 1029–1034.

Inferring Phylogeny using Maximum Likelihood in R (phangorn) - AnthroTree - DukeWiki. (n.d.). Retrieved December 19, 2018, from https://wiki.duke.edu/pages/viewpage.action?pageId=131172124

J.Charnock Simon, V. M. B. (2005). *Enzymes: Industrial and analytical applications*. Retrieved from https://www.megazyme.com/docs/default-source/analytical-applications-downloads/enzymes_indsutrial_and_analytical_appliation_eng.pdf?sfvrsn=91dce65_4

Jones, D. T., & Swindells, M. B. (2002). Getting the most from PSI-BLAST. *Trends in Biochemical Sciences*. https://doi.org/10.1016/S0968-0004(01)02039-4

Kathiresan, K., & Manivannan, S. (2006). Alpha-Amylase production by Penicillium fellutanum\nisolated from mangrove rhizosphere soil. *African Journal of Biotechnology*, *5*(10), 829–832. https://doi.org/10.5897/ajb05.373

Kirk, O., Borchert, T. V., & Fuglsang, C. C. (2002). Industrial enzyme applications. *Current Opinion in Biotechnology*, *13*(4), 345–351. https://doi.org/10.1016/S0958-1669(02)00328-2

Krieger, E., Nabuurs, S. B., & Vriend, G. (2003). *HOMOLOGY MODELING*. Retrieved from www.yasara.com

Kumar, S. (2005). Molecular clocks: four decades of evolution. *Nature Reviews Genetics*, *6*(8), 654–662. https://doi.org/10.1038/nrg1659

Laoide, B. M., & McConnell, D. J. (1989). cis sequences involved in modulating expression of Bacillus licheniformis amyL in Bacillus subtilis: effect of sporulation mutations and catabolite repression

resistance mutations on expression. *Journal of Bacteriology*, *171*(5), 2443–2450. https://doi.org/10.1128/jb.171.5.2443-2450.1989

Leisola, M., Jokela, J., Pastinen, O., Turunen, O., & Schoemaker, H. E. (2009). *Industrial Use of Enzymes*. Retrieved from https://www.eolss.net/sample-chapters/C03/E6-54-02-10.pdf

Luo, A., Qiao, H., Zhang, Y., Shi, W., Ho, S. Y., Xu, W., … Zhu, C. (2010). Performance of criteria for selecting evolutionary models in phylogenetics: A comprehensive study based on simulated datasets. *BMC Evolutionary Biology*, *10*(1). https://doi.org/10.1186/1471-2148-10-242

Machida, M., Yamada, O., & Gomi, K. (2008). Genomics of Aspergillus oryzae: learning from the history of Koji mold and exploration of its future. *DNA Research : An International Journal for Rapid Publication of Reports on Genes and Genomes*, *15*(4), 173–183. https://doi.org/10.1093/dnares/dsn020

Naidu, M. A., & Saranraj, P. (2013). Bacterial Amylase : A Review. *International Journal of Pharmaceutical & Biology Archives*, *4*(2), 274–287. https://doi.org/10.5829/idosi.ijmr.2013.4.2.75170

Nestler, E. J., & Greengard, P. (1999). Protein Phosphorylation is of Fundamental Importance in Biological Regulation. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK28063/

Orr, I. (n.d.). *Introduction to Phylogenetic Analysis*. Retrieved from https://bip.weizmann.ac.il/education/course/introbioinfo/03/lect12/phylogenetics.pdf

Pais, F. S.-M., Ruy, P. de C., Oliveira, G., & Coimbra, R. S. (2014). Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology : AMB*, *9*(1), 4. https://doi.org/10.1186/1748-7188-9-4

Paradis, E. (2012). *Paradis-2012-Analysis of Phylogenetics and Evolution*.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *Chapter 3*, Unit3.1. https://doi.org/10.1002/0471250953.bi0301s42

Peng, C. (2007). *DISTANCE BASED METHODS IN PHYLOGENTIC TREE CONSTRUCTION*. Retrieved from https://pdfs.semanticscholar.org/1388/6f47e0077240f23b55b2bc1fb7589bd85295.pdf

Pruitt, K. D., Tatusova, T., & Maglott, D. R. (2004). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, *33*(Database issue), D501–D504. https://doi.org/10.1093/nar/gki025

Reddy, N., Nimmagadda, A., & Rao, K. R. S. S. (2002). African journal of biotechnology. *African Journal of Biotechnology*, *2*(12), 645–648. https://doi.org/10.1002/smi.2619

Rothstein, D. M., Devlin, P. E., & Cate, R. L. (1986). Expression of α-amylase in Bacillus licheniformis. *Journal of Bacteriology*, *168*(2), 839–842. https://doi.org/10.1128/jb.168.2.839-842.1986

Saini, R., Singh Saini, H., Dahiya, A., & Harnek Singh Saini, C. (2017). Amylases: Characteristics and industrial applications. *~ 1865 ~ Journal of Pharmacognosy and Phytochemistry*, *6*(4), 1865–1871. Retrieved from http://www.phytojournal.com/archives/2017/vol6issue4/PartAA/6-4-407-141.pdf

Singh, R., Kumar, M., Mittal, A., & Mehta, P. K. (2016). Microbial enzymes: industrial progress in 21st century. *3 Biotech*, *6*(2), 174. https://doi.org/10.1007/s13205-016-0485-8

Soltis, D. E., & Soltis, P. S. (2003). Applying the Bootstrap in Phylogeny Reconstruction. *Statistical Science*, *18*(2), 256–267. https://doi.org/10.1214/ss/1063994980

Stam, M. R., Danchin, E. G. J., Rancurel, C., Coutinho, P. M., & Henrissat, B. (2006). Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of -amylase-related proteins. *Protein Engineering Design and Selection*, *19*(12), 555–562. https://doi.org/10.1093/protein/gzl044

Sundarram, A., & Murthy, T. P. K. (2014). α-Amylase Production and Applications: A Review. *Journal of Applied & Environmental Microbiology*, *2*(4), 166–175. https://doi.org/10.12691/JAEM-2-4-10

Sutton, S. (2008). *Multiple Sequence Alignment: A Critical Comparison of Four Popular Programs*. *Biochemistry 218 Final Project*. Retrieved from http://biochem218.stanford.edu/Projects 2008/Sutton2008.pdf

Udenwobele, D. I., Su, R.-C., Good, S. V, Ball, T. B., Varma Shrivastav, S., & Shrivastav, A. (2017). Myristoylation: An Important Protein Modification in the Immune Response. *Frontiers in Immunology*, *8*, 751. https://doi.org/10.3389/fimmu.2017.00751

Using Conserved Domains to Find Protein Homologs | NCBI Insights. (n.d.). Retrieved December 17, 2018, from https://ncbiinsights.ncbi.nlm.nih.gov/2013/02/12/using-conserved-domains-to-find-functional-homologs/

van der Kaaij, R. M., Janecek, S., van der Maarel, M. J. E. C., & Dijkhuizen, L. (2007). Phylogenetic and biochemical characterization of a novel cluster of intracellular fungal -amylase enzymes. *Microbiology*, *153*(12), 4003–4015. https://doi.org/10.1099/mic.0.2007/008607-0

Vengadaramana, A. (2014). Industrial Important Microbial alpha-Amylase on Starch-Converting Process Scholars Academic Journal of Pharmacy (SAJP) Review Article Industrial Important Microbial alpha-Amylase on Starch-Converting Process, (January 2013).

Suvd, D., Fujimoto, Z., Takase, K., Matsumura, M., & Mizuno, H. (2001). Crystal structure of Bacillus stearothermophilus alpha-amylase: possible factors determining the thermostability. *Journal of Biochemistry*, *129*(3), 461–468. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11226887

10. ANNEX

10.1. *B.licheniformis* alpha amylase data set

Attached as external annex in a txt format, file name is bacillus_dataset.txt

10.2. *A.oryzae* alpha amylase data set

Attached as external annex in a txt format, file name is aspergillus_dataset.txt

10.3 Phylogenetic analysis

A part from the attached code below, two Rmarkdowns with the code used to perform the phylogenetic analysis are included as external annex.

- Bl_ phylogenetics.rmd contains the code to generate the *B.lichenifromis* phylogenetic analysis.
- Ao_phylogenetics.rmd contains the code to generate the *A.oryzae* phylogenetic analysis.

Four pdf containing the phylogenetics trees are included as external annex.

- BLnj: Neighbor joining tree for *B.licheniformis* data set
- BLml: Maximum likelihood tree for *B.licheniformis* data set
- AOnj: Neighbor joining tree for *A.oryzae* data set.
- AOml: Maximum likelihood tree for *A.oryzae* data set.

10.3.1 *B.licheniformis* phylogenetic analysis code

```
library(ape)
library(phangorn)
library(phytools)
library(geiger)
library(devtools)
library(adegenet)
library(seqinr)
library(msa)
library(tools)
```

Multiple sequence alignment

```
# Dataset as fasta file
BLfasta<-read.fasta(file = "blicheniformis_dataset.fas",seqtype = "AA")

# Perform the multiple sequence aligment from the bacillus licheniformis
data set
Sequences<-readAAStringSet("blicheniformis_dataset.fas")
seqalign<-msaMuscle(Sequences,cluster = "upgma") # MUSCLE algorithm and
clustered by UPGMA
BLmsa <- msaConvert(seqalign, type="seqinr::alignment")

myAligment<-msaCheckNames(seqalign,replacement = "",verbose = TRUE)

# Convert the aligment file into a fasta file
alignment2Fasta <- function(alignment, filename) {
```

```
  sink(filename)

  n <- length(rownames(alignment))
  for(i in seq(1, n)) {
    cat(paste0('>', rownames(alignment)[i]))
    cat('\n')
    the.sequence <- toString(unmasked(alignment)[[i]])
    cat(the.sequence)
    cat('\n')
  }

  sink(NULL)
}
alignment2Fasta(seqalign, 'blmsa.fasta')

#Concensus sequence
printSplitString <- function(x, width=getOption("width") - 1)
{
starts <- seq(from=1, to=nchar(x), by=width)
for (i in 1:length(starts))
cat(substr(x, starts[i], starts[i] + width - 1), "\n")
}

printSplitString(msaConsensusSequence(seqalign))
```

Neighbor joining tree

```
# Compute the distance matrix from the aligment
d <- dist.alignment(BLmsa)
heatmap(x=as.matrix(d),Rowv=NA, Colv=NA, symm=TRUE)

#Compute the neighbor-joining tree
BLnj<-nj(d) #performs the neighbor-joining tree estimation by Saitou and
Nei
BLnj

#Compute bootstrap values for the nj tree
NJtree <- function(alignment, type)
  {

     # define a function for generating the distance matrix from the
previous aligment
     makemytree <- function(alignmentmat)
     {
        alignment <- ape::as.alignment(alignmentmat)
        if      (type == "protein")
        {
           mydist <- dist.alignment(alignment)
        }
        else if (type == "DNA")
        {
           alignmentbin <- as.DNAbin(alignment)
           mydist <- dist.dna(alignmentbin)
        }
        mytree <- nj(mydist)
     return(mytree)
     }
```

```
     # infer a tree
     mymat  <- as.matrix.alignment(alignment)
     mytree <- makemytree(mymat)
     # bootstrap the tree
     myboot <- boot.phylo(mytree, mymat, makemytree)
     mytree$node.label <- myboot    # make the bootstrap values be the
node labels
     return(mytree)
}


bootBL<- NJtree(BLmsa,"protein")

# Convert the tree into a Newick format
write.tree(BLnj, file = "BLnj")
write.tree(bootBL, file = "bootBLnj")
```

Maximum likelihood

```
blmsa<- as.phyDat(seqalign)
#computes a simoustanly comparation of the distances for aligned
sequences under the comon substitution  models JTT,LG and WAG.
mt <- modelTest(blmsa, model=c("JTT", "LG", "WAG"))
# Obtation of the lowest BIC and AIC value for the test computed above.
env <- attr(mt, "env")
eval(get(mt$Model[which.min(mt$BIC)], env), env)
eval(get(mt$Model[which.min(mt$AIC)], env), env)

BLdm<-dist.ml(blmsa,model = "JTT")
# Starting point to initialize the optimitzation process.
blnj<-NJ(BLdm)
blnj

#Pml provides a way to compute the likelihood of the data given a
phylogenetic tree and evolutionary model.
fitNJ<-pml(blnj, blmsa, model="JTT", k=4, inv=.2)

#Optimizes the tree topology. As well, optimizes the different model
parameters by computing the values that maximize the likelihood function.
fit <-optim.pml(fitNJ, rearrangement = "stochastic",optNni = TRUE,
                optInv=TRUE, optGamma=TRUE)
fit

#Bootstap the tree

bs<-bootstrap.pml(fit, bs=100, optNni=TRUE)

# Convert the tree into a Newick format
tree<-plotBS(fit$tree, bs)
write.tree(tree, file = "bootBLml")
```

10.3.2. *A.oryzae* phylogenetic analysis code

```
library(ape)
library(phangorn)
library(phytools)
library(geiger)
library(devtools)
library(adegenet)
library(seqinr)
library(msa)
library(tools)
```

Multiple sequence aligment

```
# Dataset as fasta file
BLfasta<-read.fasta(file ="aoryzae_dataset.fas",seqtype = "AA")

# Perform the multiple sequence aligment from the bacillus licheniformis
data set
Sequences<-readAAStringSet("aoryzae_dataset.fas")
seqalign<-msaMuscle(Sequences,cluster = "upgma") # MUSCLE algorithm and
clustered by UPGMA
AOmsa <- msaConvert(seqalign, type="seqinr::alignment")

# Convert the aligment file into a fasta file
alignment2Fasta <- function(alignment, filename) {
  sink(filename)

  n <- length(rownames(alignment))
  for(i in seq(1, n)) {
    cat(paste0('>', rownames(alignment)[i]))
    cat('\n')
    the.sequence <- toString(unmasked(alignment)[[i]])
    cat(the.sequence)
    cat('\n')
  }

  sink(NULL)
}
alignment2Fasta(seqalign, 'aomsa.fasta')

#Concensus sequence
printSplitString <- function(x, width=getOption("width") - 1)
{
starts <- seq(from=1, to=nchar(x), by=width)
for (i in 1:length(starts))
cat(substr(x, starts[i], starts[i] + width - 1), "\n")
}

printSplitString(msaConsensusSequence(seqalign))
```

Neighbor joining tree

```
# Compute the distance matrix from the aligment
d <- dist.alignment(AOmsa)
heatmap(x=as.matrix(d),Rowv=NA, Colv=NA, symm=TRUE)

#Compute the neighbor-joining tree
AOnj<-nj(d) #performs the neighbor-joining tree estimation by Saitou and
Nei
AOnj

#Compute bootstrap values for the nj tree
NJtree <- function(alignment, type)
  {

     # define a function for generating the distance matrix from the
previous aligment
     makemytree <- function(alignmentmat)
     {
        alignment <- ape::as.alignment(alignmentmat)
        if      (type == "protein")
        {
           mydist <- dist.alignment(alignment)
        }
        else if (type == "DNA")
        {
           alignmentbin <- as.DNAbin(alignment)
           mydist <- dist.dna(alignmentbin)
        }
        mytree <- nj(mydist)
     return(mytree)
     }
     # infer a tree
     mymat  <- as.matrix.alignment(alignment)
     mytree <- makemytree(mymat)
     # bootstrap the tree
     myboot <- boot.phylo(mytree, mymat, makemytree)
     mytree$node.label <- myboot    # make the bootstrap values be the
node labels
     return(mytree)
}


bootAO<- NJtree(AOmsa,"protein")

# Convert the tree into a Newick format
write.tree(AOnj, file = "AOnj")
write.tree(bootAO, file = "bootAOnj")
```

Maximum likelihood

```
aomsa<-as.phyDat(seqalign)
#computes a simoustanly comparation of the distances for aligned
sequences under the comon substitution  models JTT,LG and WAG.
mt <- modelTest(aomsa, model=c("JTT", "LG", "WAG"))
# Obtation of the lowest BIC and AIC value for the test computed above.
env <- attr(mt, "env")
```

```
eval(get(mt$Model[which.min(mt$BIC)], env), env)
eval(get(mt$Model[which.min(mt$AIC)], env), env)

aomsa<- as.phyDat(seqalign)
#computes distances for aligned sequences under the a certain
substitution model
aodm<-dist.ml(aomsa,model = "WAG")
# Starting point to initialize the optimitzation process.
aonj<-NJ(aodm)
aonj

#Pml provides a way to compute the likelihood of the data given a
phylogenetic tree and evolutionary model.
fitNJ<-pml(aonj, aomsa, model="WAG", k=4, inv=.2)
#Optimizes the tree topology. As well, optimizes the different model
parameters by computing the values that maximize the likelihood function.
fit<-optim.pml(fitNJ,  rearrangement  =  "stochastic",  optInv=TRUE,
optGamma=TRUE,optNni = TRUE)
fit

# Bootstrap the tree
bs<-bootstrap.pml(fit, bs=100, optNni=TRUE)

# Convert the tree into a Newick format
bstree<-plotBS(fit$tree,bs)
write.tree(bs, file = "AOml")
write.tree(bstree, file = "bootAOml")
```

10.3.3. Bibliography used for the phylogenetic analysis computed with R

Phylogenetic tree reconstruction . Retrieved January 2, 2019, from https://www.reconlearn.org/post/practical-phylogenetics.html

2.5.3 Selecting a substitution model with R and PHYML - AnthroTree - DukeWiki. Retrieved January 2, 2019, from https://wiki.duke.edu/display/AnthroTree/2.5.3+Selecting+a+substitution+model+with+R+and+PHYML

2.4 Inferring Phylogeny using Maximum Likelihood in R (phangorn) - AnthroTree - DukeWiki. Retrieved January 2, 2019, from https://wiki.duke.edu/pages/viewpage.action?pageId=131172124

Schliep, K. P. (2018). *Estimating phylogenetic trees with phangorn*. Retrieved from https://cran.r-project.org/web/packages/phangorn/vignettes/Trees.pdf